Gaussian Mixtures with common Variance

MIGUEL FELGUEIRAS [1,3,4], JOÃO MARTINS [2,3,5], RUI SANTOS [1,3]
[1] ESTG, Polytechnic Institute of Leiria, PORTUGAL
[2] ESS, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida, 4249-015 Porto, PORTUGAL
[3] CEAUL, Faculdade de Ciências, Universidade de Lisboa, PORTUGAL
[4] CIDMA, University of Aveiro, PORTUGAL [5] CEISUC/CIBB, Coimbra, PORTUGAL

Abstract: The interest in Gaussian mixtures has grown significantly in recent years, primarily owing to their adaptability and widespread applications across various fields of knowledge. A specific category within these mixtures is Gaussian mixtures with common variance, wherein the assumption is made that the variances of all subpopulations are equal. This study delves Gaussian location mixtures family, exploring their applications, characterizations, and the challenges associated with estimation. Following this, we introduce an approximation to the beta distribution. When addressing scenarios involving two subpopulations, a novel test for equality of variances is proposed, employing the beta distribution approximation. This paper presents a new test for variance equality which is a novelty in the Gaussian mixture context. Practical applications for the proposed test are provided and discussed.

Key-Words: Gaussian location mixtures, beta distribution, variance equality test.

Received: November 11, 2023. Revised: March 9, 2024. Accepted: March 21, 2024. Published: April 24, 2024.

1 Introduction

The history of Gaussian mixtures models goes as far as the nineteen century. In 1894, [1, 2] analysed a sample of crabs to determine the size of their foreheads. He concluded that not a single species of crabs, but a mixture of two crab species, was observed. In a remarkable work, Pearson used a Gaussian mixture to fit that data set. In this paper we consider Gaussian mixtures in Pearson's sense, that is, a Gaussian mixture is a convex mixture of Gaussian random variables when its density function is

$$f_X(x) = \sum_{j=1}^{N} w_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma_j}\right)^2\right\}, \tag{1}$$

where $\sigma_j > 0, w_j > 0, \sum_{j=1}^{N} w_j = 1$ and N denotes the number of Gaussian random variables, each with mean μ_j, standard deviation σ_j and weight w_j. Note that some works from other authors as [3, 4] deal with a different kind of Gaussian mixture, namely assuming that one of the Gaussian distribution parameters is a random variable, usually the scale parameter. This is a type of infinite Gaussian mixture that will not be tackled in this work. Independently of the type of the considered mixture, all kind of mixtures are very effective when fitting real data since they can accommodate multimodality and a wide range of density shapes. For example, [5] uses deep Gaussian mixture models to describe data in a very flexible way, since at each layer the variables follow a mixture of Gaussian distributions. In a machine learning approach for communications, [6] applies Gaussian mixtures to channel estimation. However, previous examples have a major counter back: a large number of parameters must be estimated. The increase of computational power throughout the last decades allowed the software implement of the expectation-maximization algorithm (EM) [7], used to numerically estimate the parameters, despite of some convergence constraints [8].

In this context, variance equality is an important theme since inference procedures are usually more simple and accurate under that assumption. Moreover, the previously indicated estimation issues become less relevant since the number of parameters to estimate diminishes. From a practical point of view, it is also relevant to decide whenever subpopulations variances can be considered as equal. Hence, in this work we deal with Gaussian mixtures with common variances, presenting some results under that assumption. Furthermore, a variance equality test is developed.

2 Moments and Miscellaneous for Gaussian Mixtures

Let us consider that a random variable X is a Gaussian mixture with density as defined in Equation 1. Therefore, moments can be obtained from cumulant

generation function,

$$\ln[\varphi_X(-it)] = \kappa_1 it - \kappa_2 \frac{t^2}{2!} - \kappa_3 i \frac{t^3}{3!} + \kappa_4 \frac{t^4}{4!} + O\left(t^5\right) \tag{2}$$

with

$$\kappa_1 = \mu_1'; \;\; \kappa_2 = \mu_{(2)}; \;\; \kappa_3 = \mu_{(3)}; \;\; \kappa_4 = \mu_{(4)} - 3\mu_{(2)}^2, \tag{3}$$

where $\mu_{(k)}$ stands for the k-th centered moment, μ_k' denotes the k-th raw moment and φ_X is the characteristic function.

Two standard simplifications can be considered. One corresponds to mean equality, that is, $\mu_j = \mu$ for $j = 1, ..., N$. Under mean equality, the mixture can be approximated to the t-Student distribution. Moreover, t-Student distribution can be used to evaluate the equality of means, that is, to test [9]

$$H_0 : \mu_1 = \mu_2 = ... = \mu_N.$$

The other standard simplification, that we will deal with in this work, is to consider $\sigma_j^2 = \sigma^2$ for $j = 1, ..., N$.

Theorem 1. *Let X be a Gaussian mixture where all the subpopulations have equal variance σ^2. Then*

$$X \overset{d}{=} V + Y$$

where V and Y are independent random variables, $V \sim N(0, \sigma)$ and Y is such that $P(Y = \mu_j) = w_j$, for $j = 1, ..., N$.

Proof. Recall that for any independent random variables V and Y we have $\varphi_{V+Y}(t) = \varphi_V(t)\,\varphi_Y(t)$ and that when $V \sim N(0, \sigma)$ then $\varphi_V(t) = \exp\left(-\frac{t^2\sigma^2}{2}\right)$. Consequently, the characteristic function of the sum of the independent variables V and Y defined above is

$$\begin{aligned} \varphi_{V+Y}(t) &= \varphi_V(t)\,\varphi_Y(t) = \\ &= \exp\left(-\frac{t^2\sigma^2}{2}\right)\left[\sum_{j=1}^{N} w_j \exp(it\mu_j)\right] = \\ &= \sum_{j=1}^{N} w_j \exp\left\{it\mu_j - \frac{t^2\sigma^2}{2}\right\} = \varphi_X(t), \end{aligned}$$

and by the uniqueness of the characteristic function we obtain $X \overset{d}{=} V + Y$. □

Thus, when all the Gaussian subpopulations share the same variance, the mixture can be seen as a convolution between a Gaussian noise and a discrete random variable [10]. This would take us to deconvolution problems, often study in statistics [11, 12] but beyond the scope of the present paper. The above convolution appears in many known applications. Previous work under amphibian nervous system [13, 14] concluded that the junction between primary afferent fibre and motoneurone provides joint electrical and chemical transmission. The mixed synapse can be fitted by binomial or Poisson convolutions with a Gaussian noise. In image or signal processing, convolutions between Poisson and zero mean Gaussian are also used. For example, [15] refers that astronomical images have additive uncorrelated noise. Poisson noise, due to photon arrival events, and Gaussian white noise, due to commonly used digitized photographic plates.

Unimodality and multimodality are always possible, according with different combinations of parameters [16]. If the mixture has an unimodal density function, it can be approximated to the Pearson system, according to its β_1 and β_2 values [17], where β_1 and β_2

$$\beta_1 = \frac{\mu_{(3)}}{\mu_{(2)}^{3/2}}; \qquad \beta_2 = \frac{\mu_{(4)}}{\mu_{(2)}^2} \tag{4}$$

are the skewness and the kurtosis coefficients. For Pearson type I distribution (four parameters beta), the approximation holds when

$$1.5\beta_1^2 < \beta_2 < 1.5\beta_1^2 + 3. \tag{5}$$

3 Two Subpopulations

The main goal of this work is to develop a variance equality test for Gaussian mixtures, considering only two subpopulations as the starting point. A sufficient condition for unimodality, independent from the values of w_1 and w_2 is given by [18]

$$|\mu_1 - \mu_2| \leq 2\min(\sigma_1, \sigma_2). \tag{6}$$

We will now assume that the previous condition holds. Nevertheless, in practical issues, it is often complicated to distinguish if multimodality is due to the model or to a particular sample issue [19]. When the subpopulations share the same variance, that is when $\sigma_1^2 = \sigma_2^2 = \sigma^2$, it is clear that $w_1 = w$ and $w_2 = 1 - w$. Under these circumstances, and for a wide range of values of w, the mixture can now be approximated to a beta distribution [9, 10], using equation 5.

Theorem 2. *Let X be a finite unimodal Gaussian mixture with two components with equal variance. If*

$$w \in \left[\frac{1}{2} \pm \frac{\sqrt{3}}{6}\right],$$

then the mixture can be approximated to a beta distribution.

Note that the situation where $\mu_1 = \mu_2$ (leading to a single gaussian) corresponds to the case where the w interval is tighter. In any other scenario we obtain a wider interval that contains the presented one. For example, if $|\mu_1 - \mu_2| = \sigma$ we get $w \in [0.1899; 0.8101]$. As previously stated, Theorem 2 holds for most unimodal mixtures with common variance. Mixtures with $w \notin [0.2113; 0.7887]$ correspond, roughly, to the contaminated populations problem, where a population has a few elements that do not belong to it [20]. This is also an interesting problem, for example, when dealing with infected or non infected elements from a population with some disease.

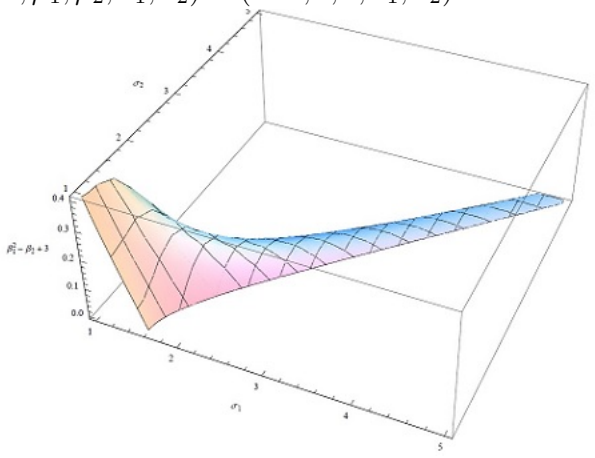
4 Testing Variance Equality

As previously stated, the mixture can be approximated by a beta distribution when $\sigma_1^2 = \sigma_2^2$, that is $X \overset{\circ}{\sim} beta(a, b, p, q)$ or

$$Y = \frac{X - a}{b - a} \overset{\circ}{\sim} beta(p, q).$$

Unfortunately, in some situations when the variances are different the approximation is still possible, even theoretically (see example in Figure 1).

Figure 1: Region where condition (5) holds for $(w, \mu_1, \mu_2, \sigma_1, \sigma_2) = (0.35, 0, 2, \sigma_1, \sigma_2)$



Hence, when testing H_0: data follows a beta distribution versus H_1: data does not follow a beta distribution, the rejection of H_0 implies that $\sigma_1^2 \neq \sigma_2^2$, but if H_0 is not rejected then the variances may or may not be equal. Even though, σ_1^2 and σ_2^2 should be at least close. Therefore, this test can be used to indirectly test $H_0 : \sigma_1^2 = \sigma_1^2 \quad vs \quad H_1 : \sigma_1^2 \neq \sigma_2^2$ or, in a equivalent way, to test $H_0 : \sigma_1 = \sigma_1 \quad vs \quad H_1 : \sigma_1 \neq \sigma_2$.

All the four parameters can be simultaneous estimated, numerically, by the maximum likelihood method [21]. This method is already implemented in some software, like the R package *ExDist* based on [22] and [23] work. However, [21] states that good results can only be achieved for large samples, since convergence to a global maximum is not guaranteed. Alternatively, straightforward estimators for a and b, based on the sample minimum and maximum $(\min X_i, \max X_i)$ are

$$\widehat{a} = \min X_i - \frac{\max X_i - \min X_i}{n}$$

$$\widehat{b} = \max X_i + \frac{\max X_i - \min X_i}{n},$$

and then the moment estimators can be defined as [23]

$$\widehat{p} = \frac{\left(\frac{\overline{X}-a}{b-a}\right)^2 \left(1 - \frac{\overline{X}-a}{b-a}\right)}{\frac{S^2}{(b-a)^2}} - \frac{\overline{X} - a}{b - a}$$

$$\widehat{q} = \frac{\left(\frac{\overline{X}-a}{b-a}\right) \left(1 - \frac{\overline{X}-a}{b-a}\right)}{\frac{S^2}{(b-a)^2}} - 1 - \widehat{p},$$

where as usual $\overline{X}$ and S^2 represent sample mean and sample variance, respectively.

5 Applications

In this section we apply the test to three real data sets, in order to understand if the results can be applied to practical situations. For all the analysed data sets, the unknown parameter vector $(\mu_1, \mu_2, \sigma_1, \sigma_2, w)$, where the parameter w corresponds to the first component weight, was estimated by the EM algorithm using $MatlabR2013b$. To compare models with common and different variances, the Bayesian information criterion (BIC) [24] was also computed. This is an information criterion that penalizes more severely over fitting than the most used Akaike information criterion. Smaller values of BIC are obtained for better fitted models.

5.1 Applications to financial data

There are several applications concerning economical data linked with Gaussian mixtures. In a very recent work, [25] uses Gaussian mixture returns for portfolio construction. In this paper we consider the model of daily log-returns, a well known problem in finance. Log-returns are defined as $x_t = \ln(X_t) - \ln(X_{t-1})$, where X_t represents the close index value of the t-th day. Previous works like [26, 27, 28] suggest a wide set of possible models, but Gaussian mixtures (with a small number of components, preferably only two, to avoid over-fitting) are a common choice. Let us consider the daily log-returns from the PSI20 stock index. The data set comprehends the time gap between 2012/03/16 and 2017/03/17, roughly five years, and a

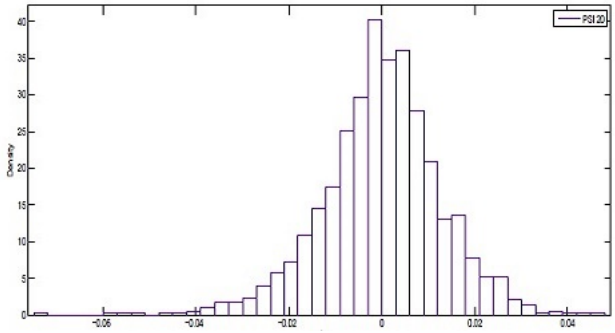
Table 1: Estimated Gaussian mixture for the PSI20 data set.

$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\sigma}_1$	$\widehat{\sigma}_2$	$\widehat{w}$
0.0011	−0.0024	0.0091	0.0176	0.6417

total of 1278 observations. Parameters estimates are displayed in Table 1.

When analysing the histogram presented in Figure 2, the data can be considered as unimodal.

Figure 2: Histogram for PSI20 log-returns data.



When testing $H_0 : \sigma_1^2 = \sigma_2^2 \; vs \; H_1 : \sigma_1^2 \neq \sigma_2^2$, we obtain a p-value $= 0.0110$ and accordingly (remember that the test is quite conservative) we should reject H_0 and conclude for different variances. The *BIC* measure, which greatly penalizes over fitting is $BIC = -7515.6568$. For a Gaussian mixture with the same variance we get $BIC = -7510.9287$ and, as expected, the model with different variances yields better results.

Next we present a similar example, for the daily log-returns from the SP500 stock index. The data set comprehends the same time gap but with a total of 1259 observations. The estimates are presented in Table 2.

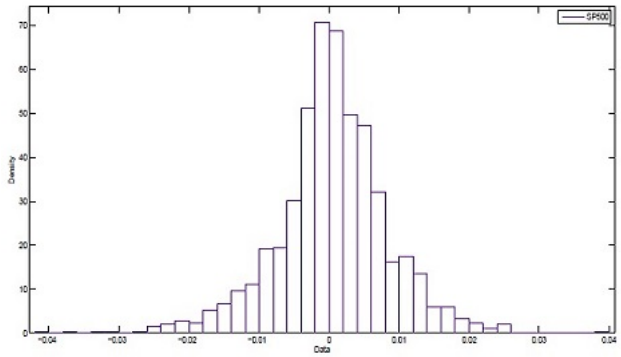
Table 2: Estimated Gaussian mixture for the SP500 data set.

$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\sigma}_1$	$\widehat{\sigma}_2$	$\widehat{w}$
0.0007	−0.0001	0.0039	0.0104	0.5263

The data is clearly unimodal, as putted in evidence by the histogram in Figure 3.

When testing $H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_1 : \sigma_1^2 \neq \sigma_2^2$, we obtain a p-value $< 10^{-4}$ and therefore we should reject H_0 and conclude for different variances. The BIC measure value when $\sigma_1^2 \neq \sigma_2^2$ is $BIC = -8654.4748$ and when $\sigma_1^2 = \sigma_2^2$ we get $BIC = -8545.2453$. Again, the model with different variances yields better results.

Figure 3: Histogram for SP500 log-returns data.



5.2 Application to Biometric Data

Let us consider the *Davis* dataset, available in the R package "car". It contains measured heights and weights of 200 adults, men (112) and women (88), engaged in regular exercise. Note that in line 12, weight and height were switched as they appear to be reversed in the original dataset. The descriptive statistics for height (in centimeters) obtained from the dataset are presented in Table 3.

Table 3: Descriptive statistics for the variable Height in the "Davis" dataset.

	Male	Female
Mean	178.0114	164.7143
Standard Deviation	6.4407	5.6591
Proportion	0.44	0.56

Firstly we tested the normality of the variable Height for both subsets (male and female) with Lilliefors normality test [24]. For the male and females subsets we obtained, respectively, p-value $= 0.8720$ and p-value $= 0.1818$.

When testing $H_0 : \sigma_M^2 = \sigma_F^2 \quad vs \quad H_1 : \sigma_M^2 \neq \sigma_F^2$ with F test, we obtain p-value $= 0.1979$. As a consequence, variances should be considered as equal for both sex.

For illustrative purpose, we will be considering for now on that the data is "mixed", that is, that we do not know if an individual is male or female. The histogram presented in Figure 4 shows an unimodal data set

If we fit a two component Gaussian mixture to the dataset, we get as estimates for the model components (Table 4).

Figure 4: Histogram for the "Davis" height data.

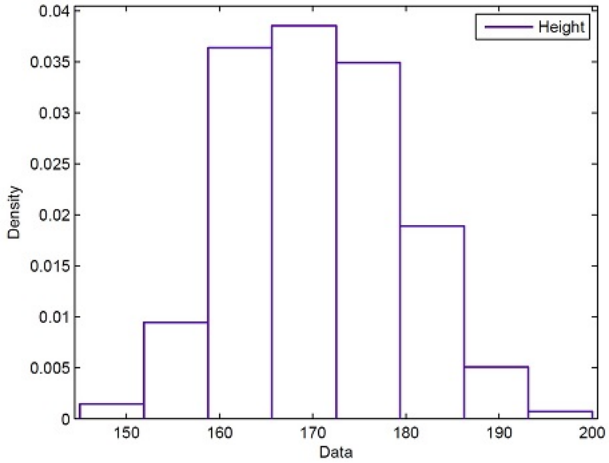


Table 4: Estimated Gaussian mixture for the "Davis" height dataset.

$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\sigma}_1$	$\widehat{\sigma}_2$	$\widehat{w}$
177.6408	164.7896	6.7108	5.7621	0.4494

When testing $H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_1 : \sigma_1^2 \neq \sigma_2^2$, using the K-S test for the beta distribution we obtain a p-value $= 0.5638$ and consequently we should not reject H_0. Therefore, evidence to reject variance equality was not found, and we can not conclude for unequal variances. The BIC statistics is $BIC = 1461.1$ if $\sigma_1^2 \neq \sigma_2^2$ and $BIC = 1457.1$ if $\sigma_1^2 = \sigma_2^2$. Thence, the model with common variance for both components yields better results. This result corroborate the obtained for the F test, that is, variances should not be considered as different.

6 Conclusion

Finite Gaussian mixtures with the same variance can be written has the convolution between a discrete variable and a zero mean Gaussian variable. It might be possible to decompose the mixture in this kind of convolution, if we have an idea about the discrete variable that is present. As stated, common examples concern Poisson or binomial data as the discrete variable, added with a Gaussian white noise.

For unimodal mixtures, Theorem 2 allows us to approximate the mixture to a beta distribution, when some conditions are fulfilled. This approximation reduce the number of unknown parameters from $2N$ to four, which can be interesting when working with a large number of subpopulations. Besides, beta distribution characterizations become available.

Finally, when only two subpopulations are considered, Theorem 2 can be used to test variance equality in unimodal mixtures. The test was applied to three different data sets with good results.

Together with the mean equality test presented in [9], the variance equality test for Gaussian mixtures can be very useful to deal with real data sets, since mean and variance equality are some of the most common hypothesis in statistics and, as far as we know, this variance equality test was not yet available for Gaussian mixtures.

References:

[1] Pearson, K (1894). Contributions to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society of London A*, 185, 71–110.

[2] Pearson, K (1895). Contributions to the mathematical theory of evolution. II. Skew variations in homogeneous material, *Philosophical Transactions of the Royal Society of London A*, 186, 343–414.

[3] Andrews, D; Mallows, C (1974). Scale Mixtures of Normal Distributions, *Journal of the Royal Statistical Society B*, 36, 1, 99–102.

[4] Bakirov, N; Székely, G (2006). Student's t-test for Gaussian scale mixtures, *Journal of Mathematical Sciences*, 139, 3, 6497–6505.

[5] Viroli, C; McLachlan, GJ (2019). Deep Gaussian mixture models. *Statistics and Computing* 29, 43–51.

[6] Fesl, B; Joham, M; Hu, S; Koller, M; Turan, N, Utschick, W (2022). Channel Estimation based on Gaussian Mixture Models with Structured Covariances. *56th Asilomar Conference on Signals, Systems, and Computers*, 533–537.

[7] Dempster, A; Laird, N; Rubin, D (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B*, 39, 1–37.

[8] Frühwirth-Schnatter, S (2006). *Finite Mixture and Markov Switching Models*, Springer, New York.

[9] Felgueiras, M; Martins, J; Santos, R (2017). Gaussian Scale Mixtures, *Journal of Numerical Analysis, Industrial and Applied Mathematics*, 11, 1–10.

[10] Felgueiras, M; Santos, R; Martins, J (2014). Some Results on Gaussian Mixtures, *AIP Conference Proceedings*, 1618, 523–526.

[11] Guerrero-Colón, J; Simoncelli, C; Portilla, J (2008). Image denoising using Mixtures of Gaussian scale mixtures, *15th IEEE International Conference on Image Processing*, 565–568.

[12] Jansson, P (1997). *Deconvolution of Images and Spectra*, Academic Press, San Diego.

[13] Grantyn, R; Shapovalov, A; Shiriaev, B (1984). Relation between structural and release parameters at frog sensory-motor synapse, *The Journal of Physiology*, 349, 459–474.

[14] Shapovalov, A; Shiriaev, B (1980). Dual mode of junctional transmission at synapses between single primary afferent fibres and motoneurones in the amphibian, *The Journal of Physiology*, 306, 1–15.

[15] Murtagh, F; Starck, J; Bijaoui, A (1995). Image restoration with noise suppression using a multiresolution support, *Astronomy and Astrophysics, Supplement Series*, 112, 179–189.

[16] Eisenberger, I (1964). Genesis of Bimodal Distributions, *Technometrics*, 6, 357–363.

[17] Johnson, N; Kotz, S; Balakrishnan, N (1994). *Continuous Univariate Distributions*, Volume I, Wiley, New York.

[18] Behboodian, J (1970). On the modes of a mixture of two normal distributions, *Technometrics*, 12, 131–139.

[19] Everitt, B; Hand, D (1981). *Finite Mixture Distributions*, Chapman & Hall, London.

[20] Karlis, D; Xekalaki, E (2003). *Mixtures Everywhere*. In Stochastic Musings: Perspectives from the Pioneers of the Late 20th Century, 78–95, Lawrence, London.

[21] Carnahan, J (1989). Maximum Likelihood Estimation for the 4-Parameter Beta Distribution, *Communications in Statistics - Simulation and Computation*, 18, 2, 513–536.

[22] Bury, K (1999). *Statistical Distributions in Engineering*, Cambridge University Press, New York.

[23] Johnson, N; Kotz, S; Balakrishnan, N (1995). *Continuous Univariate Distributions*, Volume II, Wiley, New York.

[24] Sheskin, D (2002). *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall, Boca Raton.

[25] Luxenberg, E; Boyd, S (2024). Portfolio construction with Gaussian mixture returns and exponential utility via convex optimization. Optimization and Engineering 25, 555–574.

[26] Behr, A; Pötter, U (2009). Alternatives to the normal model of stock returns: Gaussian mixture, generalized logF and generalized hyperbolic models, *Annals of Finance*, 5, 49–68.

[27] Kon, S (1984). Models of stock returns – a comparison. *Journal of Finance*, 39, 1, 147–165.

[28] Rachev, S; Mittnik, S (2000). *Stable Paretian Models in Finance*, Wiley, New York.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020. DOI: 10.54499/UIDB/00006/2020 (https://doi.org/10.54499/UIDB/00006/2020).

Conflicts of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.