computer sciences and mathematics forum

MDPI

Article

Two-component unit Weibull mixture model to analyze vote proportions

Renata Rojas Guerra[1]*, Fernando A. Peña-Ramírez[2] Charles P. Mafalda[3] and Gauss Moutinho Cordeiro[3]

[1] Universidade Federal de Santa Maria
[2] Universidad Nacional de Colombia
[3] Universidade Federal de Pernambuco
* Correspondence: renata.r.guerra@ufsm.br

Abstract: In this paper, we present a two-component Weibull mixture model. An important property is that this new model accommodates bimodality, which can appear in data representing phenomena in some heterogeneous populations. We provide statistical properties, such as the quantile function and moments. Also, the Expectation-Maximization (EM) algorithm is used to find maximum-likelihood estimates of the model parameters. Further, a Monte Carlo study is carried out to evaluate the performance of the estimators on finite samples. The new model's relevance is shown with an application referring to vote proportion for the Brazilian presidential elections runoff in 2018. The proportion of votes is an important measure in analyzing electoral data. Since it is a variable limited to the unitary interval, unit distributions should be considered to analyze its probabilistic behavior. Thus, the introduced model is suitable for describing the characteristics detected in these data, such as the asymmetric behavior, bimodality, and the unit interval as support. In the application, the superiority of the proposed model is verified when comparing the fit with the two-component beta mixture models.

Keywords: Brazilian elections, EM algorithm, mixture distributions, unit models, unit Weibull.

MSC: 60E05; 60E10; 62E10

Citation: Guerra, R.R.; Peña-Ramírez, F.A.; Mafalda, C.P.; Cordeiro, G.M Two-component unit Weibull mixture model to analyze vote proportions. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

Finite mixture models appeared in a study on the asymmetry of grouped materials not being homogeneous [1], being useful in the presence of multimodality, heavy tails, and asymmetry [2]. Many works have appeared in the literature in the context of finite mixtures. For example, [3] proposed a model for exponential mixtures. Considering Weibull mixture models, we can cite [4] for characterizations of the failure rate function and [5] for reliability approximations. Recently, [6] analyzed individual periods in combined sea waves using parametric mixture models.

In data with limited support, beta mixture models have been studied by several authors. [7] proposed a study on the beta mixture to solve problems related to correlations of gene expression levels, [8] presented a study on Bayesian analysis, and [9] studied beta mixture in regression models. The Kumaraswamy mixture model is an alternative to the beta mixture models. [10] carried out a Bayesian study on the three-component Kumaraswamy mixture.

In this paper, a new two-component mixture model is proposed as an alternative to model population heterogeneities in the unit support. We consider that each mixture component follows a unit Weibull (UW) distribution [11]. Some of the contributions of this new distribution, the so-called Weibull mixture model of the two-component unit (UWUW), are: i) all estimation routines, including simulations and applications, are performed using the expectation-maximization (EM) algorithm, and ii) applicability for electoral data modeling. The EM algorithm is a computational method used to calculate the maximum

likelihood estimator (MLE) iteratively [12]. It is widely used to estimate the maximum probability for finite mixture models [13]. Finally, the adjustment to electoral data, defined as the district's share of votes by the total number of valid votes cast in the district, the proportions of votes are useful, since the electoral districts can vary considerably in the size of the population [14]. Also, this measure can analyze other characteristics of the electoral process, such as electoral volatility [15] and nationalization of electoral change [14]. The data set used refers to the proportions of votes in the Brazilian presidential elections runoff in 2018.

The rest of the work is organized as follows. In Section 2, the new mixture model is presented. Section 3 introduces the EM algorithm to perform maximum likelihood estimation for the UWUW model. In Section 4, an application is made with electoral data. The final considerations of this work are addressed in Section 5.

2. The proposed model

In this section, the two-component unit Weibull mixture distribution, so-denoted UWUW, is introduced. Let X be a random variable with UWUW distribution. Then, its cdf is obtained as

$$\begin{aligned} F_{\text{UWUW}}(x;\boldsymbol{\Theta}) &= p\, F_{UW}(x;\boldsymbol{\theta}_1) + (1-p) F_{UW}(x;\boldsymbol{\theta}_2) \\ &= p\, \tau^{[-\log x/\log \mu_1]^{\beta_1}} + (1-p)\, \tau^{[-\log x/\log \mu_2]^{\beta_2}}, \end{aligned}$$

where $\boldsymbol{\theta}_1 = (\mu_1,\beta_1)^\top$, $\boldsymbol{\theta}_2 = (\mu_2,\beta_2)^\top$, μ_1 and $\mu_2 \in (0,1)$ are location parameters associated with the τth quantiles of each component of the mixture, β_1 and $\beta_2 > 0$ are shape parameters, and $\tau \in (0,1)$ is assumed to be known. One can note we use a parameterization based on quantiles to formulate each component of the mixture. The advantage of working with reparametrization in terms of quantiles is its flexibility to model data with heterogeneous conditional distributions [16,17]. The UWUW probability density function (pdf) is given by

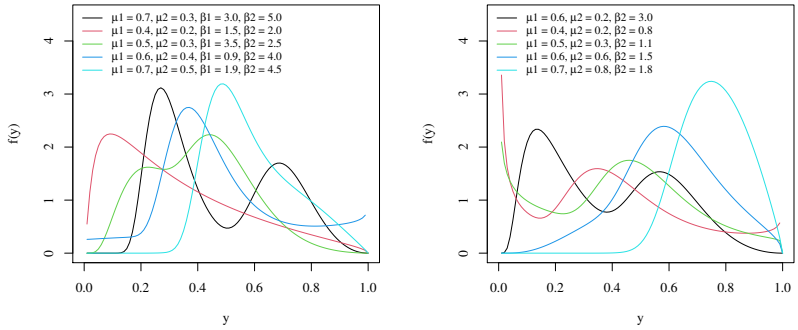
$$\begin{aligned} f_{\text{UWUW}}(x;\boldsymbol{\Theta}) &= p\, f_{\text{UW}}(x;\boldsymbol{\theta}_1) + (1-p) f_{\text{UW}}(x;\boldsymbol{\theta}_2) \\ &= p\, \frac{\beta_1 \log \tau}{x \log \mu_1} \left(\frac{\log x}{\log \mu_1}\right)^{\beta_1 - 1} \tau^{(\log x/\log \mu_1)^{\beta_1}} \\ &+ (1-p)\, \frac{\beta_2 \log \tau}{x \log \mu_2} \left(\frac{\log x}{\log \mu_2}\right)^{\beta_2 - 1} \tau^{(\log x/\log \mu_2)^{\beta_2}}. \end{aligned} \tag{1}$$

Figure 1 shows some plots of the UWUW pdf for some combinations of parameters and $\tau = 0.5$, which reveals the high flexibility of the new distribution. It accommodates bimodal, unimodal, descending, and bath forms under different asymmetric characteristics. Also, it is possible to identify a bimodal form for different values of p. Hereafter, we denote X as a random variable following a UWUW distribution, this is, $X \sim \text{UWUW}(\boldsymbol{\Theta})$.

3. Parameter estimation

An approach to the iterative computation of MLEs when the observations can be treated as incomplete data is the well-known expectation-maximization (EM) algorithm. Considering the context of two-component mixture models, let $\boldsymbol{x} = \{x_1,\ldots,x_n\}$ be a random sample of size n from a random variable X having pdf (4) with unknown parameter vector $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, p)^\top$, where $\boldsymbol{\theta}_1 = (\mu_1,\beta_1)^\top$ and $\boldsymbol{\theta}_2 = (\mu_2,\beta_2)^\top$. It is customary to call $\boldsymbol{x}$ of "incomplete data" since it is associated with a second component $\boldsymbol{z} = \{z_1,\ldots,z_n\}$ of unobserved values of a latent random variable Z. Each value z_i of Z indicates which component of the mixture belongs to the ith observation x_i such that

$$z_i = \begin{cases} 1 & \text{if } x_i \text{ has pdf } f_{\text{UW}}(x|\boldsymbol{\theta_1}), \\ 0 & \text{if } x_i \text{ has pdf } f_{\text{UW}}(x|\boldsymbol{\theta_2}), \end{cases}$$



(**a**) For $p = 0.4$. (**b**) For $p = 0.6$ and $\beta_1 = 3.0$.

Figure 1. Plots of the UWUW density for some parameter values.

where $P(Z = 1) = p$ and $P(Z = 0) = 1 - p$. The complete-data specification is determined by the joint density of (X, Z)

$$f_{X,Z}(x_i, z_i; \boldsymbol{\Theta}) = \left[p \, \frac{\beta_1 \log \tau}{x_i \log \mu_1} \left(\frac{\log x_i}{\log \mu_1} \right)^{\beta_1 - 1} \tau^{(\log x_i / \log \mu_1)^{\beta_1}} \right]^{z_i} \times \left[(1-p) \, \frac{\beta_2 \log \tau}{x \log \mu_2} \left(\frac{\log x}{\log \mu_2} \right)^{\beta_2 - 1} \tau^{(\log x / \log \mu_2)^{\beta_2}} \right]^{1 - z_i},$$

and based on it, the complete log-likelihood function, for the sample of size n, is given by

$$\begin{aligned} l_c(\boldsymbol{\Theta}) &= \sum_{i=1}^{n} \log f_{X,Z}(x_i, z_i; \boldsymbol{\Theta}) \\ &= \sum_{i=1}^{n} z_i \log \left[p \, \frac{\beta_1 \log \tau}{x_i \log \mu_1} \left(\frac{\log x_i}{\log \mu_1} \right)^{\beta_1 - 1} \tau^{(\log x_i / \log \mu_1)^{\beta_1}} \right] \\ &\quad + \sum_{i=1}^{n} (1 - z_i) \log \left[(1-p) \, \frac{\beta_2 \log \tau}{x \log \mu_2} \left(\frac{\log x}{\log \mu_2} \right)^{\beta_2 - 1} \tau^{(\log x / \log \mu_2)^{\beta_2}} \right]. \end{aligned} \tag{2}$$

The EM algorithm iterates, between two steps, to compute the MLEs of $\boldsymbol{\Theta}$. In the E-step or expectation step, due to (2) is unobservable, it is replaced by its conditional expectation with respect to the conditional distribution of Z, given $\boldsymbol{x}$ and the current parameter estimates. More specifically, in the $(k+1)$th iteration, the E-step computes

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k)}) &= E_{\boldsymbol{\Theta}^{(k)}}[l_c(\boldsymbol{\Theta}) | \boldsymbol{x}] \\ &= \sum_{i=1}^{n} \log f_{X,Z}(x_i, z_i; \boldsymbol{\Theta}) \\ &= \sum_{i=1}^{n} \bar{z}_{i1} \, \log \left[p \, \frac{\beta_1 \log \tau}{x_i \log \mu_1} \left(\frac{\log x_i}{\log \mu_1} \right)^{\beta_1 - 1} \tau^{(\log x_i / \log \mu_1)^{\beta_1}} \right] \\ &\quad + \sum_{i=1}^{n} \bar{z}_{i2} \, \log \left[(1-p) \, \frac{\beta_2 \log \tau}{x \log \mu_2} \left(\frac{\log x}{\log \mu_2} \right)^{\beta_2 - 1} \tau^{(\log x / \log \mu_2)^{\beta_2}} \right], \end{aligned} \tag{3}$$

where

$$\bar{z}_{i1} = \frac{p^{(k)} f_{UW}(x; \boldsymbol{\theta}_1^{(k)})}{p^{(k)} f_{UW}(x; \boldsymbol{\theta}_1^{(k)}) + (1 - p^{(k)}) f_{UW}(x; \boldsymbol{\theta}_2^{(k)})},$$

$$\bar{z}_{i2} = \frac{(1 - p^{(k)}) f_{UW}(x; \boldsymbol{\theta}_2^{(k)})}{p^{(k)} f_{UW}(x; \boldsymbol{\theta}_1^{(k)}) + (1 - p^{(k)}) f_{UW}(x; \boldsymbol{\theta}_2^{(k)})},$$

and $\boldsymbol{\Theta}^{(k)} = (\boldsymbol{\theta}_1^{(k)}, \boldsymbol{\theta}_2^{(k)}, p^{(k)})^\top$ are obtained from the kth iteration.

The M-step or maximization step, requires the maximization of (3) with respect to $\boldsymbol{\Theta}$. This is

$$\boldsymbol{\Theta}^{(k+1)} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(k)}). \tag{4}$$

The vector $\boldsymbol{\Theta}^{(k+1)}$ is used to initialize the next iteration. Thus, the EM algorithm is initialized by the starting values $\boldsymbol{\Theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)}, p^{(0)})^\top$ and the MLEs $\hat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$ are obtained by $\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(k+1)}$ when a convergence criterion $|\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Theta}^{(k)}| < \varepsilon$ is reached [12]. We set $\varepsilon = 10,000$. It should be noted that it is not possible to obtain analytical results from these expressions. It is necessary to perform this maximization by applying some iterative techniques, for example, Newton Raphson's method [18].

4. Application

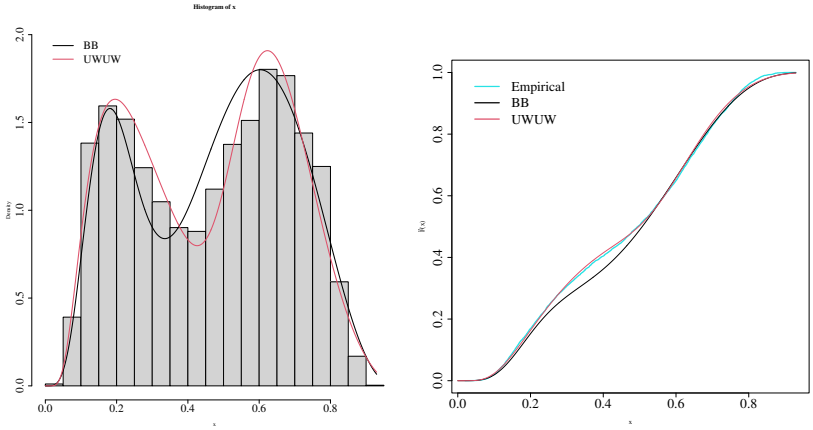
In what follows, we present a case study that illustrates the suitability of the UWUW distribution for modeling real unit data sets. The database considered is the municipality's vote proportion of the winning candidate in the Brazilian presidential elections runoff in 2018. Since it presents a bimodal shape, see Figure 2a, a unimodal distribution would not be appropriate to fit this data set. Therefore, the UWUW distribution is a suitable alternative to model these data. Its performance is compared with other double-bounded component mixtures that have already been studied in the literature: two-component beta mixture (BB) model. In this paper, the parameterization proposed by [19] is considered to define the BB model, which has pdf given by

$$f(x; \boldsymbol{\Theta}) = p \frac{\Gamma(\mu_1 + \beta_1)}{\Gamma(\mu_1)\Gamma(\beta_1)} x^{\mu_1 - 1}(1 - x)^{\beta_1 - 1} + (1 - p) \frac{\Gamma(\mu_2 + \beta_2)}{\Gamma(\mu_2)\Gamma(\beta_2)} x^{\mu_2 - 1}(1 - x)^{\beta_2 - 1}, \quad 0 < x < 1,$$

where $\boldsymbol{\Theta} = (\mu_1, \mu_2, \beta_1, \beta_2, p)^\top$, μ_1 and $\mu_2 \in (0, 1)$ are location parameters associated with the mean of each mixture component, β_1 and $\beta_2 > 0$ are precision parameters, and $p \in (0, 1)$ is the parameter that measures the weights of the mixture.

For all competitive mixture models, the parameter estimation is carried out using the EM algorithm following the steps described in Section 3. The Corrected Anderson-Darling (A^*) [20], Cramér-von Misses (W^*) [21], and the Kolmogorov Smirnov (KS) [22] statistics are calculated to assess the quality-of-fit for the three fitted models. The lower their values are, the better is the model fit. All the analysis is performed using the `R` programming language, and the goodness-of-fit measures are computed using the `AdequacyModel` [23] subroutine.

Table 1 displays the parameter estimates, standard errors, and the model comparison criteria of the three considered models. The results indicate that the UWUW distribution provides the lowest values for all goodness-of-fit statistics. The KWKW presents the worse performance, not being an adequate alternative to fit these data.



(**a**) (**b**)

Figure 2. Estimated densities (a) and empirical cdf (b) of the BB, KWKW and UWUW models.

Table 1: Parameter estimates and standard errors (given in parentheses) for the models fitted to Bolsonaro's vote proportion in Brazilian presidential elections in 2018.

	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\beta_1}$	$\hat{\beta_1}$	$\hat{p}$	W^*	A^*	KS
BB	0.5816	0.1985	9.7510	29.3260	0.7268	1.2937	7.4584	0.0477
	(0.0035)	(0.0026)	(0.3201)	(1.3521)	-			
UWUW	0.2677	0.6491	2.7011	2.9611	0.5368	**0.4119**	**3.6768**	**0.0153**
	(0.0039)	(0.0027)	(0.0545)	(0.0567)	-			

Figure 2a presents the histogram of the vote proportion data overlaid with the estimated densities of the fitted models. The bimodality of the data is confirmed, and the UWUW model provides the closest fit to the histogram. Clearly, the KWKW model is not adequate to fit these data. Further, Figure 2b gives plots of the empirical and estimated cdfs. This visual inspection favors the results in Figure 2a and Table 1, indicating that the proposed model is appropriate to fit these data. Thus, it can be an effective alternative to analyze vote proportions, being quite competitive with the BB model and providing consistently better fits than the KWKW model. Therefore, the UWUW provides a useful tool for modeling bimodal data restricted to the unit interval. Also, with the estimates of the mixture parameters, it is possible to identify that more than 50 % of the observations belong to the first mixture component. The estimated median of the first component is $\hat{\mu}_1 = 0.2677$ and the estimated median of the second component is $\hat{\mu}_2 = 0.6649$.

5. Conclusion

A two-component mixture model was defined to describe the heterogeneities of the population with the limited domain. The two-component unit Weibull mixture (UWUW) model is formulated considering that each mixture component follows the unit Weibull distribution. Some of the main properties of UWUW have been presented, such as ordinary moments. The EM algorithm was used to obtain maximum likelihood estimates for the model parameters. To evaluate the performance of the EM algorithm, Monte Carlo simulations were performed. An application to electoral data illustrates the importance and potential of the new model. The motivating data set is about the vote proportions obtained by the winning candidate in the Brazilian presidential runoff elections in 2018. The results indicate that our proposal is adequate to fit this data set since it is suitable to analyze the asymmetric and bimodal behaviors. From the mixing parameter estimate, we can conclude that 53.68% of the observations are from the first component of the mixture with estimated median at $\hat{\mu}_1 = 0.2677$. The estimated median for the municipalites from the second mixture compontent was $\hat{\mu}_2 = 0.6491$. This application proved empirically that

the UWUW performance may overcome other two-component mixture models based on other widely known unit distributions such as the beta and Kumaraswamy.

References

1. Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London, A.* **1894**, *185*, 71–110.
2. Lachos, V.H.; Moreno, E.J.L.; Chen, K.; Cabral, C.R.B. Finite mixture modeling of censored data using the multivariate Student-t distribution. *Journal of Multivariate Analysis* **2017**, *159*, 151–167.
3. Jewell, N.P. Mixtures of exponential distributions. *The Annals of Statistics* **1982**, *10*, 479–484.
4. Jiang, R.; Murthy, D. Mixture of Weibull distributions-parametric characterization of failure rate function. *Applied Stochastic Models and Data Analysis* **1998**, *14*, 47–65.
5. Bučar, T.; Nagode, M.; Fajdiga, M. Reliability approximation using finite Weibull mixture distributions. *Reliability Engineering & System Safety* **2004**, *84*, 241–251.
6. Huang, W.; Dong, S. Probability distribution of wave periods in combined sea states with finite mixture models. *Applied Ocean Research* **2019**, *92*, 101938.
7. Ji, Y.; Wu, C.; Liu, P.; Wang, J.; Coombes, K.R. Applications of beta-mixture models in bioinformatics. *Bioinformatics* **2005**, *21*, 2118–2122.
8. Bouguila, N.; Ziou, D.; Monga, E. Practical Bayesian estimation of a finite beta mixture through Gibbs sampling and its applications. *Statistics and Computing* **2006**, *16*, 215–225.
9. Grün, B.; Kosmidis, I.; Zeileis, A. Extended beta regression in R: shaken, stirred, mixed, and partitioned. Technical report, Working Papers in Economics and Statistics, 2011.
10. Khalid, M.; Aslam, M.; Sindhu, T.N. Bayesian analysis of 3-components Kumaraswamy mixture model: Quadrature method vs. Importance sampling. *Alexandria Engineering Journal* **2020**, *59*, 2753–2763.
11. Mazucheli, J.; Menezes, A.; Ghitany, M. The unit-Weibull distribution and associated inference. *Journal of Applied Probability and Statistics* **2018**, *13*, 1–22.
12. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **1977**, *39*, 1–22.
13. Redner, R.A.; Walker, H.F. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **1984**, *26*, 195–239.
14. Alemán, E.; Kellam, M. The nationalization of presidential elections in the Americas. *Electoral Studies* **2017**, *47*, 125–135.
15. Powell, E.N.; Tucker, J.A. Revisiting electoral volatility in post-communist countries: new data, new results and new approaches. *British Journal of Political Science* **2013**, *44*, 123–147.
16. Bayes, C.L.; Bazán, J.L.; De Castro, M. A quantile parametric mixed regression model for bounded response variables. *Statistics and its Interface* **2017**, *10*, 483–493.
17. Mazucheli, J.; Menezes, A.F.B.; Fernandes, L.B.; de Oliveira, R.P.; Ghitany, M.E. The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates. *Journal of Applied Statistics* **2020**, *47*, 954–974.
18. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical recipes 3rd edition: The art of scientific computing*; Cambridge University Press, 2007.
19. Ferrari, S.L.P.; Cribari-Neto, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **2004**, *7*, 799–815.
20. Chen, G.; Balakrishnan, N. A general purpose approximate goodness-of-fit test. *Journal of Quality Technology* **1995**, *27*, 154–161.
21. Durbin, J.; Knott, M. Components of Cramér-Von Mises statistics. *Journal of the Royal Statistical Society: Series B (Methodological)* **1972**, *34*, 290–307.
22. Goodman, L.A. Kolmogorov-Smirnov tests for psychological research. *Psychological Bulletin* **1954**, *51*, 160.
23. Marinho, P.R.D.; Silva, R.B.; Bourguignon, M.; Cordeiro, G.M.; Nadarajah, S. AdequacyModel: An R package for probability distributions and general purpose optimization. *PloS one* **2019**, *14*, e0221487.