Contents list available at JMCS
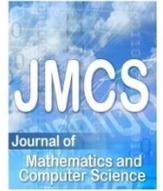
Journal of Mathematics and Computer Science

Journal Homepage: www.tjmcs.com

Goodness of Fit Test for The Skew-T Distribution

M. Maghami[1,*] , M. Bahrami[1,+]

[1]Department of Statistics, University of Isfahn, Isfahan, Iran
[*]*maghami8@gmail.com*
[+]*m.bahrami@sci.ui.ac.ir*

Abstract

In this manuscript goodness-of-fit test is proposed for the Skew-t distribution based on properties of the family of these distributions and the sample correlation coefficient. The critical values for the test can be achieved by Monte Carlo simulation method for several sample sizes and levels of significance. The power of the proposed test can be specified for different sample sizes and considering diverse alternatives.

Keywords: Sample correlation coefficient; Skew-t; Goodness-of-fit test.

1. Introduction

Let Z be a random variable, we say that Z has the *Skew-normal* distribution, denoted by $Z \sim SN(\lambda)$, if its probability density function be

$$f_z(z;\lambda) = 2\phi(z)\Phi(\lambda z)I_{(-\infty,\infty)}(z), \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denotes the density and cumulative distribution function of standard Normal distribution respectively. The skew-normal distribution was introduced by Azzalini (1985), as a family with the appealing property of strictly including the normal law, as well as a wide variety of skewed densities. We say that a random variable $W_{\lambda,v}$ has the *Skew-t* distribution with parameters $v > 0$ and $\lambda \in \mathrm{R}$ if $W_{\lambda,v} = Z_\lambda / \sqrt{V}$, where Z_λ is the skew-normal variable with pdf in (1), $V \sim \chi_v^2 / v$, and are independent. This variable is denoted by $W_{\lambda,v} \sim St(\lambda, v)$.

If a random variable Y is defined as

$$Y = \mu + \sigma W_\lambda \tag{2}$$

with $\mu \in \mathrm{R}$, $\sigma \in \mathrm{R}^+$, then $Y \sim St(\mu, \sigma, \lambda, v)$. Skew-Cauchy distribution is obtained simply as special cases of the skew-t with $v = 1$ and denoted by $SC(\mu, \sigma, \lambda)$.

Some well know properties of skew-t variables which will be useful for constructing goodness of fit test are the following (See for details [3]):

(a) If $W_{\lambda,\nu} \sim St(\lambda,\nu)$, then $-W_{\lambda,\nu} \sim St(-\lambda,\nu)$.

(b) If $W_{\lambda,\nu} \sim St(\lambda,\nu)$, then $W_{\lambda,\nu}^2 \sim F(1,\nu)$.

2. EDF-Based Tests

Perez Rodrguez and Villasenor (2010) developed a goodness of fit test for the skew normal family based on the sample correlation coefficient and showed that their test have greater power than the Empirical Distribution Function-based tests against some alternative distributions. We are interested in testing the null hypothesis

$$H_0 : Y \text{ is } St(\mu,\sigma,\lambda,\nu) \text{ for some } \lambda \in R, \mu \in R, \sigma \in R^+, \nu \in R^+ \qquad (3)$$

against general alternatives. In this section we discuss general EDF-based goodness-of-fit statistics designed to test the null hypothesis H_0 . EDF-based test statistics measure the difference between the distribution function, $F(.)$, stated in the null hypothesis and the EDF, a step function denoted by $F_n(.)$ given as

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}, \\ i/n, & x_{(i)} \le x < x_{(i+1)}, \\ 1, & x_{(n)} < x, \end{cases}$$

where $x_{(1)} \le ... \le x_{(n)}$ are the ordered statistics of the x_i's. To compare the two distribution functions, several statistics can be used that Stephens (1986) divides into two families. The Cramér-von Mises family contains the Cramér-von Mises statistic, W^2 , Watson's U^2 statistic, and the Anderson-Darling statistic, A^2 , defined as:

$$W^2 = n\int_{-\infty}^{\infty} \left(F_n(x) - F(x)\right)^2 dF(x)$$

$$U^2 = n\int_{-\infty}^{\infty} \left[F_n(x) - F(x) - \int_{-\infty}^{\infty}\left(F_n(t) - F(t)\right)dF(t)\right]^2 dF(x)$$

$$A^2 = n\int_{-\infty}^{\infty} \left(F_n(x) - F(x)\right)^2 \left[F(x)\left(1 - F(x)\right)\right]^{-1} dF(x)$$

The Kolmogorov-Smirnov family contains the statistics D^+ , D^- , the Kolmogorov-Smirnov statistic, D , and the Kuiper statistic, V , defined as:

$$D^+ = \sup_x \left(F_n(x) - F(x)\right), \quad D^- = \sup_x \left(F(x) - F_n(x)\right).$$

$$D = \max\left(D^+, D^-\right), \qquad V = D^+ + D^-.$$

Stephens (1986) provides the following simple formulae for calculating these statistics:

$$W^2 = \sum_{i=1}^{n} \left(P_{(i)} - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n} \qquad (4)$$

$$U^2 = W^2 - n\left(\overline{p} - \frac{1}{2}\right)^2 \tag{5}$$

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log p_{(i)} + \log\left(1 - p_{(n+1-i)}\right)\right] \tag{6}$$

$$D^+ = \max_i\left(\frac{i}{n} - p_{(i)}\right) \tag{7}$$

$$D^- = \max_i\left(p_{(i)} - \frac{i-1}{n}\right) \tag{8}$$

$$D = \max\left(D^+, D^-\right) \tag{9}$$

$$V = D^+ + D^-, \tag{10}$$

Where $p_{(i)} = F\left(x_{(i)}\right)$ and $\overline{p} = \sum_{i=1}^{n} p_{(i)}/n$. Large values of a given statistic indicate significant differences between the empirical and hypothesized distribution functions and thus that we should reject the null hypothesis.

In general, when the parameter values of the hypothesized distribution are completely specified, the sampling distribution of any of these EDF statistics is known exactly, and tables of percentage points are available (see Stephens (1986), Table 4.2). However, when the values taken by the parameters of the distribution are unknown and have to be estimated from the sample, the sampling distribution of any EDF statistic depends on the distribution being tested, sample size, true values of the unknown parameters, and method used to estimate the parameters.

Now we describe the parametric bootstrap techniques used to estimate the quantiles of the test statistic T when the hypothesized distribution is skew-t with parameter values estimated from the data. Maximum likelihood methods can be employed to estimate the parameters of the skew-t distribution. Since analytic expressions do not exist for these estimators, numerical methods must be used to compute them. Note that when the unknown parameters are location or scale parameters, and they are estimated using location and scale equivariant estimators (as are maximum likelihood estimators), the sampling distributions of the EDF statistics do not depend on the true values of those parameters. (see Eastman and Bain (1973)). Therefore the values of $\mu = 0$ and $\sigma = 1$ were used for simplicity because of the sampling distributions of the statistics being invariant to changes in the location and scale parameters.

Since however the asymptotic null distribution of the test statistic depends upon the unknown value of λ and v , a parametric bootstrap versions of the test is performed:

1. Given the sample $y_1,..., y_n$ compute the maximum likelihood estimator $\hat{\lambda}$ and $\hat{v}$ of λ and v .

2. Calculate the value of the chosen test statistic, T , using the appropriate formula(e) from Eqs. (4) - (10) , where $F(.)$ denoted the distribution function of $St\left(\hat{\lambda}, \hat{v}\right)$.

(a) Generate a bootstrap sample of size n from $St\left(\hat{\lambda}, \hat{v}\right)$.

(b) Given the bootstrap sample generated previously, compute the ML estimators of λ and ν say $\hat{\lambda}^*$ and $\hat{\nu}^*$.

(c) Compute the value of the test statistic, say T^* using $\hat{\lambda}^*$, $\hat{\nu}^*$ and the bootstrap sample.

3. Repeat steps (a), (b) and (c) 1000 times to get T_j^* , $j = 1,...,1000$.

4. Obtain $T_n(0.05)$ as $T_{(950)}^*$, where $T_{(j)}^*$, $j = 1,...,1000$ denotes the ordered T_j^* values.

3. Correlation Goodness-of-Fit Test

In this section, we introduced goodness-of fit test for skew-t distribution with sample correlation coefficient. The test procedure is based on property (b). From Eq. (2):

$$Y - \mu = \sigma W_{\lambda,\nu},$$

where $Y \sim St(\mu, \sigma, \lambda, \nu)$ and $W_{\lambda,\nu} \sim St(\lambda,\nu)$, then

$$X := (Y - \mu)^2 = \sigma^2 W_{\lambda,\nu}^2 \tag{11}$$

By property (b):

$$\frac{X}{\sigma^2} := W_{\lambda,\nu}^2 \sim F(1,\nu). \tag{12}$$

From (12) parameter λ has been eliminated from the problem.

For fixed μ and ν , say $\mu = \mu_0$ and $\nu = \nu_0$, X has a scale distribution,

$$P(X \leq x) = G(\frac{x}{\sigma^2})$$

where G is the distribution function of a $F(1,\nu_0)$ random variable. So given the sample $y_1,.., y_n$ and μ_0 , calculate $x_1,..., x_n$ by using (11). A consistent estimator for $P(X \leq x)$ is the empirical distribution function, then

$$G(\frac{x}{\sigma^2}) \approx F_n(x)$$

therefore

$$u := G^{-1}\big(F_n(x)\big) \approx \frac{x}{\sigma^2} \tag{13}$$

Since (13) is established, we should expect a strong linear relationship between x_i's and u_i's under the null hypothesis stated in (3). If μ and ν are estimated by consistent estimators, say μ and ν then it is expected that the linear relationship in (14) still holds. To test if there is a strong linear relationship between x_i's and u_i's the sample correlation coefficient statistic is used, which is given by

$$C_n = Corr(X,U) = \frac{\sum\limits_{i=1}^{n}(X_i - \overline{X})(U_i - \overline{U})}{\sqrt{\sum\limits_{i=1}^{n}(X_i - \overline{X})^2(U_i - \overline{U})^2}} \tag{14}$$

The null hypothesis (3) is rejected at the level of significance α if $C_n \leq C_n(\alpha)$, where $C_n(\alpha)$ is such that

$$\alpha = \max_{\lambda,\nu} P(\text{Re } ject \ H_0 \mid H_0) = \max_{\lambda,\nu} P(C_n \leq C_n(\alpha)). \tag{16}$$

The distribution of C_n under the null hypothesis for each fixed value of v and λ can be obtained by Monte Carlo simulation. Note that C_n is scale invariant and the distribution of X in (11) does not depend on μ , therefore we will fix $\mu = 0$ and $\sigma = 1$. If the random sample comes from a distribution function different from the skew-t distribution for which property (b) does not hold then it means that (13) does not hold. Therefore the sample correlation coefficient in (14) can not be near 1, hence C_n should be lower than the critical value, since under H_0 the distribution of C_n will be concentrated close to 1.

Therefore we use the following procedure to obtain the critical values:

1. Fix n , λ , v , $\mu = 0$, $\sigma = 1$.

2. Simulate a sample of size n from $St(\mu, \sigma, \lambda, v)$

3. Calculate the maximum likelihood estimator of parameter μ .

4. Calculate x_i , $i = 1,...,n$ using Eq. (11).

5. Sort x_i's into ascending order.

6. Calculate $u_i := G^{-1}(F_n(x_i))$, $i = 1,...,n$, where G^{-1} is the quantile function of the $F(1,v)$ distribution.

7. Calculate C_n using Eq. (14) and the data x_i , u_i generated in steps 5 and 6.

8. Repeat steps 2-7 B times.

Upon finishing the simulation process, we have B realizations of C_n for a given value of λ and v .

Therefore the value of the critical constant $C_n(\alpha)$ is determined with the quantiles 100α from the empirical distribution of C_n .

For example, fig. 1 presents graph of $C_n(\alpha)$ as a function of λ , $\alpha = \{0.1, 0.05, 0.025\}$ and $n = 50$ for $v = 2$, which shows that the distribution of the test statistic C_n under H_0 not depends on the value of the unknown parameter λ . Our simulations show this fact indefeasible for arbitrary v .

Note that we have limited our attention to $Y \sim St(\mu, \sigma, \lambda, v)$ with $\lambda \geq 0$ since $-Y \sim St(-\mu, \sigma, -\lambda, v)$ by property (a). Therefore distribution of C_n does not depend on the sign of λ , hence the critical constant $C_n(\alpha)$ in (15) is such that

$$\alpha = \max_{\lambda, v} P(C_n \leq C_n(\alpha)) = \max_{\lambda \geq 0, v} P(C_n \leq C_n(\alpha)).$$

For arbitrary v , simulations show that the values of the critical constant $C_n(\alpha)$ are determined with the quantiles 100α from the empirical distribution of C_n obtained by simulation with arbitrary λ . Fig. 1 show this fact for $v = 2$.
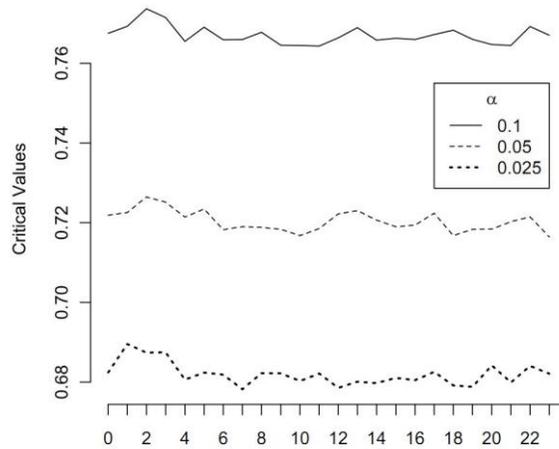
Figure 1. Critical values as a function of λ for $n = 50$, $v = 2$, $B = 5000$ for the statistic C_n .

Given a random sample of n data values, the steps necessary to carry out a given test can be summarized as follows:

1. Calculate the MLEs of the μ and v , using library `sn' (Azzalini (2008)) in R (R Development Core and Team, 2008) and denote by $\hat{\mu}$ and $\hat{v}$.

2. Calculate the value of the test statistic, C_n , using the Eq. (14).

3. For a given significance level, α , identify the quantile $C_n(\alpha)$ of the test statistic corresponding to $v = \hat{v}$ and n .

4. If $C_n \leq C_n(\alpha)$, the null hypothesis is rejected at the α significance level.

4. Simulation studies

4-1. Tests size

The results of size estimations of tests presented in Table 1 and 2, obtained by simulation for $\alpha = 0.05$. The selected sample sizes were 50 and 100 the value of parameter $(\lambda, v) = \{(0,4),(1,1),(2,5),(3,7),(5,20),(7,10),(10,0.5),(20,30)\}$, $\mu = 0$, $\sigma = 1$.

From Table 1 and 2 it can be seen that the estimated tests sizes are very close to the nominal significance level α .

Table 1: Test size estimates using the statistics obtained by simulation with $B = 10000$ Monte Carlo samples of size $n = 50$ with $\alpha = 0.05$.

Statistic	(λ, v)							
	$(0,4)$	$(1,1)$	$(2,5)$	$(3,7)$	$(5,20)$	$(7,10)$	$(10,0.5)$	$(20,30)$
A^2	0.024	0.047	0.035	0.053	0.042	0.002	0.041	0.056
W^2	0.057	0.054	0.023	0.015	0.011	0.011	0.029	0.017
U^2	0.024	0.057	0.043	0.025	0.011	0.012	0.038	0.041
D	0.043	0.039	0.037	0.012	0.036	0.043	0.023	0.054
V	0.021	0.018	0.054	0.037	0.024	0.037	0.054	0.048
C_n	0.047	0.053	0.033	0.023	0.01	0.015	0.056	0.007

Table 2: Test size estimates using the statistics obtained by simulation with $B = 10000$ Monte Carlo samples of size $n = 100$ with $\alpha = 0.05$.

Statistic	(λ, v)							
	$(0,4)$	$(1,1)$	$(2,5)$	$(3,7)$	$(5,20)$	$(7,10)$	$(10,0.5)$	$(20,30)$
A^2	0.014	0.021	0.017	0.034	0.054	0.027	0.037	0.021
W^2	0.021	0.031	0.020	0.008	0.041	0.034	0.022	0.034
U^2	0.023	0.041	0.053	0.049	0.033	0.024	0.029	0.017
D	0.045	0.051	0.037	0.047	0.041	0.051	0.038	0.042
V	0.035	0.027	0.053	0.042	0.026	0.021	0.022	0.031
C_n	0.04	0.054	0.023	0.015	0.011	0.011	0.049	0.009

4-2. Tests power

To analyze the behavior of the proposed tests, alternatives different to the skew-t were considered. The distributions selected for this were: skew-slash (SSL), Logistic, Exponential, Chi squared, Weibull, Gumbel, Log Normal, and Stable (see Nolan (1999)). We also considered some bimodal distributions. The results are shown in Tables 3 and 4, from which it can be seen that the proposed test C_n show the highest powers for several of the considered alternatives.

Table 3: Power estimates of the A^2, W^2, U^2, D, V, C_n statistics for some alternatives with $n = 50$, $\alpha = 0.05$, $B = 5000$.

Alternative	A^2	W^2	U^2	D	V	C_n
SSL(0,1,1,1)	0.524	0.624	0.413	0.341	0.271	**0.745**
Standard logistic	**0.421**	0.314	0.217	0.248	0.312	0.417
Standard exp.	0.321	0.624	0.221	0.541	0.347	**0.652**
Chisquared(4)	0.078	**0.095**	0.074	0.023	0.047	0.092
Weibull(0.75,1)	0.231	0.712	0.124	0.224	0.317	**0.841**
Standard Gumbel	0.529	0.427	0.471	0.521	0.321	**0.537**
Log-Normal(0,0.5)	**0.712**	0.562	0.541	0.321	0.271	0.694
Sta(1.6,0.25,1,0;0)	0.314	0.421	0.124	0.213	0.119	**0.451**
0.5N(4.5,0.5)+0.5N(-4.5,0.5)	0.817	0.910	0.514	0.617	0.419	**0.913**
0.9N(4.5,0.5)+0.1N(-4.5,0.5)	**0.952**	0.514	0.248	0.512	0.642	0.934
0.5N(1,(1/3))+0.5N(-1,(1/3))	0.521	0.612	0.217	0.274	0.415	**0.664**
0.9N(1,(1/3))+0.1N(-1,(1/3))	0.412	**0.574**	0.321	0.217	0.326	0.291

Table 4: Power estimates of the A^2, W^2, U^2, D, V, C_n statistics for some alternatives with $n = 100$, $\alpha = 0.05$, $B = 5000$.

Alternative	A^2	W^2	U^2	D	V	C_n
SSL(0,1,1,1)	0.617	0.714	0.321	0.412	0.319	**0.753**
Standard logistic	0.502	0.421	0.312	0.231	0.345	**0.518**
Standard exp.	0.345	0.715	0.431	0.614	0.457	**0.761**
Chisquared(4)	0.093	**0.117**	0.082	0.051	0.071	0.103
Weibull(0.75,1)	0.417	0.629	0.396	0.428	0.627	**0.867**
Standard Gumbel	**0.746**	0.641	0.681	0.629	0.471	0.727
Log-Normal(0,0.5)	**0.875**	0.641	0.508	0.441	0.325	0.793
Sta(1.6,0.25,1,0;0)	0.424	0.517	0.229	0.487	0.247	**0.673**
0.5N(4.5,0.5)+0.5N(-4.5,0.5)	0.803	**0.945**	0.673	0.741	0.576	0.944
0.9N(4.5,0.5)+0.1N(-4.5,0.5)	0.961	0.621	0.323	0.549	0.673	**0.967**
0.5N(1,(1/3))+0.5N(-1,(1/3))	0.574	0.723	0.315	0.417	0.529	**0.801**
0.9N(1,(1/3))+0.1N(-1,(1/3))	0.521	**0.672**	0.421	0.321	0.312	0.384

6. Numerical example

To illustrate how the test procedure works with real data, we use data collected at the Australian Institute of Sport (AIS) (Cook & Weisberg (1994)) containing 102 male athletes of body mass index (BMI). Table 5 reports maximum likelihood estimators of some skew models, considering the full $St(\lambda_1, \lambda_2, \nu)$ model and two special cases: Skew-normal and Skew-cauchy. The Akaike information criterion (AIC) is used to compare the estimated models (Leroux (1992)). As is well known, a model with a minimum AIC value is to be preferred. Therefore the St fit appears to be preferable. These points are further illustrated in Figure 3, where a histogram of the data is plotted together with the fitted densities.

Table 5: MLE estimates and Log-likelihood values.

Model	SN	SC	St
μ	20.797082	22.7597	21.370376
σ^2	4.137343	1.38549	2.972035
λ	3.62291	0.5401	2.140406
ν	-	-	5.613827
Log-likelihood	-237.8347	-247.5523	-235.9303
AIC	481.6694	501.1046	479.8606

$$AIC = -2\log(L) + 2k .$$

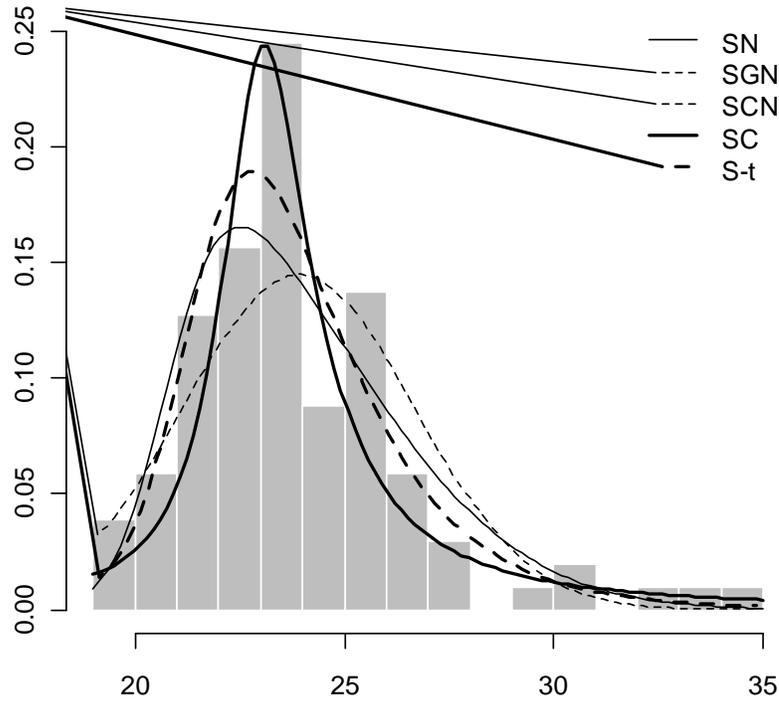L and k are the maximized log likelihood and number of parameters.

Figure 2: Histogram of BMI of 102 Australian athletes. The lines represent distributions fitted using maximum likelihood estimation.

However, the goodness of fit test for this data Skew-Normal and Skew-Cauchy rejection of SN and SC models, but we can not reject the hypothesis of an underlying skew-t population for data set. (See for details [8]). The results are summarized in Table 6 . The critical points, the corresponding value of the test statistics and range of $P-value$ given in Table 6 .

Table 6: Critical points and values of the test statistics for the BMI data

Model	SN		SC	St
Test statistics	$r_n = 0.9461928$	$r_n^* = 0.9754747$	$Rn = 0.7682234$	$Cn = 0.976186$
1%	0.9114055	0.9738395	0.7342562	0.7389985
2.5%	0.9374606	0.9798094	0.7750845	0.7833815
5%	0.9520816	0.9834587	0.8154029	0.8275803
10%	0.9666230	0.9868670	0.8645460	0.8802931
15%	0.9742942	0.9885898	0.8909212	0.9097902
0.5%	0.9895192	0.9936363	0.9714623	0.9748160
$P-value \in$	(0.025,0.05)	(0.025,0.05)	(0.025,0.05)	(0.5,1)

It is important to mention that all the calculations shown in this work were obtained using routines written in R. This routines uses the *sn package* and are freely available up on request.

References

[1] A. Azzalini, "*A class of distributions which includes the normal ones*", Scandinavian Journal of Statistics. 12   (1985) 171–178.

[2] A. Azzalini, "*R package sn: the skew-normal and skew-t distributions*" (version 0.4–6), Universita di Padova   (2008).

[3] A. Azzalini, A. Capitanio, "*Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution*", Journal of the Royal Statistical Society: Series B (Statistical Methodology). 65(2)   (2003) 367–389.

[4] R.D. Cook, S. Weisberg, "*An Introduction to Regression Graphics*", New York: Wiley, (1994).

[5] J. Eastman, L.J. Bain, "*A property of maximum likelihood estimators in the presence of location-scale nuisance parameters*", Commun. Statist. 2   (1973) 23–28.

[6] B.G. Leroux, "*Consistent estimation of a mixing distribution*", Annals of Statistics. 20(3)   (1992) 1350–1360.

[7] J.P. Nolan, "*Stable distributions*", University Washington, DC, (1999).

[8] P. Perez Rodrguez, J.A. Villasenor, "*On testing the skew normal hypothesis*", Journal of Statistical Planning and Inference. 140   (2010) 3148–3159.

[9] R Development Core Team, "*R: a language and environment for statistical computing*", R Foundation for Statistical Computing, Vienna, Austria; ISBN 3-900051-07-0.   (2008).

[10] M.A., "*Stephens Tests based on EDF statistics*", In: D'Agostino, R. B., Stephens, M. A., eds. Goodness-of-Fit Techniques. New York: Marcel Dekker   (1986).