USING ASYMMETRIC DISTRIBUTIONS FOR MODELING GENE EXPRESSION DATA

Walkiria Maria de Oliveira MACERAU[1] [2]
Luis Aparecido MILAN[1]

▪ ABSTRACT: We present a short review of the asymmetric distributions α-stable, skew normal, skew Student's t and skew Laplace. We compare the performance for these distributions, in general, are used to model asymmetric data, using AIC and BIC. These criterias were able to selecting the best model for each data set. We also apply these models to gene expression data and we verify these distributions are qualified to model these observations.

▪ KEYWORDS: Gene expression, asymmetric distributions, α-stable, skew Laplace, skew normal, skew Student's t.

1 Introduction

Skewed distributions have been used in modeling financial, economics, medical and genetic data sets. These distributions can adjust data with asymmetric structure.

In analysis of data with asymmetric structure it is common transformation in to the variables. However, transforming the variables may cause problems such as the difficult interpretation of the results, (AZZALINI and CAPITANIO, 1999). Also, transformations do not always eliminate completely the asymmetry.

The most used asymmetric distribution is the skew normal. This class of analytically treated distributions can model the skewness of the data and has the normal distribution as a special case.

[1] Universidade Federal de São Carlos - UFSCar, Departamento de Estatística, CEP: 13565-905, São Carlos, SP, Brasil. E-mail: *walkiriamacerau@gmail.com; dlam@ufscar.br*

[2] Universidade Estadual de Maringá - UEM, Departamento de Estatística, CEP: 87.020-900, Maringá, PR, Brasil. Email: *wmomacerau@uem.br*

The skew normal distribution was proposed by Azzalini (1985) and has received the attention from researchers as Genton *et al.* (2001), Gupta *et al.* (2004) and Arellano-Valle *et al.* (2005), among others.

The skew Student's t distribution is an extension of the skew normal distribution, Azzalini & Capitanio (2003). It has been applied to data with asymmetric structure and extreme observations. Fernández & Steel (1998) and Jones & Faddy (2003) bring examples of applications of skew Student's t distributions in financial, astronomy, biological and engineering data sets.

The skew Laplace distribution was constructed using the method described by Fernandez *et al.* (1995).

Applications of skew Laplace distribution to biological data are found in Julia & Vive-Rego (2008) and Rubio & Steel (2010).

The α-stable distribution defines a class of asymmetric distributions which was characterized by Paul Lévy by the year 1920 in his studies about the sum of random variables identically distributed, *apud* Nolan (2009). In general, this distribution does not have a closed probability density function, being defined through its characteristic function. Three particular cases are normal, Cauchy and Lévy distributions. Applications of the α-stable distribution can be found in Nolan (2003), Rachev (2003) and Rachev and Mittinik (2000) and Gonzalez *et al.* (2009) in the context of financial returns and genetic data.

The choice of a probabilistic model is a important factor in data analysis. For selection of the model the criteria most used are: the Akaike information criterion - AIC (AKAIKE, 1974), the Schwarz Bayesian criterion - BIC(SCHWARZ, 1978) and the likelihood ratio test - TRV (BOZDOGAN, 1987; WOLFINGER, 1993; LITTELL *et al.*, 2002). We develop a simulation experiment to verify the capability of the AIC and BIC to identify the best fitting distribution.

In Section 2, we make a brief descripton of the asymmetric distributions considered here. In Section 3, we describe the inferences for the parameters as well as the procedures used in the simulation. the criteria for model selection, the software and the functions used in this study. In Section 4, we report the results of applications to a set of gene expression data. Finally, in Section 5, we present the final considerations.

2 Methodology

Following we describe the α-stable, the skew normal, the skew Student's t and the skew Laplace distributions.

2.1 The α-stable distribution

The term *α-stable* refers to distributions whose sum of identically and independently distributed (i.i.d.) random variables belongs to the same family than their components. In other words, if $X_1, X_2 \ldots, X_n$ are α-stable i.i.d. random variables, then for every n belonging to the set of natural numbers,

$$X_1 + X_2 + \cdots + X_n \overset{d}{=} c_n X + d_n, \tag{1}$$

where X is also a random variable with α-stable distribution, for any constant $c_n > 0$ and $d_n \in \Re$, and the symbol $\overset{d}{=}$ indicates equality in distribution. The distribution (1) is called *strictly stable* if $d_n = 0$ for all n, Nolan (2009).

The α-stable distribution in its general case has no closed expression for the probability density function and the cumulative distribution function, being expressed by the characteristic function.

The α-stable distribution is described by four parameters $(\alpha, \beta, \gamma, \delta)$ and can be presented in two parameterizations denoted by $AE(\alpha, \beta, \gamma, \delta_0, 0)$ and $AE(\alpha, \beta, \gamma, \delta_1, 1)$.

Definition 2.1. A random variable X has distribution $AE(\alpha, \beta, \gamma, \delta_0, 0)$ if its characteristic function is

$$E\left[e^{itX}\right] = \begin{cases} e^{-\gamma^{\alpha}|t|^{\alpha}\left[1+i\beta tan\left(\frac{\pi\alpha}{2}\right)sen(t)\left(\gamma|t|^{1-\alpha}-1\right)\right]+i\delta_0 t} & \text{for } (\alpha \neq 1), \\ e^{-\gamma|t|\left[1+i\beta\frac{2}{\pi}sen(t)ln(\gamma|t|)\right]+i\delta_0 t} & \text{for } (\alpha = 1). \end{cases} \tag{2}$$

Definition 2.2. A random variable X has distribution $AE(\alpha, \beta, \gamma, \delta_1, 1)$ if its characteristic function is

$$E\left[e^{itX}\right] = \begin{cases} e^{-\gamma^{\alpha}|t|^{\alpha}\left[1-i\beta tan\left(\frac{\pi\alpha}{2}\right)sen(t)\right]+i\delta_1 t} & \text{for } (\alpha \neq 1), \\ e^{-\gamma|t|\left[1+i\beta\frac{2}{\pi}sen(t)ln|t|\right]+i\delta_1 t} & \text{for } (\alpha = 1). \end{cases} \tag{3}$$

The location parameters δ_0 and δ_1 are related, and given by

$$\delta_0 = \begin{cases} \delta_1 + \beta\gamma tan\left(\frac{\pi\alpha}{2}\right), & (\alpha \neq 1), \\ \delta_1 + \beta\frac{2}{\pi}\gamma\ ln(\gamma), & (\alpha = 1), \end{cases} \quad \delta_1 = \begin{cases} \delta_0 - \beta\gamma\ tan\left(\frac{\pi\alpha}{2}\right), & (\alpha \neq 1), \\ \delta_0 - \beta\frac{2}{\pi}\gamma\ ln(\gamma), & (\alpha = 1). \end{cases} \tag{4}$$

Here we use the parameterization $AE(\alpha, \beta, \gamma, \delta_0, 0)$.

Parameters α, β and γ have the same interpretation in both parametrizations while parameter δ has not, Nolan (2009).

The parameter α is the *stability index* or *exponent characteristic* and it defines the *local intensity level*, that is, the degree of concentration of the observations of the surrounding medium distribution; $\alpha \in (0, 2]$. The parameter β defines the asymmetry of the distribution, if $\beta = 0$ the distribution is symmetric, if $\beta > 0$ the distribution is skewed to the right, and if $\beta < 0$ the distribution is skewed to the left; $\beta \in [-1, +1]$. The parameters α and β determine the shape of the distribution. The parameter γ defines the dispersion or distribution range, $\gamma \geq 0$ and parameter δ sets the location of the distribution, $\delta \in (-\infty, +\infty)$.

If $\beta = 0$ the parameterizations coincide, when $\beta \neq 0$ and $\alpha \neq 1$ the parameterizations differ for $+$ or - $\beta\gamma \tan\left(\frac{\pi\alpha}{2}\right)$, and when $\beta \neq 0$ and $\alpha = 1$ the parameterizations differ for $+$ or - $\beta\frac{2}{\pi}\gamma \ln(\gamma)$.

$AE(\alpha, \beta, \gamma, \delta_0, 0)$ has continuous distribution function for all four parameters. The parameterization $AE(\alpha, \beta, \gamma\delta_0, 0)$ is a model belonging to a family of distributions location and scale for $\alpha \in (0, 2]$.

The mode of a random variable $X \sim AE(\alpha, \beta, \gamma, \delta_1, 1)$ tends to $+\infty$ if $[sin(\alpha - 1)\beta] > 0$ or tends to $-\infty$ if $\alpha \to 1$. The parameter $AE(\alpha, \beta, \gamma, \delta_1, 1)$ has no continuous distribution function for $\alpha = 1$ (NOLAN, 2009).

There are at least three particular cases where it is possible to write the expression of probability density function: they are the normal, Cauchy and Lévy distributions (NOLAN, 2009).

2.2 Skew Normal distribution

The skew normal distribution, introduced by Azzalini (1985), is a class of continuous probability distributions that extends the normal distribution allowing the presence of asymmetry. Its probability density function is

$$f_Z(z; \lambda) = 2\phi(z)\Phi(\lambda z), \tag{5}$$

where $\phi(z)$ is the density of a standard normal distribution and $\Phi(\lambda z)$ is the normal cumulative distribution function on λz, and $\lambda \in \Re$ is the asymmetry parameter.

The normal distribution can be recovered in Equation (5) when $\lambda = 0$. When $\lambda > 0$ the distribution is skewed to the right and when $\lambda < 0$ is skewed to the left.

Suppose a random variable Z with skew normal distribution with parameter λ, denoted by $Z \sim SN(\lambda)$. Some basic properties of the skew normal distribution given by Azzalini (1985) are

1. $SN(0) = N(0, 1)$;
2. If $Z \sim SN(\lambda)$, then $-Z \sim SN(-\lambda)$;
3. When $\lambda \to +\infty$, $Z \xrightarrow{d} |Y|$, and when $\lambda \to -\infty$, $Z \xrightarrow{d} -|Y|$, where $Y \sim N(0, 1)$;
4. If $Z \sim SN(\lambda)$, then $Z^2 \sim \chi^2_1$.

We use the linear transformation to add the location and scale parameters,

$$X = \xi + \omega Z, \tag{6}$$

where X is a random variable with asymmetric normal distribution with parameters (ξ, ω, λ), i.e., $X \sim SN(\xi, \omega, \lambda)$, and Z is a random variable with density function given by (5).

The probability density function of X is given by

$$f_X(x; \xi, \omega, \lambda) = \frac{2}{\omega}\phi\left(\frac{x - \xi}{\omega}\right)\Phi\left(\lambda\left(\frac{x - \xi}{\omega}\right)\right), \tag{7}$$

where ξ is the location parameter, $\xi \in (-\infty, +\infty)$; ω is the scale parameter, $\omega > 0$; λ is the shape parameter, $\lambda \in (-\infty, +\infty)$, also called asymmetry parameter. The asymmetry of the distribution is limited to the interval $(-1, 1)$.

The expression (7) can be rewritten as

$$f_X(x; \xi, \omega, \lambda) = \frac{1}{\omega\pi}\exp\left\{-\frac{1}{2}\left(\frac{x-\xi}{\omega}\right)^2\right\}\int_{-\infty}^{\lambda\left(\frac{x-\xi}{\omega}\right)} \exp\left\{-\frac{t^2}{2}\right\} dt. \tag{8}$$

2.3 Skew Student's t distribution

The skew Student's t distribution is an extension of the skew normal distribution. Its probability density function is given by

$$f_Z(z; \lambda) = 2t_\nu(z)T(\lambda z), \tag{9}$$

where $t_\nu(z)$ is the density of a Student's t distribution with ν degrees of freedom, $T(\lambda z)$ is the cumulative distribution function of λz and $\lambda \in \Re$ is the asymmetry parameter, see Azzalini and Capitanio (2003). We can recover the Student's t distribution in (9) by setting $\lambda = 0$.

The skew Student's t distribution emerges as a mixture in the opposite scale of an skew normal distribution with a gamma distribution.

Lemma 2.3. *(AZZALINI and CAPITANIO, 2003) If a random variable $V \sim Gamma(\psi, \lambda)$ has mean ψ/λ and variance ψ/λ^2, then for any $a, b \in \Re$*

$$E\left(\Phi\left(a\sqrt{V} + b\right)\right) = P\left(T \leq a\sqrt{\alpha/\beta}\right), \tag{10}$$

where T has Student's t distribution not centered with 2ψ degrees of freedom and $-b$ non-centraly parameter.

Azzaline and Capitanio (2003) applies the Lemma 2.3 to a variable whit distribution $Gamma(\nu/2, \nu/2)$ and defines the density function ot the skew Studet's t distribution given by

$$f_X(x; \xi, \omega, \lambda, \nu) = \frac{2}{\omega}t_\nu(y)T_{\nu+1}\left(\lambda y\sqrt{\frac{\nu+1}{\nu+y^2}}\right), \tag{11}$$

where $y = \frac{x-\xi}{\omega}$ and $y \in \Re$, then X has skew Student's t (ST) distribuion whith parameter $(\xi, \omega, \lambda, \nu)$, i.e., $X \sim ST_\nu(\xi, \omega, \lambda)$ $\lambda \in (-\infty, +\infty)$.

In this case ξ is the location parameter, $\xi \in (-\infty, +\infty)$; ω is the scale parameter, $\omega > 0$; λ is the asymmetry parameter, $\lambda \in (-\infty, +\infty)$, and ν is the degreee of freedom, $\nu \geq 1$.

Some properties of the skew Student's t distribution, as given by Azzalini and Capitanio (2003), are

1. $ST_\nu(0) = T_\nu(0, 1)$.

2. When $\lambda \to +\infty$, $ST_\nu(\lambda)$ tends to the density of the Student's t distribution truncated, $TT_\nu(0,1)$.

3. If $Z \sim ST_\nu(\lambda)$, then $\xi + \omega Z \sim ST_\nu(\xi, \omega, \lambda)$.

2.4 Skew Laplace distribution

The skew Laplace distribution is obtained by the conversion of the Laplace distribution to an asymmetric density function, as described in Kotz *et al.* (2001). This distribution is a continuous probability distribution with three parameters (ξ, ω, λ) and its probability density function is

$$f_X(x; \xi, \omega, \lambda) = \frac{\sqrt{2}}{\omega} \frac{\lambda}{1+\lambda^2} \begin{cases} \exp\left(-\frac{\sqrt{2}\lambda}{\omega}(x-\xi)\right), & \text{para} \ \ x \geq \xi, \\ \exp\left(\frac{\sqrt{2}}{\omega\lambda}(x-\xi)\right), & \text{para} \ \ x < \xi, \end{cases} \tag{12}$$

where ξ is the location parameter, $\xi \in (-\infty, +\infty)$; ω is the scale parameter, $\omega > 0$; and λ is the asymmetry parameter, $\lambda > 0$. The notation used for this distribution is $SL(\xi, \omega, \lambda)$.

The Laplace distribution or double exponential distribution is a particular case of skew Laplace distribution when $\lambda = 1$.

3 Results

In the study of simulation we estimate the parameters of the α-stable, skew normal, skew Student's t and skew Laplace distributions, using the maximum likelihood method, and we calculated the corresponding standard errors of maximum likelihood estimates (EMV's) using Fisher's expected information matrix, $I(\theta_0)$.

Taking the inverse of Fisher's expected information matrix, $[I(\theta_0)]^{-1}$, we have on its main diagonal the values s_{ii}, which are the correspondents standard errors of θ_i, and calculated asymptotic confidence intervals with a coefficient of 95% confidence, for the parameter vector, θ, of the distributions, through expression: $\hat{\theta}_i \pm 1.96\sqrt{s_{ii}}$.

We used the Akaike Information Criterion - AIC (AKAIKE, 1974), and the Schwarz Bayesian Criterion - BIC (SCHWARZ, 1978) to select the best model.

We simulate each random variable X for each distribution using functions avaliable in *packages fBasics, VGAM and sn* of the *R software* (R 2.11.1, 2011).

We simulated data sets using one set of parameters from each of the distributions considered for several sample sizes.

The data was simulate from a α-stable distribution with parameters $\alpha = 1.65$, $\beta = 0.4$, $\gamma = 0.3$ and $\delta = 0.2$, and with sample sizes, $n = 30$, $n = 100$, $n = 1000$ and $n = 10000$, and we fit the model to these observations using α-stable, skew normal, skew Student's t and skew Laplace distributions.

Table 1 - Estimates of parameters and corresponding standard errors

Generating distribution	Parameters values	Estimates of parameters (standard errors)			
		$n = 30$	$n = 100$	$n = 1000$	$n = 10000$
α-stable	α=1.65	1.6375 (0.3554)	1.6740 (0.1821)	1.6749 (0.0443)	1.6462 (0.0153)
	β=0.40	-0.3220 (0.6439)	0.6897 (0.4065)	0.4748 (0.1189)	0.3895 (0.0364)
	γ=0.30	0.2326 (0.0523)	0.3040 (0.0319)	0.2783 (0.0082)	0.3018 (0.0030)
	δ=0.20	0.1558 (0.2975)	0.1602 (0.0668)	0.1950 (0.0187)	0.2010 (0.0066)
Skew normal	ξ=-0.20	-0.5060 (0.6526)	-0.0313 (3.4313)	-0.1793 (0.1254)	-0.2274 (0.0257)
	ω=0.50	0.5612 (0.3224)	0.4164 (0.0364)	0.4831 (0.0533)	0.5108 (0.0134)
	λ=0.70	0.9999 (2.8502)	0.0070 (16.1544)	0.2352 (1.6372)	0.6443 (0.1252)
Skew Student's t	ξ=-0.20	-0.4658 (0.0921)	-0.3180 (0.1105)	-0.2061 (0.0635)	-0.2050 (0.0216)
	ω=0.50	0.7850 (0.2181)	0.5795 (0.1143)	0.4833 (0.0322)	0.4989 (0.0109)
	λ=0.60	8.6902 (8.8983)	1.8165 (0.8752)	0.5993 (0.1983)	0.6098 (0.0659)
	ν=5	40.6564 (323.4978)	4.2818 (1.7717)	4.3650 (0.6135)	4.9473 (0.2425)
Skew Laplace	ξ=-0.20	-0.2730 (0.0170)	-0.2110 (0.0171)	-0.1844 (0.0110)	-0.2069 (0.0046)
	ω=0.50	0.4401 (0.0850)	0.4084 (0.0450)	0.4817 (0.0166)	0.5007 (0.0057)
	λ=0.60	0.6159 (0.0903)	0.5672 (0.0507)	0.6159 (0.0190)	0.5910 (0.0065)

We repeated the same procedure applied above for skew normal, skew Student's t and skew Laplace distribubions, i.e., we simulate the data sets from a distribution and fitting all four distributions.

The data sets was simulate from a skew normal distribution with parameters $\xi = -0.2$, $\omega = 0.5$, and $\lambda = 0.7$; and from a skew Student's t distribution with parameters $\xi = -0.2$, $\omega = 0.5$, $\lambda = 0.6$ and $\nu = 5$; and from a skew Laplace distribution with parameters, $\xi = -0.2$, $\omega = 0.5$ and $\lambda = 0.6$, all of them with sample sizes given above.

Table 1 shows the true values of parameters, their estimates and corresponding standard errors for sample sizes from 30 to 10000, for replica of simulated data from α-stable, skew normal, skew Student's t and skew Laplace distributions.

We observe that as the sample size increases the parameter estimates of all parameters get closer to the true values and the corresponding standard errors tend to zero.

These results indicate that the estimators are consistent and also validate the computational procedure.

The exceptions are the estimates of asymmetry parameter λ of the skew normal distribution and the degrees of freedom of the skew Student's t distribution, ν, for sample size of $n = 100$.

In the study of simulation we also identified not convergencie of the estimation of the skew normal distribution parameters when the observations are generated from a α-stable distribution, with parameter $\alpha = 0.1$ and large sample size, n $=$ 10000. Also, when the α-stable distribution parameter approaches 2.0, the program is unable to estimate this parameter.

Simulating data sets from each distribution and using AIC and BIC criteria to identify the best fitting model produced the results presented in Table 2. We can observe that for sample sizes $n \geq 100$ the model is correctly identified for all distributions considered. For $n = 30$ data sets generated from α-stable and skew Student's t are identified as skew Laplace and data sets generated as skew Laplace and skew normal are identified as skew normal distribution. Those results indicate that for large sample sizes criteria AIC and BIC are effective in identifying the best model (n>100 in our experiments).

Table 2 - Results of model selection using *AIC* and *BIC* criteria

Generation Distribution	Modeled Distributions	$n = 30$		$n = 100$		$n = 1000$		$n = 10000$	
		AIC	*BIC*	*AIC*	*BIC*	*AIC*	*BIC*	*AIC*	*BIC*
α-stable	α-stable	36.9	56.1	**163.3**	**192.1**	**1352.3**	**1399.6**	**15415.2**	**15481.0**
	normal(*)	35.3	49.7	359.7	381.4	1895.8	1931.2	25573.8	25623.1
	Student"s t(*)	36.3	55.5	166.2	195.0	1364.2	1411.4	15497.6	15563.3
	Laplace (*)	**34.2**	**48.6**	194.9	216.5	1438.7	1474.2	16602.6	16651.9
normal(*)	α-stable	50.6	69.8	116.6	145.43	1200.6	1247.8	12301.8	12097.5
	normal(*)	**48.5**	**63.0**	**114.6**	**136.2**	**1198.3**	**1233.7**	**12018.9**	**12068.2**
	Student's t(*)	50.6	70.0	116.6	145.4	1200.5	1247.5	12020.7	12086.4
	Laplace(*)	51.1	65.5	151.7	173.3	1292.5	1327.9	12954.4	13003.7
Student's t(*)	α-stable	91.7	111.0	165.8	194.7	1709.7	1757.0	18472.6	18538.2
	normal(*)	49.5	63.9	187.7	209.3	1806.8	1842.3	17836.0	17885.3
	Studen's t(*)	43.3	62.5	**165.2**	**194.1**	**1701.9**	**1749.2**	**16937.5**	**17003.1**
	Laplace(*)	**41.5**	**55.9**	188.8	216.54	1720.2	1755.6	17224.7	17273.94
Laplace(*)	α-stable	46.6	65.8	177.7	206.6	1502.5	1549.7	20362.6	20428.3
	normal(*)	**42.0**	**56.4**	154.1	175.8	1721.3	1756.8	18596.1	18645.4
	Student's t(*)	42.5	61.7	126.8	154.5	1478.4	1525.7	15925.5	15991.2
	Laplace(*)	44.5	58.9	**125.7**	**148.5**	**1464.9**	**1500.3**	**15747.8**	**15797.1**

(*)Skew

Table 3 - Coverage of Confidencial Intervals of 95%

Generation Distributions	Parameters Values	Coverage (%)			
		$n = 30$	$n = 100$	$n = 1000$	$n = 10000(*)$
α-stable	α=1.65	92.6	96.2	96.2	94.7
	β=0.40	93.4	95.5	94.3	92.8
	γ=0.30	88.3	93.2	89.8	95.6
	δ=0.20	99.3	96.1	99.2	93.6
skew normal	ξ=-0.20	99.9	100.0	99.8	95.4
	ω=0.50	83.2	81.6	82.9	96.5
	λ=0.70	100.0	100.0	99.8	96.1
Skew Student's t	ξ=-0.20	87.3	92.4	94.4	95.3
	ω=0.50	88.4	94.1	94.0	94.9
	λ=0.60	100.0	99.9	95.6	95.3
	ν=5	87.6	91.5	94.9	95.0
Skew Laplace	ξ=-0.20	95.3	95.3	84.2	84.2
	ω=0.50	86.6	90.5	94.9	95.5
	λ=0.60	72.1	80.3	86.7	91.1

(*) For this size of sample replicates were made 640

Using the same true values of parameters given the above the simulation process was replicated 1000 times for $n \in \{30, 100, 1000\}$ and was replicated 640 times for $n = 10000$. For each simulated data set we estimate 0.95 confidence intervals. The proportion of times the interval contains the true value are presented in Table 3.

As we observe, the intervals for dispersion/scale parameters are under dimensioned, γ for α-stable distributon and ω for skew normal, Student's t and Laplace distributions. The same occurs with the asymmetry and location parameters of the skew Laplace, λ and ξ respectively. For small sample sizes there are several parameters with under dimensioned confidence intervals but results are not conclusive.

3.1 Application to gene expression data

The gene expression data are from an experiment known in the *microarray* literature as "Swirl Zebrafish" (*Danio rerio*), Ferreira and Leandro (2009). This experiment was conducted using the fish "Zebrafish" as a model organism for the study of growth in vertebrates.

A goal of the Swirl experiment is to identify genes with altered expression in the swirl mutant compared to wild-type zebrafish. Two sets experiments were performed, for a total of four replicate hybridizations. For each of these hybridizations, target cDNA from the swirl mutant was labeled using one of the Cy3 or Cy5 dyes and the target cDNA wild-type mutant was labeled using the other dye. Target cDNA was hybridized to microarrays containing 8.448 cDNA probes. The data sample consists of 33.792 observations, and the data were removed from the site$<http://bioinf.wehi.edu.au/limmaGUI/DataSets.html>$.

We estimate the parameters of the α-stable, skew normal, skew Student's t and skew Laplace distributions for gene expression data, and calculate the asymptotic Confidence Intervals for these parameters, as described in Mood and Graybill (1974). Table 4 presents the results. The significance test for the asymmetry parameters of the distributions indicates that there is evidence for asymmetry, β of the α-stable (p-value $<$0.0002), λ of the skew normal (p-value $<$0.0001), skew Student's t distribution (p-value $<$0.001) and skew Laplace (p-value $<$0.0001) distributions are all significant at 0.05 significance level.

Figure 1(a) illustrate the histogram and the fitted distributions. Figure 1(b) shows the estimated densities for the considered distributions: α-stable (solid line), skew normal (dashed line), skew Student's t distribution (dotted line) and skew Laplace (dashed-dotted line).

The model selection criteria *AIC* and *BIC* indicate the α-stable distribution as the best fitting model. These results are in Table 5.

Table 4 - Parameters estimates, standard errors and 0.95 confidence intervals for the “Swirl Zebrafish” data

Distributions	Parameters	Estimates	Standart Errors	2.5%	97.5%
α-stable	α	1.9160	0.0055	1.9053	1.9267
	β	**0.2257**	0.0560	0.1160	0.3354
	γ	0.3103	0.0014	0.3076	0.3130
	δ	-0.2888	0.0026	-0.2940	-0.2837
skew normal	ξ	-0.6112	0.0064	-0.6265	-0.5958
	ω	0.5789	0.0048	0.5694	0.5884
	λ	**0.9999**	0.0314	0.9393	1.0605
skew Student’s t	ξ	-0.3673	0.0199	-0.4064	-0.3282
	ω	0.4187	0.0042	0.4105	0.4269
	λ	**0.2208**	0.0580	0.1071	0.3344
	ν	8.5805	0.3043	7.9841	9.1770
skew Laplace	ξ	-0.2607	0.0042	-0.2689	-0.2525
	ω	0.5203	0.0029	0.5148	0.5259
	λ	**1.0350**	0.0071	1.0211	1.0489

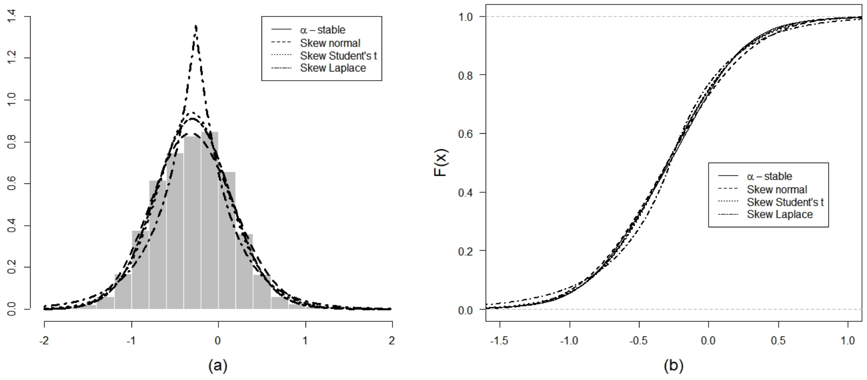

Figure 1 - Histogram and pdf’s of fitted densities (a) and fitted cdf’s (b) for “Swirl Zebrafish” data.

Table 5 - "Swirl Zebrafish" data: Model selection AIC's and BIC's

Distributions	AIC	BIC
α-stable	**43768.28**	**43843.70**
skew normal	45901.85	45958.41
skew Student's t	44190.40	44265.83
skew Laplace	46898.73	46955.30

Discusions

We develop a simulation study to explore practical aspects of asymmetric distributions α-stable, skew normal, skew Student's t and skew Laplace.

We have that as the sample size increses the standard errors are reduced, indicating consistency of the estimators used and also validating the computational procedures.

We also verify the capability of AIC and BIC to identify the right model. In our study of simulate the choice of the right model happened for sample sizes $n \geq 100$ for all distributions considered but this is parameter dependent.

We detected that the confidence intervals for dispersion/scale parameters are under dimensioned for all sample sizes considered while for small sample sizes the results are not conclusive.

The application to gene expression data results identify the α-stable distribution as the best fitting model indicating the relevance of this distribution.

This result agrees with Gonzalez *et al.*, (2009) in which the authors apply the *alpha*-stable distribution for 4 different cDNA dual dye microarray datasets, including "Swirl zerafish".

In general, gene expression data sets are composed of sample sizes $n \geq 100$, like the data sets analyzed by Gonzalez *et al.*, (2009). Therefore, the results found in this simulation study are valid to analyze gene expression data.

There were also difficulties in the estimation of the parameters. This happens when the true values of parameters are near the border of the region where the parameters were defined.

MACERAU, W. M. O.; MILAN, L. A. Usando distribuições assimétricas para modelagem de dados de expressão gênica. *Rev. Bras. Biom.,* Lavras, v.39, n.4, p.266-278, 2021.

- *RESUMO: Apresentamos uma breve revisão das distribuições assimétricas α-estável, normal, t de Student e Laplace. Comparamos o desempenho dessas distribuições, em geral, usadas para modelar dados assimétricos, usando AIC e BIC. Esses critérios foram capazes de selecionar o melhor modelo para cada conjunto de dados. Também aplicamos esses modelos a dados de expressão gênica e verificamos que essas distribuições são qualificadas para modelar essas observações.*

▪ *PALAVRAS-CHAVE: Expressão gênica, distribuições assimétricas, α-estável, normal assimétrica, t de Student assimétrica, Laplace assimétrica.*

References

ARELLANO-VALLE, R. B.; BOLFARINE, H.; LACHO, V. H. Skew-normal Linear Mixed Models. *Journal of Data Science*, v.3, p.415-438, 2005.

AKAIKE, H. A new look at the statistical model identifications. *IEEE Transations on Automatic Control*, v.19, n.6, p.716-723, 1974.

AZZALINI, A. A Class of Distributions which Includes the normal Ones. *Scand Journal Statistical*, v.12, p.171-178, 1985.

AZZALINI, A.; CAPITANIO, A. Robustness in real life: a study of clinical laboratory data. *Biometrics*, v.38, p.377-396, 1999.

AZZALINI, A.; CAPITANIO, A. Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew t Distribution. *Journal of the Royal Statistical Society*, v.65, p.367-389, 2003.

BOZDOGAN, H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, v.52, n.3, p.345-370, 1987.

FERNÁNDEZ, C.; OSIEWALSKI, J.; STEEL, M. F. J. Modeling and Inference With *v*-Spherical Distribution. *Journal of the American Statistical Association*, v.90, p.1331-1340, 1995.

FERNÁNDEZ, C.; STEEL, M. F. J. On Bayesian Modeling of Fat Tails and Skewness. *Journal of the American Statistical Association*, v.93,n.441, p.359-371, 1998.

FERREIRA FILHO, D.; LEANDRO, R. A. *Análise de Microarray usando o R e o Biocondutor*. Tutorial apresentado no 54°RBRAS e 13°SEAGRO, 2009.

GENTON, M. G.; HE, L.; LIU, X. Moments of skew-normal randon vectors and their quadratic forms. *Statistics and Probability Letters*, v.51, p.319-325, 2001.

GONZALEZ, D. S.; KURUOGLU, E. E.; RUIZ, D. P. Modelling and Assessing Differential Gene Expression Using the Alpha Stable Distribution. *The International Journal of Biostatistics*, v.5, n.1, p.16, 2009.

GUPTA, A. K.; NGUYEN, T. T.; SANQUI, J. A. T. Characterization of the Skew-normal Distribution. *The Institute of Statistical Mathematics*, v.56, p.351-360, 2004.

JONES, M. C.; FADDY, M. J. A skew extension of the *t*-distribution, with aplications. *Journal of the Royal Statistical Society*, v.6,n.1, p.159-174, 2003.

JULIÀ, O.; VIVES-REGO, J. A microbiology application of the Skew-Laplace distribution. *Sort*, v.2, p.141-150, 2008.

KOTZ, S.; KOZUBOWSKI, T. J.; PODGÓRSKI, K. *The Laplace distribution and generalizations. A revisit with applications to Communications, Economics, Engineering and Finance.* Birkhaüser, 2001, 349p.

LEE, P.M. *Bayesian Statistical: and Introduction.* 2nd Edition, Edward Arnold, 1996, 344p.

LITTELL, R. C.; MILIKEN, G. A.; STROUP, W. W.; WOLFINGER, R. D. *SAS System for Mixed Models.* Cary: Statistical Analysis System Institute, v.633, 2002.

MOOD, A. M.; GRAYBILL, F. A. *Introduction to the Theory of Statistics.* McGraw-Hill International Editions, 1974, 480p.

NOLAN, J. P. *Stable Distributions: Models for Heavy Tailed Data*, 2009. Capítulo 1. Disponível em: <http://academic2.american.edu/ jpnolan/stable/chap1.pdf.> Acesso em: 1 mar. 2011.

NOLAN, J. P. *Modeling financial distributions with stable distributions.* Volume 1 of Handbooks in Finance, section 3, pp. 105-130. Amsterdam: Elsevier, 2003. $http://bioinf.wehi.edu.au/limmaGUI/DataSets.html$.

RACHEV, S. T.; MITTINIK, S. *Stable Paretian Models in Finance.* New York, NY. Wiley, 2000, 874p.

RACHEV, S. T. *Hanndbook of Heavy Tailed Distributions in Finance.* Amsterdam. Wiley, 2003, 704p.

RUBIO, F. J.; STEEL, M. F. J. Inference for grouped data with a truncated skew-Laplace distribution. *University of Warwick Institutional Repository*, v.10, p.20, 2010.

R. *R-Project Software Version 2.11.1.* Disponível em:<http://www.r-project.org>.

SCHWARZ, G. E. Estimating the dimension of a model. *Annals of Statistics*, v.6, n.2, p.461-464, 1978.

WOLFINGER, R. D. Covariance estruture selection in general mixed models. *Comunications in Statistics*, v.22, p.1079-1106, 1993.

Received on 18.03.2020.

Approved after revised on 18.02.2021.