INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Statistical Framework for Modeling Asymmetrical Data with Dual Peaks

K M Sakthivel[1], **G Vidhya**[2]*

1 Professor, Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu, India
2 Research Scholar, Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu, India

OPEN ACCESS

Received: 07-05-2024

Accepted: 24-06-2024

Published: 16-07-2024

Citation: Sakthivel KM, Vidhya G (2024) Statistical Framework for Modeling Asymmetrical Data with Dual Peaks. Indian Journal of Science and Technology 17(27): 2829-2840. https://doi.org/10.17485/IJST/v17i27.1540

* **Corresponding author**.

vidhyastatistic96@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To create a comprehensive framework that effectively identifies the most suitable model for asymmetrical data based on its unique characteristics. **Methods:** This study proposed a new model named Gompertz-Gumbel distribution (GGD) based on the results from the framework which utilizes various statistical tools, as well as information criteria. A dip test is used to check the modality of the data. To propose a new model, the finite mixture model concept was employed. The location, scale, shape, and weight parameters of the GGD were estimated using the maximum likelihood estimation method. **Findings:** The suggested framework exhibits superior performance in developing a suitable model for the asymmetrical data with dual peaks, resulting in the best fit for the data. To validate the effectiveness of the proposed model, it has been compared with various models like Gaussian models and two-component mixture models. The GGD's properties have also been determined. The various shapes of the GGD were also analyzed. **Novelty:** A novel framework is proposed to identify the appropriate model for the asymmetrical data with dual peaks that outperform the existing models. It shows the significance of the framework.

Keywords: Lifetime distributions; Mixture models; Information Criteria; Goodness of fit; Asymmetrical data

1 Introduction

Data plays an important role in many fields in the modern era, including reliability, economics, finance, medicine, engineering, and so on. However, not all data follow symmetrical patterns; most belong to asymmetrical patterns, such as being skewed to the left or right or showing multiple peaks. In such a case, we are unable to determine which of the current distributions best fits the data. In order to choose the appropriate model for the asymmetrical data with dual peaks based on the properties of the data, we designed a model selection procedure as a framework. We incorporate the finite mixture model in the model selection procedure to develop a new model. Even though many techniques, such as the transformation method, composite method, transformed-transformer method, α-power transformation, method of adding a new parameter, etc., are described in the literature to generate new distributions, the finite mixture model method is one of the best methods for unimodal as well as bimodal datasets.

The first noteworthy investigation of the finite mixture model was presented by renowned biometrician Karl Pearson[1]. He created a model that combined two normal probability distributions with different means and variances in a proportionate way. After that, several researchers have since created a proportionate mix of probability distributions with different parameters. Subsequently, Lindley[2] developed the Lindley distribution by combining exponential and gamma distributions with varying proportions. Following that, several researchers have developed the one-parameter mixture model by combining exponential and gamma distributions with different proportions. The study's authors first developed a model, which they subsequently applied to real-time data.

To determine the importance of the one-parameter mixture model, Vijay Sharma[3] did a comparison study on many one-parameter mixture models. The Kpenadidum distribution is a one-parameter mixture model created by Barinaadaa[4] using $exponential\,(\theta)$, $gamma\,(3,\theta)$, $gamma\,(4,\theta)$ with different mixing proportions. Ganaie[5] exponentiated the one-parameter Aradhana mixture model to boost the model's flexibility, whereas Tesfalem Eyob[6] changed the mixture model by adding more parameters to the distribution. Iwok Iberedem Aniefiok[7] combined two gamma distributions to create a single-parameter model and examined their characteristics. Mohammed Benrabia[8] introduced a two-parameter mixture model by combining exponential and gamma distributions with some proportions. Geoffrey J McLachlan[9] gives a brief note on finite mixture models and their importance in statistical analysis. In order to model the continuous data, Pinho[10] created a continuous cumulative distribution function (cdf) using the PIPE algorithm as a tool. The R package for building and assessing the mixture models is provided by Lukas Sablica[11]. Numerous writers developed and worked on mixture probability models.

Even though much work has been done in a mixture model with various combinations, we often don't know which model will suit the data perfectly. It is impractical to test all possible distributions to find the best fit for the data. Moreover, determining the correct number of the mixture components can be challenging in the classic mixture model. The mixture models typically assume a specific form for each component (e.g. Gaussian). If the true distribution differs significantly, the model may not fit well.

In the exponential-gamma mixture models, there are many distributions, such as the Rama distribution[12], Janardan distribution[13], Sushila distribution[14], and Exponential-Gamma distribution[15] among others. These distributions are designed for unimodal datasets and often provide a good fit. However, it raises the question: why is the combination of exponential and gamma distributions particularly effective for making good mixture models?

To address this issue, we incorporated other standard models with diverse properties and shapes to create a mixture of models. We developed a framework to select the best fitting model for the given data, eliminating the struggle of choosing the most appropriate model.

Previous works have focused on either symmetrical, unimodal, or bimodal datasets, while we work with data that is unimodal but also has dual peaks. Dealing with unimodal asymmetric data that contains dual peaks; thus, it is challenging to determine which distribution would be most appropriate in this scenario. So, we developed an approach to address this issue and identify which model most closely matches the actual data. This necessitated first determining the characteristics of the data through a thorough analysis and then selecting the model that would best fit these attributes. Through this process, we developed a mixture model that accurately represents the data and estimates it reliably.

2 Methodology

A systematic approach to the analysis of dual-peak asymmetric data is provided by this algorithm. This model selection procedure separates the data into two groups by applying clustering techniques. After the data has been partitioned, the next step is to fit a probability distribution to each partition. This is done to determine the best-fit model for each group of data. The best-fit model is then chosen based on factors like goodness of fit, model complexity, and interpretability of the results. A mixture model is produced by combining the best-fit models; This model thoroughly comprehends the data and spots patterns and trends that might not be obvious when examining the data.

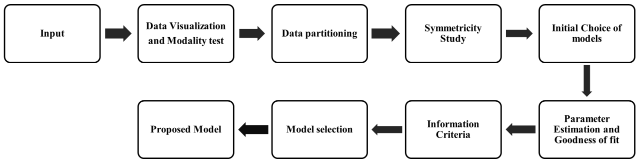

Fig 1. Framework for selecting a model

This algorithm consists of the following steps:

Step 1: Consider the data.

Step 2: To determine the pattern of the random variable, visualize the data. likewise, verify the data's modality.

Step 3: Applying the clustering algorithm, split the data into two groups.

Step 4: To identify the asymmetry, determine the skewness for each component of the data.

Step 5: Take into consideration the fundamental distributions for modeling according to the properties of the data.

Step 6: Use Maximum Likelihood Estimation (MLE) to estimate the parameter values for the first portion of the data using appropriate probability distributions and determine the model's adequacy by computing the goodness of fit and information metrics.

Step 7: Repeat the process for the second half of the data.

Step 8: Based on the minimized values of -2LogLikelihood(-2LL), AIC, BIC, and AICC, select a better model among the distributions that were taken into consideration in steps 6 and 7.

Step 9: Propose a new model by combining the selected models from Step 8.

The model selection process is refined using goodness of fit tests like Cramer Von Mises (CVM)[16], Anderson Darling (AD) test[16], Kolmogorov Smirnov (KS) test[16], and information criteria like Akaike Information Criteria (AIC)[17],[18], Bayesian Information Criteria (BIC)[18], and Corrected Akaike Information Criteria (AICC)[18], ensuring accurate and reliable insights from the data. The framework provides useful insights that may be applied in a variety of fields, including engineering, finance, marketing, and healthcare. It does this by effectively analyzing complicated data sets using various analytical tools and statistical approaches.

It is essential to take into consideration a few assumptions before utilizing the framework. First and foremost, asymmetrical, dual-peak data needs to be used for the analysis. Furthermore, the framework is especially intended for scenarios in which there is a significant deviation from a normal distribution. The data must show some degree of skewness, either positive or negative, to be appropriate for analysis using this approach. Moreover, the framework works best with data that permits outliers to exist in the dataset and has heavy tails. It's crucial to remember that this framework was created specifically for univariate data analysis.

It is important to keep in mind that there are certain limitations to this procedure. This framework is specially designed to handle asymmetrical data with dual peaks and requires dividing the data into two distinct groups. Moreover, the method encompasses a constrained set of fundamental distributions, not exceeding twenty models that are rudimentary and widely employed across diverse disciplines. The manual manipulation of this framework with substantial data volumes may engender numerous challenges and protracted processing times. Subsequent research endeavors are poised to mechanize this process by leveraging statistical software and generating outcomes via uncomplicated algorithms. This strategic approach is poised to rationalize the process and enhance efficiency when handling larger datasets.

3 Results and discussion

Using the following real-time application, the inner workings of this framework are thoroughly discussed.

3.1 Applications

The dataset provides comprehensive information on the percentage of land area covered by forests in different countries. With 210 observations, it accurately reflects the forested areas in each country in 2010. The data is presented as a percentage, making it easy to compare and analyze. The dataset was obtained from http://data.un.org/Data.aspx?d=MDG\&f=seriesRowID\%3a567.

Table 1. Summary of the Forest Coverage Area (FCA) data

Minimum	1st Quartile (Q1)	Median	Mean	3rd Quartile (Q3)	Maximum	Skewness	Kurtosis
0.01	11.36	32.72	32.93	48.24	98.32	0.4759	2.4751

The results in Table 1 make it evident that there are differences between the mean and median values in our sample. The data is most likely skewed and not symmetric based on the disparity in central tendency measures. Given that the kurtosis suggests that the data is light-tailed, and the presence of skewness reinforces the positive skewness of the data, we may adopt the light-tailed-natured model for both data sections. Despite this non-normality, the data can still be used to create a statistical model. All things considered, these findings provide insight into the characteristics of our datasets and can guide future studies and modeling endeavors.

The data's graphical representation displays two peaks. Moreover, the kernel density plot indicates its asymmetric nature with a non-symmetric distribution. To confirm the data's modality, we employed the Hartigan dip test as proposed by Hartigan

(1985), yielding a dip statistic of 0.02967 (p-value = 0.218). The Hartigan dip test result confirms that the data is unimodal. The data was then split into two parts using K-mean clustering as part of Step 3 in our framework. The first segment consists of 118 observations, while the second contains 92 observations. The first segment is positively skewed, with a mean of 15.691, a median of 13.215, and a skewness value of 0.2284.

The most appropriate distribution for the data is chosen in our study by using fundamental models such as symmetric, heavy-tailed, light-tailed, positively skewed, and negatively skewed nature models. Since the initial portion of the data is positively skewed, we may choose a model that exhibits this property. The computation of the tools listed above (in Section 2) for the different distributions completes Step 6. All the findings were obtained using the R program, which is shown in Table 2 below. The model is filtered using the findings to choose a superior model.

Table 2. Estimated parameters value and Goodness of fit for the first part of the dataset

Models	Estimated parameter	-2LL	CVM	AD	KS	AIC	BIC	AICC
Logistic	$\widehat{a}$ =15.242 $\widehat{b}$ =7.0614	923.47	0.3716 (0.086)	2.7956 (0.055)	0.1037 (0.158)	927.47	933.01	927.58
Gumbel	$\widehat{k}$=10.0998 $\widehat{\lambda}$ =9.6648	907.89	0.3332 (0.109)	2.5511 (0.057)	0.0967 (0.220)	911.89	917.43	911.99
Weibull	$\widehat{k}$=1.069 $\widehat{\lambda}$ =16.025	885.04	0.4565 (0.051)	3.2367 (0.051)	0.1168 (0.080)	889.04	894.59	889.15
Gompertz	$\widehat{k}$=0.0436 $\widehat{\lambda}$ =0.0349	866.70	0.3523 (0.097)	3.4606 (0.076)	0.109 (0.121)	870.70	876.25	870.81

Based on the statistical analysis, we can confidently say that a better fit between our model and the data is suggested by lower KS, CVM, and AD test statistic values. It is a fact that lower KS, CVM, and AD values indicate a preferable model, whereas higher values suggest that the model does not fit well with the data and may not be a good option. We can compare the p-values from the KS, CVM, and AD tests for the many possible distributions. A p-value $\rightarrow 0$ denotes a weaker fit, whereas a p-value $\rightarrow 1$ denotes a better fit. After careful analysis, we have concluded that Rayleigh, Lindley, Exponential, Gamma, Lognormal, Lomax, Laplace, Cauchy, and Pareto distributions have greater KS, AD, and CVM statistic values and lower p-values, and hence, we have ignored them. The best fit for the data is then chosen after computing the information criteria. We have thoroughly examined the information criteria, and we are confident that the distribution with the lowest values of AIC, BIC, and AICC is the best fit. Table 2 simplifies the process of selecting the appropriate model. We can state with confidence that the Gompertz distribution is the best match for the first part of the data based on the parameters listed in the table.

To analyze the second part of the data, we followed the same process as we did for the first part of the data. The results of this analysis have been tabulated in Table 3. Upon analyzing this part of the data, we found that its mean value is 55.03 and the median is 51.23. This indicates that the data is asymmetric, with a positive skewness value of 0.8165. This positive skewness value further confirms the skewed distribution of the data. Given these findings, we recommend selecting a distribution that closely matches the positive skewness of the data while also being a light-tailed model, as we have previously mentioned. This will enable us to accurately model and analyze the data to draw meaningful conclusions.

Table 3. Estimated parameters value and Goodness of fit for the second part of the dataset

Model	Estimated parameter	-2LL	CVM	AD	KS	AIC	BIC	AICC
Logistic	$\widehat{a}$ =53.553 $\widehat{b}$ =8.789	764.97	0.23982 (0.202)	1.7879 (0.121)	0.11421 (0.1813)	768.97	774.013	769.10
Weibull	$\widehat{k}$=3.7444 $\widehat{\lambda}$ =60.846	765.93	0.30817 (0.1281)	2.0558 (0.086)	0.12514 (0.1121)	769.93	774.98	770.07
Lognormal	$\widehat{\mu}$ =3.9718 $\widehat{\sigma}$ =0.265	747.53	0.17248 (0.3282)	1.078 (0.319)	0.08434 (0.5296)	751.53	756.58	751.67
Gamma	$\widehat{\alpha}$=14.0315 $\widehat{\beta}$=0.255	751.09	0.20913 (0.2507)	1.3045 (0.231)	0.09308 (0.4028)	755.09	760.14	755.23
Gumbel	$\widehat{k}$=48.046 $\widehat{\lambda}$ =11.507	745.19	0.16738 (0.3411)	1.0561 (0.329)	0.080826 (0.5849)	749.19	754.23	749.32

The results from Table 3 indicated that the Gumbel distribution was the most suitable model for the second part of the data. With the two necessary components identified and obtained for both parts of the data, we can proceed to the next step.

In step 8 of our analysis, we examined the findings from steps 6 and 7, which revealed that the Gompertz distribution was a better fit for the first part of the data, while the Gumbel distribution was more suitable for the second part of the data. As we wanted to combine these two distributions, we used the finite mixture model to create a new distribution called the Gompertz-Gumbel distribution (GGD) in step 9. The function of the GGD is provided below, which can be used to model the entire dataset accurately.

In order to ensure the effectiveness and precision of our proposed model in handling data processing tasks, it is of utmost importance to conduct rigorous testing. Through this testing process, we will be able to showcase that our framework can generate superior results and surpass the performance of existing models.

To compare the proposed model with previously filtered models and frequently used mixture models such as the two-component normal distribution and the two-component log-normal distribution, we have repeated step 6 for the proposed model. The results of this testing process are presented in Table 4, providing a comprehensive overview of the model's performance and efficiency.

We also conducted an analysis using the "mclust" package in the R program to generate a Gaussian Mixture Model (GMM) for our data. The model indicated that the data could be represented using four-component Gaussian mixture models. However, after comparing the results with our proposed model, we found that our model performed significantly better than GMM. Our findings suggest that the proposed model is a superior alternative for modeling the data, as it provides more accurate results. These results have important implications for further research in this field and can be leveraged to develop more effective models for similar datasets.

Table 4. Estimated parameters value, Goodness of fit, and Model selection criteria for the data

Model	Estimated parameter	-2LL	CVM	AD	KS	AIC	BIC	AICC
Gamma	$\widehat{\alpha}$=1.0541 $\widehat{\beta}$=0.0320	1887.22	0.95051 (0.0032)	5.0943 (0.003)	0.1407 (0.0004)	1891.22	1897.92	1891.28
Weibull	$\hat{k}$=1.1408 $\widehat{\lambda}$ =34.212	1882.84	0.68611 (0.0137)	4.3118 (0.007)	0.1238 (0.0032)	1886.84	1893.53	1886.90
Logistic	$\hat{a}$ =31.711 $\hat{b}$ =13.866	1934.40	2.4044 (0.201)	2.1504 (0.076)	0.0923 (0.056)	1938.40	1945.09	1938.46
Normal	$\widehat{\mu}$ =32.925 $\widehat{\sigma}$ =23.618	1923.99	0.26448 (0.1708)	2.149 (0.076)	0.08356 (0.107)	1927.99	1934.69	1928.05
Gompertz	$\hat{k}$=0.0190 $\widehat{\lambda}$ =0.0176	1854.10	0.1958 (0.2759)	2.1504 (0.076)	0.0675 (0.295)	1858.10	1864.79	1858.16
Gumbel	$\hat{k}$=21.6896 $\widehat{\lambda}$ =19.471	1909.31	0.31553 (0.1222)	2.127 (0.078)	0.0756 (0.182)	1913.31	1920.01	1913.37
Two Component Log-Normal	$\widehat{w}$ =0.3739 $\widehat{\mu}_1$=1.6758 $\widehat{\sigma}_1$=1.6726 $\widehat{\mu}_2$=3.7104 $\widehat{\sigma}_2$=0.4346	1859.96	0.0684 (0.761)	0.6557 (0.596)	0.0561 (0.523)	1869.96	1886.70	1870.25
Two Component Normal	$\widehat{w}$ =0.0985 $\widehat{\mu}_1$=32.9279 $\widehat{\sigma}_1$=23.6643 $\widehat{\mu}_2$=32.9232 $\widehat{\sigma}_2$=23.6137	1923.992	0.2643 (0.1711)	2.1481 (0.076)	0.083529 (0.1067)	1933.99	1950.73	1934.28
Gompertz-Gumbel	$\widehat{w}$ =0.9821 $\hat{b}$=0.0208 $\hat{\eta}$=0.0161 $\widehat{\mu}$=10.736 $\widehat{\beta}$=0.0162	1846.97	0.1609 (0.358)	1.5934 (0.155)	0.0506 (0.656)	1856.94	1873.67	1857.23

Table 4 and Figure 2 make it clear that our suggested mixed probability model yields the best results, and our approach helps select a more suitable model for the asymmetrical with dual peak data.

3.2 Gompertz-Gumbel Distribution:

The two-component finite mixture model can be created using

$$f(x) = w_1 g_1(x) + w_2 g_2(x) \tag{1}$$

Where, $g_1(x) \sim \ Gompertz \ (b, \eta)$, $g_2(x) \sim \ Gumbel\,(\mu, \beta)$ and $w_1 = \omega$; $w_2 = 1 - w_1 = 1 - \omega$.

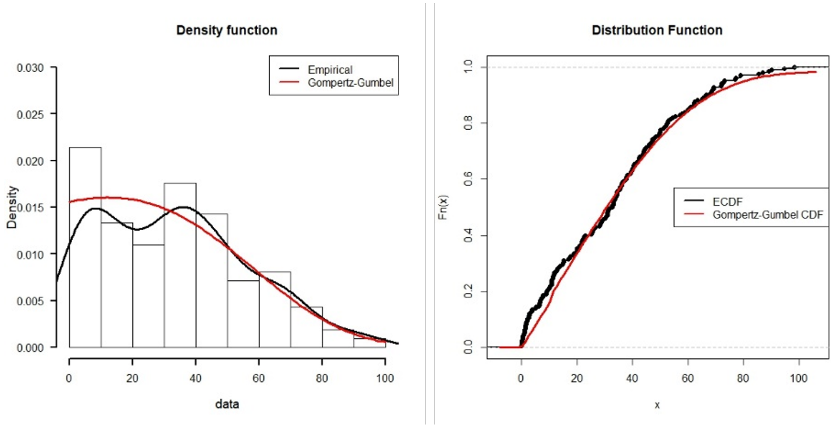

Fig 2. Comparison fit for the data

Let X ~ GGD $(\omega, b, \eta, \mu, \beta)$ then the Probability density function (pdf) and cumulative distribution function (cdf) for the GGD are given below and the different shapes of the model are figured in Figure 3 and Figure 4. The pdf of GGD is

$$f(x)=\frac{1}{(e^{\eta}-1)\omega+1}\left(\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)}+(1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}\right) \tag{2}$$

And the cdf of GGD is,

$$F(x)=\int_{-\infty}^{x}f(x)\,dx=\frac{e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}-\eta e^{bx}}\left(\left(\omega e^{\eta e^{bx}+\eta}-e^{\eta}\omega\right)e^{e^{\frac{\mu}{\beta}-\frac{x}{\beta}}}+(1-\omega)e^{\eta e^{bx}}\right)}{(e^{\eta}-1)\omega+1}$$

Simplified and rewrite

$$F(x)=\frac{(1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}}+e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta}-e^{\eta}\omega\right)}{(e^{\eta}-1)\omega+1} \tag{3}$$

$$For,\ b>0,\ \beta>0,\ \eta>0, 1>\ \omega>0, \mu\in R; -\infty<x<\infty$$

3.2.1 Properties of the GGD

The Reliability function of X is

$$S(x)=\frac{(e^{\eta}-1)\omega+1-(1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}}-e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta}-e^{\eta}\omega\right)}{(e^{\eta}-1)\omega+1} \tag{4}$$

The hazard function of X is

$$h(x)=\frac{\beta\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)}+(1-\omega)e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta\left((e^{\eta}-1)\omega+1-(1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}}-e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta}-e^{\eta}\omega\right)\right)} \tag{5}$$

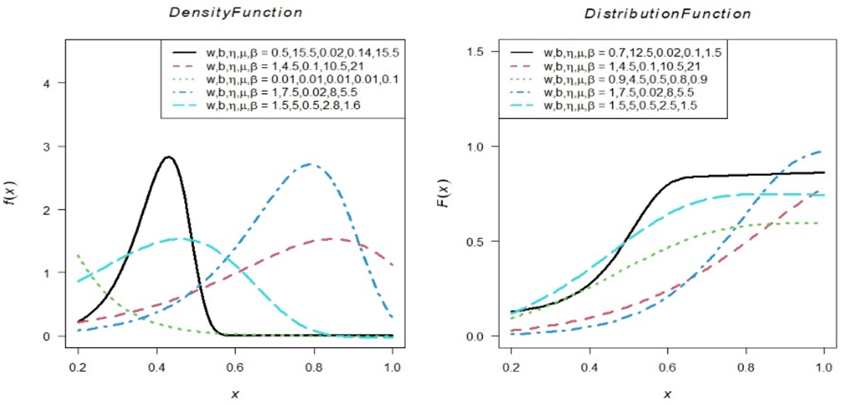

Fig 3. Various shapes of the GGD's Density Function and cumulative Distribution Function for various parameter values

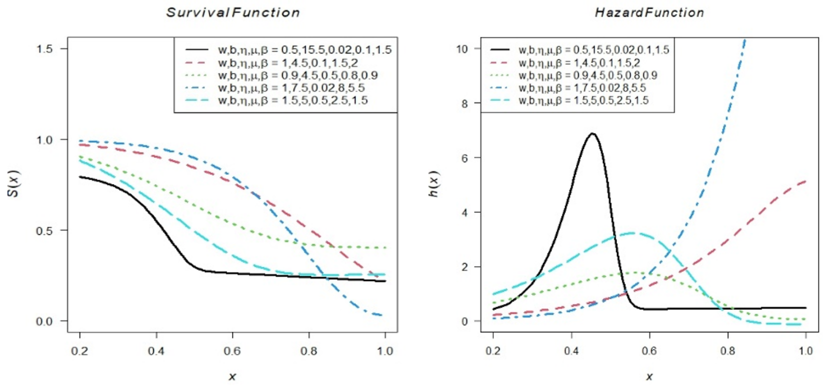

Fig 4. Various shapes of the GGD's Survival Function and Hazard Function for various parameter values

Reverse hazard rate of X is

$$\tau(x) = \frac{\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)} + (1-\omega)\,\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}}{e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}-\eta e^{bx}}\left(\left(\omega e^{\eta e^{bx}+\eta}-e^{\eta}\omega\right)e^{e^{\frac{\mu}{\beta}-\frac{x}{\beta}}}+(1-\omega)\,e^{\eta e^{bx}}\right)} \tag{6}$$

The cumulative hazard rate of X is

$$H(x) = -log\left(\frac{\left(e^{\eta}-1\right)\omega+1-e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}-\eta e^{bx}}\left(\left(\omega e^{\eta e^{bx}+\eta}-e^{\eta}\omega\right)e^{e^{\frac{\mu}{\beta}-\frac{x}{\beta}}}+(1-\omega)\,e^{\eta e^{bx}}\right)}{\left(e^{\eta}-1\right)\omega+1}\right) \tag{7}$$

The odds function of X is

$$\pi_O(x) = \frac{(1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}} + e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta} - e^{\eta}\omega\right)}{\left(e^{\eta}-1\right)\omega + 1 - (1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}} - e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta} - e^{\eta}\omega\right)} \tag{8}$$

The log-odds function is

$$\begin{aligned} LO(x) = {} & log\left((1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}} + e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta} - e^{\eta}\omega\right)\right) \\ & -log\left(\left(e^{\eta}-1\right)\omega + 1 - (1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}} - e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta} - e^{\eta}\omega\right)\right) \end{aligned} \tag{9}$$

The log-odds rate is defined as

$$LOR(x) = \frac{\dfrac{\beta\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)} + (1-\omega)e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta\left(\left(e^{\eta}-1\right)\omega + 1 - (1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}} - e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta} - e^{\eta}\omega\right)\right)}}{\dfrac{\left(e^{\eta}-1\right)\omega + 1 - (1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}} - e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta} - e^{\eta}\omega\right)}{\left(e^{\eta}-1\right)\omega+1}} \tag{10}$$

n^{th} order statistics

$$\begin{aligned} f_{X_{(n)}}(x) = n & \left(\frac{1}{\left(e^{\eta}-1\right)\omega+1}\left(\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)} + (1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}\right)\right) \\ & \left[\frac{(1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}} + e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta} - e^{\eta}\omega\right)}{\left(e^{\eta}-1\right)\omega+1}\right]^{(n-1)} \end{aligned} \tag{11}$$

1st order statistics

$$\begin{aligned} f_{X_{(1)}}(x) = \ n & \left(\frac{1}{\left(e^{\eta}-1\right)\omega+1}\left(\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)} + (1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}\right)\right) \\ & \times\left[\frac{\left(e^{\eta}-1\right)\omega + 1 - (1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}} - e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta} - e^{\eta}\omega\right)}{\left(e^{\eta}-1\right)\omega+1}\right]^{(n-1)} \end{aligned} \tag{12}$$

The pdf of a median of order statistic is

$$f_{m+1:n}(x) = \frac{(2m+1)}{m!m!}\left(\frac{1}{(e^{\eta}-1)\omega+1}\left(\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)}+(1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}\right)\right)$$
$$\times\left[\frac{(1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}}+e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta}-e^{\eta}\omega\right)}{(e^{\eta}-1)\omega+1}\right]^{m}$$
$$\times\left[\frac{(e^{\eta}-1)\omega+1-(1-\omega)e^{-e^{\frac{\mu}{\beta}-\frac{x}{\beta}}}-e^{-\eta e^{bx}}\left(\omega e^{\eta e^{bx}+\eta}-e^{\eta}\omega\right)}{(e^{\eta}-1)\omega+1}\right]^{m} \quad (13)$$

3.2.2 Estimation

The maximum likelihood estimates of the GGD $(\omega, b, \eta, \mu, \beta)$ parameters. Consider the following log-likelihood function l of a random sample $X_1, X_2, X_3, \ldots, X_n$ from a population following GGD$(\omega, b, \eta, \mu, \beta)$ with pdf (2)

$$l = -nlog\left((e^{\eta}-1)\omega+1\right)+\sum_{i=1}^{n} log\left(\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)}+(1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}\right) \quad (14)$$

On differentiating Equation (14) with respect to the parameters $\omega, b, \eta, \mu,\ and\ \beta$ and equating to zero, we obtain the following likelihood equations.

$$\frac{\partial l}{\partial \omega} = -\frac{n(e^{n}-1)}{\omega(e^{n}-1)+1}+\sum_{i=1}^{n}\frac{b\eta e^{\left(\eta+bx-\eta e^{bx}\right)}+\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}}{\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)}+(1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}} = 0 \quad (15)$$

$$\frac{\partial l}{\partial b} = \sum_{i=1}^{n}\frac{\eta\omega b\left(x-\eta x e^{xb}\right)e^{\left(\eta+bx-\eta e^{bx}\right)}+\eta\omega e^{\left(\eta+bx-\eta e^{bx}\right)}}{\omega b\eta e^{\left(\eta+bx-\eta e^{bx}\right)}+(1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}} = 0 \quad (16)$$

$$\frac{\partial l}{\partial \mu}=\sum_{i=1}^{n}\frac{(1-\omega)\left(\frac{1}{\beta}-\frac{e^{-\frac{x-\mu}{\beta}}}{\beta}\right)e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta\left(\omega b\eta e^{(\eta+bx-\eta e^{bx})}+(1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}\right)}=0 \tag{17}$$

$$\frac{\partial l}{\partial \beta}=\sum_{i=1}^{n}\frac{\frac{(1-\omega)\left(\frac{x-\mu}{\beta^2}-\frac{(x-\mu)e^{-\frac{x-\mu}{\beta}}}{\beta^2}\right)e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}-\frac{\frac{(1-\omega)e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta^2}}{\beta}}{\omega b\eta e^{(\eta+bx-\eta e^{bx})}+(1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}}=0 \tag{18}$$

$$\frac{\partial l}{\partial \eta}=-\frac{n\omega e^{\eta}}{\omega(e^{n}-1)+1)}+\sum_{i=1}^{n}\frac{b\omega(1-e^{bx})\eta e^{\eta+bx-\eta e^{bx}}+b\omega e^{\eta+bx-\eta e^{bx}}}{\omega b\eta e^{(\eta+bx-\eta e^{bx})}+(1-\omega)\frac{e^{-e^{-\frac{x-\mu}{\beta}}-\frac{x-\mu}{\beta}}}{\beta}}=0 \tag{19}$$

Now the MLEs $\hat{\omega}$, $\hat{b}$, $\hat{\eta}$, $\hat{\mu}$ *and* $\hat{\beta}$ of the parameters $\omega, b, \eta, \mu, \beta$ of GGD with pdf (2) can be obtained by solving the likelihood Equations (15), (16), (17), (18) and (19) with the help of statistical software R.

3.2.3 Simulation Studies

A simulation study is used to assess how well ML estimations perform. For this, we replicate the Gompertz-Gumbel Distribution (GGD) parameters 1000 times using different sample sizes, ranging from 25 to 250. Using R programming, we generated a random sample of GGD by utilizing the Monte Carlo simulation approach. The performance of the MLEs is evaluated for each sample by computing the mean value, average bias, and root-mean-square error (RMSE), which are displayed in Table 5.

From Table 5, it is observed that the sample size of n increases, and the bias and RMSE tend to decrease. Therefore, a larger sample size indicates more accurate results

Table 5. Simulation analysis: Mean, Bias, and RMSE values for GGD with various sample sizes

		Case (i): ω=0.1, b =0.1, η=0.2, μ=0.5, β=1.5			**Case (ii): ω=0.1, b =0.5, η=0.1, μ=2.5, β=0.1**		
n	**Parameters**	**Mean**	**Average Bias**	**RMSE**	**Mean**	**Average Bias**	**RMSE**
25	ω	0.5181	0.1600	0.1865	0.5184	0.1598	0.1867
	b	1.1359	0.3556	0.4658	1.1440	0.3466	0.4658
	η	0.0049	0.0049	0.0124	0.0065	0.0003	0.0147
	μ	0.8989	0.0092	0.0378	2.5820	0.0086	0.0113
	β	0.0046	4.1e^{-03}	0.0708	0.0013	1.09e^{-03}	0.0081
50	ω	0.5101	0.0502	0.0894	0.5076	-0.0325	0.0756
	b	1.0696	0.2517	0.3273	1.089	0.3148	0.4244
	η	0.0047	0.0043	0.0114	0.0051	0.0001	0.0116
	μ	0.8917	0.0022	0.0024	2.5606	0.0050	0.0058

Continued on next page

Table 5 continued

	β	0.0003	$8.08e^{-04}$	0.0008	0.0003	$7.65e^{-04}$	0.0008
	ω	0.5045	0.0467	0.0864	0.5028	-0.0471	0.0682
	b	1.0525	0.1592	0.3050	1.0495	0.0239	0.1644
100	η	0.0036	0.0031	0.0083	0.0034	0.0001	0.0086
	μ	0.8915	0.0019	0.0022	2.3674	0.0021	0.0025
	β	0.0003	$3.64e^{-05}$	0.0001	0.0003	$1.74e^{-04}$	0.0002
	ω	0.5004	0.0161	0.0549	0.5025	-0.0494	0.0317
	b	1.0463	0.0778	0.1615	1.0189	-0.0500	0.1151
250	η	0.0020	0.0018	0.0062	0.0012	0.0001	0.0033
	μ	0.8912	0.0012	0.0021	2.0739	0.0015	0.0018
	β	0.0002	$1.65e^{-05}$	0.0001	0.0003	$1.39e^{-05}$	0.0001

4 Conclusion

This study presents an algorithm that deals with the challenges of asymmetric data with dual peaks. The proposed algorithm is based on a unique framework that combines various probability distributions, which helps us to determine the best mixture of probability models for the given data. The new approach has been thoroughly evaluated by subjecting it to numerous goodness of fit tests and information criteria to ensure that it outperforms existing models. It relies on a finite mixture model, which combines multiple probability models to provide a more accurate representation of the data. To accurately estimate the parameters for our proposed model and select the most appropriate model for the data, we use the maximum likelihood estimation method. A thorough analysis was done on the statistical characteristics of our proposed mixture model and evaluated their performance against existing models. We use Geospatial data to demonstrate the performance of the proposed algorithm, which has shown accuracy and efficiency in selecting the best model for asymmetric data with dual peaks, making it a valuable tool for researchers and analysts working in various domains. The proposed model provides a better fit than the two-component normal, two-component lognormal, and four-component Gaussian models. Additionally, it fits the data better than the Gamma, Weibull, logistic, Gompertz, normal, and Gumbel models.

References

1) Pearson K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society, Series A*. 1894;185:71–110. Available from: https://doi.org/10.1098/rsta.1894.0003.
2) Lindley DV. Fiducial distributions and Bayes theorem. *Journal of the Royal Statistical Society, Series B*. 1958;20(1):102–107. Available from: https://doi.org/10.1111/j.2517-6161.1958.tb00278.x.
3) Sharma V, Shanker R, Shanker R. On some one parameter lifetime distributions and their applications. *Annals of Biostatistics & Biometric Applications*. 2019;3(2). Available from: http://dx.doi.org/10.33552/ABBA.2019.03.000556.
4) Nwikpe BJ, Cleopas IE. Kpenadidum distribution: statistical properties and application. *Asian Journal of Pure and Applied Mathematics*. 2022;4(1):759–765. Available from: https://globalpresshub.com/index.php/AJPAM/article/view/1699.
5) Ganaie RA, Rajagopalan V. Exponentiated Aradhana distribution with properties and applications in engineering sciences. *Journal of Scientific Research*. 2022;6(1):316–325. Available from: http://dx.doi.org/10.37398/JSR.2022.660134.
6) Eyob T, Shanker R, Shukla KK, Leonida TA. Weighted quasi-Akash distribution: properties and applications. *American Journal of Mathematics and Statistics*. 2019;9(1):30–43. Available from: http://dx.doi.org/10.5923/j.ajms.20190901.05.
7) Aniefiok II, Nwikpe BJ. The Iwok-Nwikpe distribution: statistical properties and application. *Asian Journal of Probability and Statistics*. 2021;15(1):35–45. Available from: https://doi.org/10.9734/ajpas/2021/v15i130347.
8) Benrabia M, Alzoubi MAL. Alzoubi distribution: properties and applications. *Journal of Statistics Applications & Probability- An International Journal*. 2022;11(2):625–640. Available from: https://doi.org/10.18576/jsap/110221.
9) Mclachlan G, Lee SX, Rathnayake S. Finite mixture models. *Annual Review of Statistics and Its Application*. 2019;6(1):355–378. Available from: https://doi.org/10.1146/annurev-statistics-031017-100325.
10) Pinho GBL, Nobre JS, Cordeiro GM. Continuous probability distributions are generated by the PIPE algorithm. *Anais da Academia Brasileira de Ciencias*. 2022;94(3). Available from: https://doi.org/10.1590/0001-3765202220201542.
11) Sablica L, Hornik K. mistr: A computational framework for mixture and composite distributions. *The R Journal*. 2020;12(1):283–299. Available from: https://doi.org/10.32614/rj-2020-003.
12) Shanker R. Rama distribution and its application. *International Journal of Statistics and Applications*. 2017;7(1):26–35. Available from: http://article.sapub.org/10.5923.j.statistics.20170701.04.html.
13) Shanker R, Sharma S, Shanker U, Shanker R. Janardan distribution and its application to waiting times data. *Indian Journal of Applied Research*. 2013;3(8):500–502. Available from: https://doi.org/10.15373/2249555x/aug2013/157.
14) Shanker, Sharma S, Shanker U, Shanker R. Sushila distribution and its application to waiting times data. *Opinion- International Journal of Business Management Special Issue on Role of Statistics in Management and Allied Sciences*. 2013;3(2):1–11. Available from: https://www.researchgate.net/publication/323265030_SUSHILA_distribution_and_its_application_to_waiting_times_data.
15) Adewusi O, Ogunwale OD, Ayeni TM. Exponential-Gamma distribution. *International Journal of Emerging Technology and Advanced Engineering*. 2019;9(10):245–249. Available from: https://www.researchgate.net/publication/337008547_Exponential-Gamma_Distribution.

16) Stephens MA. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*. 1974;69(347):730–737. Available from: https://doi.org/10.2307/2286009.

17) Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974;19(6):716–723. Available from: https://doi.org/10.1109/TAC.1974.1100705.

18) Wit E, Heuvel EVD, Romeijn JW. 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*. 2012;66(3):217–236. Available from: https://doi.org/10.1111/j.1467-9574.2012.00530.x.