On the robustness of the skew-t model

Márcia D'Elia Branco

Universidade de São Paulo
Instituto de Matemática e Estatística
mbranco@ime.usp.br

Skew Workshop 2026 - Padova

Introduction

- A general notion of robustness, is the insensitivity to small deviations from the assumptions of a model.
- This concept gets closer to outlier resistance, as it is not desirable that estimates are too affected by the presence of atypical points.
- For a long time, the Student-t distribution has been used as an alternative to the normal distribution for robustness purpose.
- Distributions with heavier tails deal better with the problem of the influence of outliers on estimates.

Introduction

- A general notion of robustness, is the insensitivity to small deviations from the assumptions of a model.
- This concept gets closer to outlier resistance, as it is not desirable that estimates are too affected by the presence of atypical points.
- For a long time, the Student-t distribution has been used as an alternative to the normal distribution for robustness purpose.
- Distributions with heavier tails deal better with the problem of the influence of outliers on estimates.
- In the skew world, it seems natural to expect that the skew-t distribution has the same robustness properties.
- For example, see Azzalini and Genton (2008) paper in *International Statistical Review.*

Robustness of the student-t

- Lange, Little and Taylor (1989) is one of the oldest references about the use of the t distribution for outlier resistance purpose.
- An analytical strategy based on maximum likelihood for a general model with independent Student-t errors is suggested and applied to a variety of problems, including linear and nonlinear regression.

Robustness of the student-t

- Lange, Little and Taylor (1989) is one of the oldest references about the use of the t distribution for outlier resistance purpose.
- An analytical strategy based on maximum likelihood for a general model with independent Student-t errors is suggested and applied to a variety of problems, including linear and nonlinear regression.
- It is important to point here the difference between the independent and the dependent (no correlation only) regression t-model (see Arellano-Valle, 1994). Only the first, has the outlier resistance property.
- The dependent t regression model has another kind of robustness property. See, p.e., Osiewalski and Steel (1993) and Breusch, Robertson and Welsh (2001).

Robustness of the student-t

- Lucas (1997) argues that only when the degrees of freedom are known the Student-t distribution is robust.
- Using the concept of M-estimator, the author proves that the influence functions for the position and scale parameters are limited when the degrees of freedom are fixed.
- On the other hand, when the degrees of freedom (ν) are estimated only the influence function of the position parameter is limited.

Robustness of the student-t

- Lucas (1997) argues that only when the degrees of freedom are known the Student-t distribution is robust.
- Using the concept of M-estimator, the author proves that the influence functions for the position and scale parameters are limited when the degrees of freedom are fixed.
- On the other hand, when the degrees of freedom (ν) are estimated only the influence function of the position parameter is limited.
- Moreover, the IF for the scale and ν are negative and decreasing. This means, for example, that the estimate of ν can be negatively biased.
- For a long time, how to estimate the degree of freedom has been a problem.

Influence function and M-estimators

- The influence function (IF) measures the impact of an infinitesimal fraction of outliers on an estimator.

$$IF(x; T, F) = \lim_{t \to 0} \frac{T[(1-t)F + t\Delta_x] - T(F)}{t}$$

- IF is an asymptotic version of the Sensitivity Curve

$$S(x_0) = T(x_1, \ldots, x_n, x_0) - T(x_1, \ldots, x_n).$$

- This measure the difference between the estimates with and without x_0, a possible outlier.

Influence function and M-estimators

- Let $\rho(\theta; x_i)$ be a differentiable function with respect to θ and $\psi(\theta, x_i) = \frac{\partial \rho(\theta; x_i)}{\partial \theta}$ a derivative vector. An estimator $\hat{\theta}$ is called an M-estimator if it satisfies

$$\sum_{i=1}^{n} \psi(\theta; x_i) = 0$$

- If we call $\rho(\theta, x_i) = -\log(f_\theta(x_i))$, the maximum likelihood estimator (MLE) is a particular case of M-estimator.
- If $\hat{\theta}$ is the MLE of a parameter vector θ then

$$IF(x; \hat{\theta}, F) = [\mathrm{B}(\hat{\theta})]^{-1} \psi(\hat{\theta}, x)$$

where B is the Fisher Information matrix and $\psi(\theta, x)$ is the negative of the score function.

My history on Skew World

- During my doctoral studies at USP, I worked with the Elliptical distribution, under the supervision of Heleno Bolfarine and Pilar Iglesias. During this time, I began my collaboration with Reinaldo Arellano-Valle.
- After finishing my doctorate, I went to UCONN for my post-doctoral studies to work with Dipak Dey.
- Chen, Dey and Shao, 1999. A New Skewed Link Model for Dichotomous Quantal Response Data. JASA.
- Azzalini and Dalla-Valle, 1997. The multivariate skew-normal distribution. Biometrika.
- Branco and Dey, 2001. A General Class of Multivariate Skew-Elliptical Distributions. JMVA.
- In 2003, I met Marc Genton at the regression school in Conservatória, Rio de Janeiro.

The skew-t distribution

- Following the proposal given by Branco and Dey (2001) and deeply discussed in Azzalini and Capitanio (2003), the multivariate Skew-t distribution is a special case of the Skew-elliptical distribution.

The skew-t distribution

- Following the proposal given by Branco and Dey (2001) and deeply discussed in Azzalini and Capitanio (2003), the multivariate Skew-t distribution is a special case of the Skew-elliptical distribution.
- The original construction given by Branco and Dey, uses the conditional method.
- Let $X = (X_0, X_1, \ldots, X_k)^T$ be an Elliptical r.v. , with some parameters and generation function g^{k+1}, then $Y = [X \mid X_0 > 0]$ has a skew-elliptical distribution.
- Notation $Y \sim SE(\xi, \Omega, \lambda; g^{k+1})$.

The skew-t distribution

- Following the proposal given by Branco and Dey (2001) and deeply discussed in Azzalini and Capitanio (2003), the multivariate Skew-t distribution is a special case of the Skew-elliptical distribution.
- The original construction given by Branco and Dey, uses the conditional method.
- Let $X = (X_0, X_1, \ldots, X_k)^T$ be an Elliptical r.v. , with some parameters and generation function g^{k+1}, then $Y = [X \mid X_0 > 0]$ has a skew-elliptical distribution.
- Notation $Y \sim SE(\xi, \Omega, \lambda; g^{k+1})$.
- A subclass of SE is the Scale Mixture of Normal of which the Skew-t is a particular case.

The skew-t distribution

- A convenient expression for the Skew-elliptical pdf is given by

$$f(y) = 2|\Omega|^{-1/2} \int_{-\infty}^{\lambda^T(y-\xi)} g^{k+1}(r^2 + q(y))dr$$

where g^{k+1} is the generation function of the Elliptical distribution and $q(y) = (y-\xi)^T\Omega^{-1}(y-\xi)$

The skew-t distribution

- A convenient expression for the Skew-elliptical pdf is given by

$$f(y) = 2|\Omega|^{-1/2} \int_{-\infty}^{\lambda^T(y-\xi)} g^{k+1}(r^2 + q(y))dr$$

  where g^{k+1} is the generation function of the Elliptical distribution and $q(y) = (y-\xi)^T\Omega^{-1}(y-\xi)$

- Considering g^{k+1} the generation function of the scale mixture of normal, we get

$$f(y) = 2\int_0^{\infty} \phi(y;\xi, K(\eta)\Omega)\Phi\left(\frac{\lambda^T(y-\xi)}{K(\eta)^{1/2}}\right) dH(\eta).$$

  where η is a mixing variable; ϕ and Φ the pdf and cdf of a normal distribution.

The skew-t distribution

- The Skew-t case follows by considering $K(\eta) = 1/\eta$ and $H(\eta)$ a Gamma distribution with both parameters equals to $\nu/2$ (ν is the degree of freedom).

The skew-t distribution

- The Skew-t case follows by considering $K(\eta) = 1/\eta$ and $H(\eta)$ a Gamma distribution with both parameters equals to $\nu/2$ (ν is the degree of freedom).
- The final expression obtained by Branco and Dey (2001) was

$$f(y) = 2f_{\nu,\tau}(y;\xi,\Omega)F_{\nu^*,\tau^*}(y;\lambda^T(y-\xi))$$

  where f and F are the pdf and the cdf of a generalized t-distribution, $\nu^* = \nu + k$ and $\tau^* = \tau + (y-\xi)^T\Omega^{-1}(y-\xi)$.
- In fact, $F_{\nu^*,\tau^*}(y;\lambda^T(y-\xi))$ is not really a cdf, but only a skewness function.
- A more convinient expression was obtained by Azzalini and Capitanio, 2003.

The univariate skew-t distribution

Folowing Azzalini and Capitanio (2003), the univariate Skew-t distribution is characterized by its probability density function:

$$f(y) = 2t(z;\nu)T\left(\alpha z\sqrt{\frac{\nu+1}{\nu+z^2}};\nu+1\right),$$

where $z = (y-\xi)/\omega$ with t and T are the pdf and cdf of the Student-t distribution.

Notation $Y \sim ST(\xi,\omega^2,\alpha,\nu)$.

The parameters are called ξ location, ω scale, α shape and ν degrees of freedom.

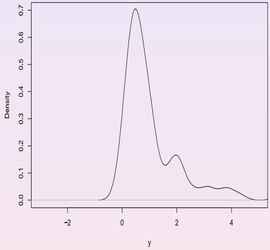
Illustrative simulated example

- The Skew-t model, as claimed before, is an interesting alternative to tackle the problem of atypical observations, because this family has parameters to control the tails.
- However, the side of outliers can cause problems. If the atypical observation is on the opposite side of the skewness, the influence on the estimates can be relevant.

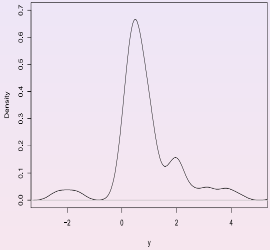
Illustrative simulated example

- The Skew-t model, as claimed before, is an interesting alternative to tackle the problem of atypical observations, because this family has parameters to control the tails.
- However, the side of outliers can cause problems. If the atypical observation is on the opposite side of the skewness, the influence on the estimates can be relevant.
- We generated 100 observations from the skew-t distribution with $\xi = 0$, $\omega = 1$, $\alpha = 5$ and $\nu = 4$, which means that the generated sample is positively skewed. Then, we introduced five contaminant points $-2.50, -2.25, -2.00, -1.75, -1.50$.

Illustrative Simulated Example



(a) *Sample with 100 observations*



(b) *Inclusion of 5 contaminants*

Figure 2.3: *Empirical densities for a sample of 100 observations of* $ST(0, 1, 5, 4)$ *and a contaminated sample.*

Illustrative Simulated Example

Tabela: Estimates (and standard errors) for parameters and quantities in the original 100 sample of $ST(0, 1, 5, 4)$ and the contaminated sample (with 105 observations). Mean= 0.93.

	Without contaminants		**With contaminants**	
	Skew-normal	**Skew-t**	**Skew-normal**	**Skew-t**
ξ	-0.050 (0.061)	0.082 (0.071)	-0.283 (0.177)	0.279 (0.085)
ω	1.634 (0.123)	0.815 (0.165)	1.803 (0.174)	0.511 (0.099)
α	16.417 (9.149)	5.859 (2.857)	1.899 (0.431)	1.350 (0.518)
ν	–	2.169 (0.679)	–	1.244 (0.248)
Mean*	1.251 (0.096)	0.948 (0.104)	0.990 (0.125)	0.810 (0.092)

- The presence of outliers diminished the estimates of α.

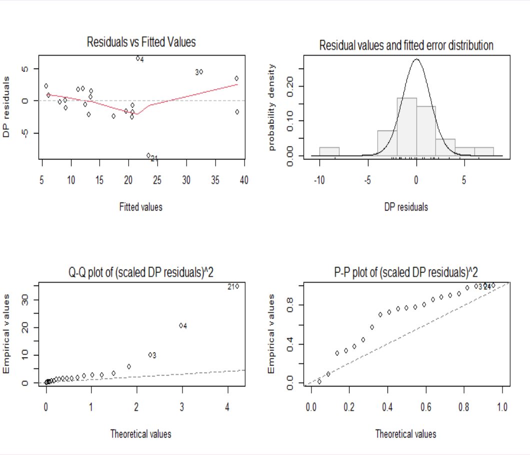
Real data set: stack-loss

- The stack-loss data (Dodge, 1996) refers to n=21 days of observations on a chemical process. The author called this data "the Guinea Pig"for robustness purpose.
- The interest variable is $y = Stack\ loss$, with explanatory variables $x_1 = Air\ flow$, $x_2 = Water\ temperature$ and $x_3 = Acid\ concentration$.
- This data set was used in Azzalini and Genton (2008) to show that the skew-t linear model has a satisfactory behavior with respect to other robust methods.

Real data set: stack-loss

- The stack-loss data (Dodge, 1996) refers to n=21 days of observations on a chemical process. The author called this data "the Guinea Pig"for robustness purpose.
- The interest variable is $y = Stack\ loss$, with explanatory variables $x_1 = Air\ flow$, $x_2 = Water\ temperature$ and $x_3 = Acid\ concentration$.
- This data set was used in Azzalini and Genton (2008) to show that the skew-t linear model has a satisfactory behavior with respect to other robust methods.
- They obtained a very low value for the shape parameter estimate, $\hat{\alpha} = 0.28$, and conclude that there is no asymmetry.
- The residuals plot, using the normal errors are presented on the next page.

Stack-Loss data - residuals with normal error

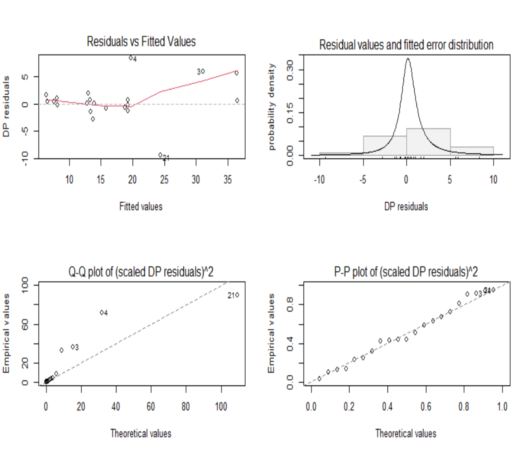


Stack-Loss data

- There are at least three evident outliers, with one of them on the opposite side of the asymmetry.
- If we remove this observation, the new shape estimate is $\hat{\alpha} = 1.74$, showing a right skewness for the remaining data.
- The Table shows the estimates with and without the outlier (point 21).

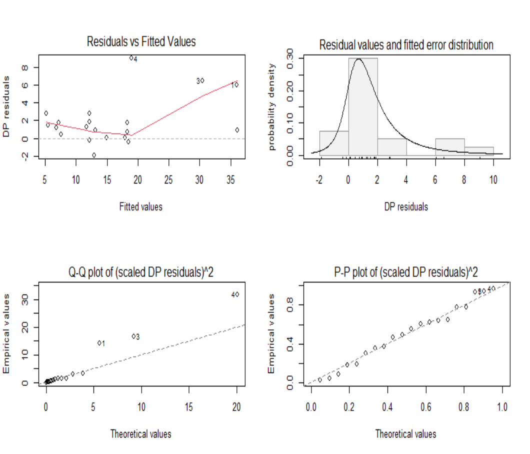
Parameter	Complete data	Without outlier
Scale	0.98	1.60
Shape	0.28	1.74
Df	1.14	1.97

Complete data



Stack-Loss data - residuals with skew-t error

Data without observation 21



Influence function of the skew-t distribution

The IF depends on the behavior of the score function, it is proportional to the score function with opposite signal.

Consider $z_* = z + \epsilon$ with $z = (y - \xi)/\omega$.
The following results were obtained by Harnik (2023):

$$
\begin{aligned}
\lim_{\epsilon \to \infty} \frac{\partial l(\theta, z_*)}{\partial \xi} &= 0, \\
\lim_{\epsilon \to \infty} \frac{\partial l(\theta; z_*)}{\partial \omega} &= \frac{\nu}{\omega}, \\
\lim_{\epsilon \to \infty} \frac{\partial l(\theta; z_*)}{\partial \alpha} &= (\nu + 1)^{1/2} \frac{t(\alpha(\nu + 1)^{1/2}; \nu + 1)}{T(\alpha(\nu + 1)^{1/2}; \nu + 1)}
\end{aligned}
$$

Influence function of the skew-t distribution

- The score function (also the IF) are limited to location, scale and shape parameters.

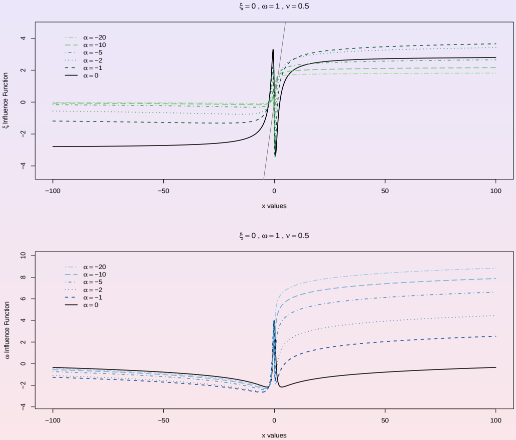
Influence function of the skew-t distribution

- The score function (also the IF) are limited to location, scale and shape parameters.
- However, the score function related to the shape parameter depends on α and the ratio of the pdf and cdf of the Student-t distribution.

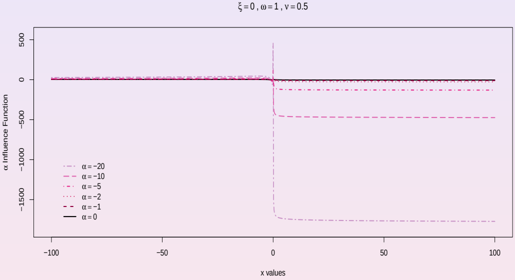
Influence function of the skew-t distribution

- The score function (also the IF) are limited to location, scale and shape parameters.
- However, the score function related to the shape parameter depends on α and the ratio of the pdf and cdf of the Student-t distribution.
- When α is negative, the denominator tends to a quantity that is close to zero for small values of α and the $IF(\alpha)$ tends to be high.
- When α is positive, the denominator tends to 1 for big values of α and the $IF(\alpha)$ gets under control.

Influence function behavior for location and scale



Influence function behavior for the shape



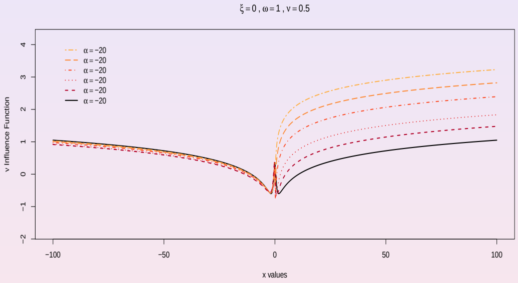
Influence function of the skew-t distribution

To complete the result, we study the behavior of the score function for the degree of freedom, showing that it is not limited.

$$\lim_{\epsilon\to\infty} \frac{\partial l(\theta; z^*)}{\partial \nu} = K + \frac{1}{2}\left\{\lim_{\epsilon\to\infty}\left[-\log\left(1+\frac{(z^*+\epsilon)^2}{\nu}\right)\right]\right\} = -\infty.$$

where K does not depend on ϵ.

Influence function behavior for degree of freedom



Conclusions

- The behavior of the $IF(\alpha)$ shows a higher influence to one extreme observation at the opposite side of the asymmetry. Usually α will be underestimated.

Conclusions

- The behavior of the $IF(\alpha)$ shows a higher influence to one extreme observation at the opposite side of the asymmetry. Usually α will be underestimated.
- The behavior of the $IF(\nu)$ shows a negative bias in the estimates of ν, similar to what occurs in the symmetric model. However, this is even worse when the outlier is in the opposite side of the asymmetry.

Conclusions

- The behavior of the $IF(\alpha)$ shows a higher influence to one extreme observation at the opposite side of the asymmetry. Usually α will be underestimated.
- The behavior of the $IF(\nu)$ shows a negative bias in the estimates of ν, similar to what occurs in the symmetric model. However, this is even worse when the outlier is in the opposite side of the asymmetry.
- This bad behavior of the influence function is not a particularity of the Skew-t distribution. As shown by Harnik (2023), this is extended to the class of distribution with asymmetry generator mechanism given by $fy(y) = 2f_0(x)G(w(x, \alpha))$.

Weighted Likelihood Estimator

- Harnik (2023) proposed some new estimators to deal with the outliers influence under the skew-t distribution.
- None of them proved completely satisfactory. However, the best was the Weighted Likelihood Estimator (WLE).

Weighted Likelihood Estimator

- Harnik (2023) proposed some new estimators to deal with the outliers influence under the skew-t distribution.
- None of them proved completely satisfactory. However, the best was the Weighted Likelihood Estimator (WLE).
- Agostinelli and Greco (2012) developed an approach to reach robust estimation by introduces a set of weights into the original likelihood function with the aim of lowering the influence of extreme observations.
- The log weighted likelihood is given by $\sum_{i=1}^{n} w(x_i) l(\theta, x_i)$. The challenge is to specify the weights $w(x_i)$.
- We applied a strategy proposed by Majumder et al. (2021) (for details see Harnik, 2023).

Weighted Likelihood Estimator

The general idea to define the weights is the following

(i) Choose a proportion $0 < p < 0.5$ of observations that should have their weight reduced in the final estimate.

(ii) For these observations, the weight are a function of the rate between the empirical and theoretical cdf. Such that, less weight will be given to values where these functions differ more.

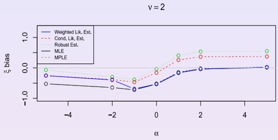
(iii) Assign weight one to the remaining observations.

Weighted Likelihood Estimator

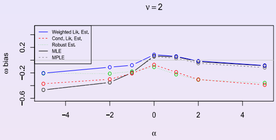
- A simulation study considering 90 observations from the skew-t distribution with 10 contaminants in the right tail was carried out.
- The location and scale are fixed in zero and one, respectively. We created 28 scenarios varying α and ν.
- The Figure shows the median bias for the estimators varying α for the case $\nu = 2$. The blue lines represent the WLE and the black lines the MLE.

Weighted Likelihood Estimator

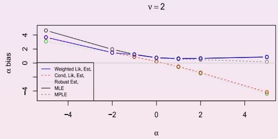
- A simulation study considering 90 observations from the skew-t distribution with 10 contaminants in the right tail was carried out.
- The location and scale are fixed in zero and one, respectively. We created 28 scenarios varying α and ν.
- The Figure shows the median bias for the estimators varying α for the case $\nu = 2$. The blue lines represent the WLE and the black lines the MLE.
- For all parameters we can see the superiority of the WLE, when $\alpha < 0$. When $\alpha > 0$, the WLE bias is very similar to the MLE.
- In general, we can observe a clear negative bias for ν. One the other hand, for α we note a positive bias.



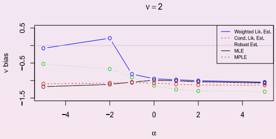
(a) *Bias for* ξ



(b) *Bias for* ω



(c) *Bias for* α



(d) *Bias for* ν

Figure 4.3: *Median bias for different estimators (Weighted Likelihood Estimator, CLE, Robust Measures Estimator, MLE and MPLE) for skew-t distribution with* $\xi = 0$, $\omega = 1$ *and different values of* α *and* ν. *The case of* $\alpha = 0$ *corresponds to the Student-t(0,1).*

Additional Results

- Under the Bayesian approach, O'Hagan (1979) brought the ideas of outlier-prone referring to the processes that generated the data.
- Suppose $x_1, \ldots, x_n$ a random sample from a location family $f(x - \theta)$. The density $f(.)$ is said to be right outlier-prone of order n, if for $x_{n+1} \to \infty$,

$$P(\theta \leq c \mid x_1, \ldots, x_n, x_{n+1}) \to P(\theta \leq c \mid x_1, \ldots, x_n)$$

- He proves that the Student-t distribution is outlier-prone for the location parameter and the normal is not.

Additional Results

- We showed that the Skew-t distribution is also outlier-prone for the location parameter, as long as, the others parameters are fixed.
- However, in general, the others parameters are unknown and estimated. In this case, the Skew-t distribution loses its property.

Final Comments

- For one-dimensional data, such as those explored in this presentation, a simple Exploratory Data Analysis can be very helpful in identifying the side of the outliers and determining the best strategy to use for modeling and estimation.
- However, for the multi-dimensional case, it is not so easy. In this sense, we believe it is very important to generalize the results obtained here to the multivariate case.

Final Comments

- For one-dimensional data, such as those explored in this presentation, a simple Exploratory Data Analysis can be very helpful in identifying the side of the outliers and determining the best strategy to use for modeling and estimation.
- However, for the multi-dimensional case, it is not so easy. In this sense, we believe it is very important to generalize the results obtained here to the multivariate case.
- My last question: Is estimating the α and ν parameters in the skew-t model still a problem or not?

References

1. Azzalini and Genton, 2008. International Statistical Review.
2. Lange, Little and Taylor, 1989. JASA.
3. Arellano-Valle, 1994. Thesis, IME-USP.
4. Osiewalski and Steel, 1993. Journal of Econometrics.
5. Breusch, Robertson and Welsh, 2001. Statistica Neerlandica.
6. Lucas, 1997. Communications in Statistics - Theory Methods.
7. Branco and Dey, 2001. JMVA.
8. Azzalini and Capitanio, 2003. JRSS-B.
9. Dodge, 1996. In Robust Statistics, Data Analysis and Computer Intensive Methods.
10. Harnik, 2023. Thesis, IME-USP.

References

1. Agostinelli and Greco, 2012. In Proceedings of the 46th Scientific Meeting of the Italian Statistical Society.
2. Majumder et al. ,2021. Metrika.
3. O'Hagan, 1979. JRSS-B.