Nonlinear Regression Models Based on Scale Mixtures of Skew-Normal Distributions

Aldo M. Garay[1], Víctor H. Lachos*,[1]

Departamento de Estatística, Universidade Estatual de Campinas, Rua Sérgio Buarque de Holanda, 651, Cidade Universitária Zeferino Vaz, Campinas, São Paulo, Brazil. CEP 13083-859 – Caixa Postal 6065

C. A. Abanto-Valle[2]

Departamento de Estatística, Universidade Federal de Rio de Janeiro, Caixa Postal 68530, CEP: 21945-970, Rio de Janeiro-RJ, Brazil

Abstract

An extension of some standard likelihood based procedures to nonlinear regression models under scale mixtures of skew-normal distributions is developed. This novel class of models provides a useful generalization of the symmetrical nonlinear regression models since the error distributions cover both skewness and heavy–tailed distributions such as the skew-t, skew-slash and the skew-contaminated normal distributions. The main advantage of these class of distributions is that they have a nice hierarchical representation which allows easy implementation of inference. A simple EM-type algorithm for iteratively computing maximum likelihood estimates is presented and the observed information matrix for obtaining the asymptotic covariance matrix is derived analytically. With the aim of identifying atypical observations and/or model misspecification a brief discussion of the standardized residuals is given. Finally, an illustration of the methodology is given considering a data set previously analyzed under skew-normal nonlinear regression models. Our analysis indicates that a skew-t nonlinear regression model with 3 degrees of freedom seems to fit the data better than the skew-normal nonlinear regression model as well as other asymmetrical nonlinear models in the sense of robustness against outlying observations.

Key words: EM algorithm, Skew-normal distribution, Scale mixtures of skew-normal distributions, Nonlinear regression models.

1. Introduction

Normal nonlinear regression models (N-NLM) are usually applied in sciences and engineering to model symmetrical data for which nonlinear functions of unknown parameters are used in order

*Corresponding author

Email addresses: amedina@ime.unicamp.br (Aldo M. Garay), hlachos@ime.unicamp.br (Víctor H. Lachos), cabantovalle@im.ufrj.br (C. A. Abanto-Valle)

[1]The authors acknowledge the partial financial support from Fundação de Amparo à Pesquisa do Estado de São Paulo and CNPq

[2]The author acknowledges the partial financial support from FAPERJ

to explaining or describing the phenomena under study. But N-NLM suffers from the same lack of robustness against departures from distributional assumptions as other statistical models based on the Gaussian distribution and may be too restrictive to provide an accurate representation of the structure that is present in the data. To deal with this problem, some proposals have been made in the literature by replacing the assumption of normality by a class of symmetrical distributions that cover both light-and heavy-tailed distributions such as Student-t, logistic, power exponential, (see Cysneiros and Vanegas, 2008; Cordeiro et al., 2009, among others). Recently, Cancho et al. (2009) and Xie et al. (2009) have shown the advantage of using the skew-normal distribution in the context of nonlinear regression models (SN-NLM). In this article, we extend the SN-NLM by assuming that the models errors follow scale mixtures of skew-normal distributions– hereafter SMSN (Branco and Dey, 2001)– which deal simultaneously with skewness and heavy-tails. Interestingly, this rich class contains the entire family of scale mixtures of normal distributions (Lange and Sinsheimer, 1993). In addition, the skew–normal (SN) and skewed versions of some classical symmetric distributions are SMSN members: for example, The skew–t (ST), the skew–slash (SSL) and the skew contaminated normal (SCN). They seem to be a reasonable choice for robust inference and some of the advantages of our approach are to offer efficient algorithms to model estimation and the practical interpretation of the parameters.

The rest of the paper is organized as follows. In Section 2, we present some properties of the univariate SMSN family. Section 3 outlines the asymmetric model as well as some inferential results. In Section 4 an EM-type algorithm for maximum likelihood estimation is developed. Additionally, some model selection criteria and the use of standardized residuals in these asymmetrical nonlinear models are discussed. Finally, in Section 5, we illustrate the methodology considering an application with a real data set.

2. Scale mixtures of skew-normal distributions

2.1. Preliminaries

First, we make some remarks about the class of scale mixtures of skew-normal distributions, as introduced by Branco and Dey (2001); see also Arellano-Valle et al. (2006).

As defined by Azzalini (1985), a random variable Z has skew-normal distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ, if its density is given by

$$f(z) = 2\phi(z;\mu,\sigma^2)\Phi\left(\frac{\lambda(z-\mu)}{\sigma}\right), \tag{1}$$

where $\phi(\cdot;\mu,\sigma^2)$ denotes the density of the univariate normal distribution with mean μ and variance $\sigma^2 > 0$ and $\Phi(\cdot)$ is the distribution function of the standard univariate normal distribution. We denote it by $Z \sim SN(\mu,\sigma^2,\lambda)$.

Let $Z \sim SN(0,\sigma^2,\lambda)$. A random variable Y is in the SMSN family if it can be written as

$$Y = \mu + \kappa^{1/2}(U)Z, \tag{2}$$

where μ is a location parameter, $\kappa(u)$ is a positive function of u, U is a random variable with distribution function $H(\cdot;\boldsymbol{\nu})$ and density $h(\cdot;\boldsymbol{\nu})$ and $\boldsymbol{\nu}$ is a scalar or vector parameter indexing the distribution of U. Although we can deal with any κ function, in this paper we restrict our attention to the case in that $\kappa(u) = 1/u$, since it leads to good mathematical properties.

The name of the class becomes clear when we note that the conditional distribution of Y given $U = u$ is skew-normal. Specifically, we have that $Y|U = u \sim SN(\mu, u^{-1}\sigma^2, \lambda)$. Thus, the density of Y is given by

$$f(y) = 2\int_0^\infty \phi(y;\mu,u^{-1}\sigma^2)\Phi\left(\frac{u^{1/2}\lambda(y-\mu)}{\sigma}\right) dH(u;\boldsymbol{\nu}), \tag{3}$$

that is, $f(\cdot)$ is an infinite mixture of skew-normal densities, being U the *scale factor* and its distribution function $H(\cdot;\boldsymbol{\nu})$, the *mixing distribution*.

We use the notation $Y \sim SMSN(\mu,\sigma^2,\lambda;H)$. When H is degenerate, with $u = 1$, we obtain the $SN(\mu,\sigma^2,\lambda)$ distribution.

2.2. Moments

Arnold et al. (1993) show an interesting method of moments to obtain estimators with closed form expressions for a skew–normal random variable. In this section we extend their method to obtain the moments estimators of the parameters of a SMSN distribution. First, we present the following result

Lemma 1. *Let* $Y \sim SMSN(\mu,\sigma^2,\lambda;H)$.

a) If $E[U^{-1/2}] < \infty$, *then* $E[Y] = \mu + \sqrt{\frac{2}{\pi}}k_1\Delta$;

b) If $E[U^{-1}] < \infty$, *then* $Var[Y] = \sigma^2 k_2 - \frac{2}{\pi}k_1^2\Delta^2$;

where $\Delta = \sigma\delta$, $\delta = \dfrac{\lambda}{\sqrt{1+\lambda^2}}$ *and* $k_m = E[U^{-m/2}]$.

For $\boldsymbol{\nu}$ fixed, from Lemma 1, we can find the moments estimator of $\boldsymbol{\theta} = (\mu,\sigma^2,\delta)^\top$, which we denote by $\widetilde{\boldsymbol{\theta}} = (\widetilde{\mu},\widetilde{\sigma}^2,\widetilde{\delta})^\top$. It is given by

$$\begin{aligned} M_3(k_2 - \frac{2}{\pi}k_1^2\widetilde{\delta}^2)^{3/2} &= (M_2)^{3/2}(a_1 + a_2\widetilde{\delta}^2)\widetilde{\delta}, \\ \widetilde{\sigma}^2 &= \frac{M_2}{(k_2 - \frac{2}{\pi}k_1^2\widetilde{\delta}^2)} \end{aligned}$$

and

$$\widetilde{\mu} = M_1 - k_1\sqrt{\frac{2}{\pi}}\widetilde{\sigma}\widetilde{\delta},$$

where $a_1 = 3\sqrt{\frac{2}{\pi}}(k_3 - k_1k_2)$, $a_2 = 2(\frac{2}{\pi})^{3/2}k_1^3 - \sqrt{\frac{2}{\pi}}k_3$, $M_1 = \frac{1}{n}\sum_{i=1}^n y_i$, $M_2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$ and $M_3 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^3$. Although we do not have a closed form expression for $\widetilde{\delta}$, we can apply some computational procedures (such as the Newton-Raphson method) to obtain numerical solutions. However, when $U = 1$, the equations above reduce to the equations obtained by Arnold et al. (1993); see also Lin et al. (2007b).

For a SMSN random variable Y, a convenient stochastic representation is given next. It can be used to simulate realizations of Y, to implement the EM algorithm and also to study some of its properties. The proof follows easily from Henze (1986) and the stochastic representation given in (2).

Lemma 2. *A random variable $Y \sim SMSN(\mu, \sigma^2, \lambda; H)$ has a stochastic representation given by*

$$Y = \mu + \Delta T + U^{-1/2}\Gamma^{1/2}T_1,$$

where $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$, $\Delta = \sigma\delta$, $\Gamma = (1-\delta^2)\sigma^2$, $T = U^{-1/2}|T_0|$, T_0 and T_1 are independent standard normal random variables and $|\cdot|$ denotes absolute value.

2.3. Examples of SMSN distributions

In this section we consider some particular cases of SMSN distributions. For each SMSN distribution, we compute the conditional expectations

$$\kappa_r = E[U^r|y], \quad \tau_r = E[U^{r/2}W_\Phi(U^{1/2}A)|y],$$

where $A = \dfrac{\lambda(y-\mu)}{\sigma}$ and $W_\Phi(x) = \phi(x)/\Phi(x)$, $x \in \mathbb{R}$. These quantities will be useful when implementing the EM algorithm.

- *The skew–t distribution with ν degrees of freedom.* In this case we consider $U \sim Gamma(\nu/2, \nu/2)$, $\nu > 0$, in definition (2) – where $Gamma(a, b)$ denotes the gamma distribution with mean a/b. The density of Y takes the form

$$f(y) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{d}{\nu}\right)^{-\frac{\nu+1}{2}} T\left(\sqrt{\frac{v+1}{d+\nu}}A; \nu+1\right), \quad y \in \mathbb{R}, \tag{4}$$

  where $d = (y-\mu)^2/\sigma^2$ and $T(\cdot;\nu)$ denotes the distribution function of the standard Student–t distribution, with location zero, scale one and ν degrees of freedom, namely $t(0, 1, \nu)$. We use the notation $Y \sim ST(\mu, \sigma^2, \lambda; \nu)$. A particular case of the skew-t distribution is the skew–Cauchy distribution, when $\nu = 1$. Also, when $\nu \to \infty$, we get the skew-normal distribution as the limiting case.

  We have that

$$k_m = \left(\frac{\nu}{2}\right)^{m/2}\frac{\Gamma(\frac{\nu-m}{2})}{\Gamma(\frac{\nu}{2})}.$$

  Thus, from Proposition 1 in Lachos et al. (2009), we obtain

$$\kappa_r = \frac{2^{r+1}\nu^{\nu/2}\Gamma(\frac{\nu+2r+1}{2})(d+\nu)^{-\frac{\nu+2r+1}{2}}}{f(y)\Gamma(\nu/2)\sqrt{\pi}\sigma} T(\sqrt{\frac{\nu+2r+1}{d+\nu}}A; \nu+2r+1)$$

  and

$$\tau_r = \frac{2^{(r+1)/2}\nu^{\nu/2}\Gamma(\frac{\nu+r+1}{2})(d+\nu+A^2)^{-\frac{\nu+r+1}{2}}}{f(y)\Gamma(\nu/2)\sqrt{\pi}^2\sigma}.$$

  Applications of the skew–t distribution in robust estimation can be found in Lin et al. (2007a) and Azzalini and Genton (2008).

- *The skew–slash distribution.* In this case we have $U \sim Beta(\nu, 1)$ – where $Beta(a, b)$ denotes the beta distribution with parameters a and b – with positive shape parameter ν, and use the notation $Y \sim SSL(\mu, \sigma^2, \lambda; \nu)$. The density of Y is given by

$$f(y) = 2\nu \int_0^1 u^{\nu-1} \phi(y; \mu, u^{-1}\sigma^2)\Phi(u^{1/2}A)du, \quad y \in \mathbb{R}, \tag{5}$$

and we have that

$$k_m = \frac{2\nu}{2\nu - m}, \quad \nu > m/2.$$

In this case, the conditional expectations are given by

$$\kappa_r = \frac{2^{\nu+r+1}\nu\Gamma(\frac{2\nu+2r+1}{2})P_1\left(\frac{2\nu+2r+1}{2}, \frac{d}{2}\right) d^{-\frac{2\nu+2r+1}{2}}}{f(y)\sqrt{\pi}\sigma} E[\Phi(S^{1/2}A)]$$

and

$$\tau_r = \frac{2^{\nu+r/2+1/2}\nu\Gamma(\frac{2\nu+r+1}{2})}{f(y)\sqrt{\pi}^2\sigma}(d + A^2)^{-\frac{2\nu+r+1}{2}} P_1\left(\frac{2\nu+r+1}{2}, \frac{d+A^2}{2}\right),$$

where $P_x(a, b)$ denotes the distribution function of the $Gamma(a, b)$ distribution evaluated at x and $S \sim Gamma\left(\frac{2\nu+2r+1}{2}, \frac{d}{2}\right)\mathbb{I}_{(0,1)}$, a truncated gamma distribution on $(0, 1)$, with the parameters values in parenthesis before truncation. The skew-slash is a heavy-tailed distribution having as limiting distribution the skew-normal one (when $\nu \to \infty$). Applications can be found in (Wang and Genton, 2006).

- *The skew contaminated normal distribution.* Here U is a discrete random variable taking one of two states. The probability function of U is given by

$$h(u|\boldsymbol{\nu}) = \nu\mathbb{I}_{(u=\gamma)} + (1 - \nu)\mathbb{I}_{(u=1)}, \;\; 0 < \nu < 1, \, 0 < \gamma \leq 1,$$

where $\boldsymbol{\nu} = (\nu, \gamma)^\top$. We denote it by $Y \sim SCN(\mu, \sigma^2, \lambda; \nu, \gamma)$. Also, we have

$$k_m = \frac{\nu}{\gamma^{m/2}} + 1 - \nu.$$

It follows immediately that

$$f(y) \;\; = \;\; 2\{\nu\phi(y; \mu, \gamma^{-1}\sigma^2)\Phi(\gamma^{1/2}A) + (1 - \nu)\phi(y; \mu, \sigma^2)\Phi(A)\}.$$

The parameters ν and γ can be interpreted as the proportion of outliers and a scale factor, respectively. The skew contaminated normal distribution reduces to the skew-normal distribution when $\gamma = 1$. In this case, we have that

$$\kappa_r = \frac{2}{f(y)}[\nu\gamma^r\phi_1(y; \mu, \gamma^{-1}\sigma^2)\Phi(\gamma^{1/2}A) + (1 - \nu)\phi(y; \mu, \sigma^2)\Phi(A)]$$

and

$$\tau_r = \frac{2}{f(y)}[\nu\gamma^{r/2}\phi(y; \mu, \gamma^{-1}\sigma^2)\phi(\gamma^{1/2}A) + (1 - \nu)\phi(y; \mu, \sigma^2)\phi(A)].$$

3. The SMSN nonlinear regression model

The nonlinear regression model based on SMSN distributions–hereafter SMSN-NLM– is defined as

$$Y_i = \eta(\boldsymbol{\beta}, \mathbf{x}_i) + \varepsilon_i, \;\; i = 1, \ldots, n, \tag{6}$$

where the Y_i are responses, $\eta(.)$ is an injective and twice continuously differentiable function with respect to the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\mathbf{x}_i$ is a vector of explanatory variable values and the random errors $\varepsilon_i \sim SMSN(-\sqrt{\frac{2}{\pi}}k_1\Delta, \sigma^2, \lambda; H)$ that corresponds to the regression model where the error distribution has mean zero. When they exist, from Lemma 1, we have that

$$E[Y_i] = \eta(\boldsymbol{\beta}, \mathbf{x}_i), \; Var[Y_i] = k_2\sigma^2 - b^2\Delta^2,$$

where $b = -\sqrt{\frac{2}{\pi}}k_1$ and $Y_i \sim SMSN(\eta(\boldsymbol{\beta}, \mathbf{x}_i) + b\Delta, \sigma^2, \lambda; H)$, for $i = 1, \ldots, n$. In order to avoid difficulties in estimating the parameter $\boldsymbol{\nu}$ of the mixing variable, we fixed it previously, as recommended by Lange et al. (1989) and Berkane et al. (1994).

The log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \lambda)^\top$ given the observed sample $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is given by $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})$, where

$$\ell_i(\boldsymbol{\theta}) = \log 2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 + \log K_i,$$

with

$K_i = \int_0^\infty u_i^{1/2} \exp\{-\frac{1}{2}u_i d_i\}\Phi(u_i^{1/2}A_i)dH(u_i)$ and $d_i = (y_i - \eta(\boldsymbol{\beta}, \mathbf{x}_i) - b\Delta)^2/\sigma^2$ and $A_i = d_i^{1/2}\lambda$. The score function is given by $U(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} U_i(\boldsymbol{\theta})$, where $U_i(\boldsymbol{\theta}) = \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \big(U_i(\boldsymbol{\beta})^\top, U_i(\sigma^2), U_i(\lambda)\big)^\top$ and $U_i(\boldsymbol{\gamma})$, for $\boldsymbol{\gamma} = \boldsymbol{\beta}, \sigma^2$ or λ, has the form

$$U_i(\boldsymbol{\gamma}) = \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}} = -\frac{1}{2}\frac{\partial \log\sigma^2}{\partial \boldsymbol{\gamma}} + \frac{1}{K_i}\frac{\partial K_i}{\partial \boldsymbol{\gamma}}, \;\; i = 1, \ldots, n, \tag{7}$$

where

$$\frac{\partial K_i}{\partial \boldsymbol{\gamma}} = I_i^{\phi}(1)\frac{\partial A_i}{\partial \boldsymbol{\gamma}} - \frac{1}{2}I_i^{\Phi}\left(\frac{3}{2}\right)\frac{\partial d_i}{\partial \boldsymbol{\gamma}}$$

and the observed information matrix $\mathbf{J}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} = -\sum_{i=1}^{n}\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}$, have elements given by $J_{\gamma\tau} = -\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\tau}^\top}$, for $\boldsymbol{\gamma}, \boldsymbol{\tau} = \boldsymbol{\beta}, \sigma^2$ or λ, where

$$\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\tau}^\top} = -\frac{1}{2}\frac{\partial^2 \log\sigma^2}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\tau}^\top} - \frac{1}{K_i^2}\frac{\partial K_i}{\partial \boldsymbol{\gamma}}\frac{\partial K_i}{\partial \boldsymbol{\tau}^\top} + \frac{1}{K_i}\frac{\partial^2 K_i}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\tau}^\top},$$

and

$$\begin{aligned}\frac{\partial^2 K_i}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\tau}^\top} \;=\; & \frac{1}{4}I_i^{\Phi}\left(\frac{5}{2}\right)\frac{\partial d_i}{\partial \boldsymbol{\gamma}}\frac{\partial d_i}{\partial \boldsymbol{\tau}^\top} - \frac{1}{2}I_i^{\Phi}\left(\frac{3}{2}\right)\frac{\partial^2 d_i}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\tau}^\top} - \frac{1}{2}I_i^{\phi}(2)\left(\frac{\partial A_i}{\partial \boldsymbol{\gamma}}\frac{\partial d_i}{\partial \boldsymbol{\tau}^\top} + \frac{\partial d_i}{\partial \boldsymbol{\gamma}}\frac{\partial A_i}{\partial \boldsymbol{\tau}^\top}\right) \\ & -I_i^{\phi}(2)A_i\frac{\partial A_i}{\partial \boldsymbol{\gamma}}\frac{\partial A_i}{\partial \boldsymbol{\tau}^\top} + I_i^{\phi}(1)\frac{\partial^2 A_i}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\tau}^\top},\end{aligned}$$

with,

$$I_i^{\Phi}(w) = \int_0^{\infty} u_i^w \exp\left(-\frac{1}{2}u_i d_i\right) \Phi_1(u_i^{1/2} A_i) dH(u_i)$$

and

$$I_i^{\phi}(w) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} u_i^w \exp\left(-\frac{1}{2}u_i(d_i + A_i^2)\right) dH(u_i).$$

Notice that we can also write $K_i = I_i^{\Phi}(\frac{1}{2})$. Direct substitution of H in the integrals above yields immediately the following results for each distribution considered, viz.,

- *Skew–t:*

$$\begin{aligned} I_i^{\Phi}(w) &= \frac{2^w \nu^{\nu/2} \Gamma(w + \nu/2)}{\Gamma(\nu/2)(\nu + d_i)^{\nu/2+w}} T\left(\sqrt{\frac{\nu + 2w}{\mathrm{d_i} + \nu}}\, \mathrm{A}_i; \nu + 2w\right) \text{ and} \\ I_i^{\phi}(w) &= \frac{2^w \nu^{\nu/2} \Gamma(\frac{\nu+2w}{2})}{\sqrt{2\pi}\Gamma(\nu/2)(d_i + \mathrm{A}_i^2 + \nu)\sqrt{\frac{\nu + 2w}{2}}}. \end{aligned}$$

- *Skew–slash:*

$$\begin{aligned} I_i^{\Phi}(w) &= \frac{\nu 2^{\nu+w} \Gamma(\nu + w)}{d_i^{\nu+w}} P_1\left(\nu + w, \frac{d_i}{2}\right) E\{\Phi(S_i^{1/2} \mathrm{A}_i)\} \text{ and} \\ I_i^{\phi}(w) &= \frac{\nu 2^{\nu+w} \Gamma(\nu + w)}{\sqrt{2\pi}(d_i + \mathrm{A}_i^2)^{\nu+w}} P_1\left(\nu + w, \frac{d_i + \mathrm{A}_i^2}{2}\right), \end{aligned}$$

  where $S_i \sim Gamma(\nu + w, \frac{d_i}{2})\mathbb{I}_{(0,1)}$.

- *Skew contaminated normal:*

$$\begin{aligned} I_i^{\Phi}(w) &= \sqrt{2\pi}\{\nu\gamma^{w-1/2}\phi_1\left(\sqrt{d_i}|0, \frac{1}{\gamma}\right)\Phi(\gamma^{1/2} A_i) + (1-\nu)\phi_1(\sqrt{d_i}|0,1)\Phi(A_i)\} \text{ and} \\ I_i^{\phi}(w) &= \nu\gamma^{w-1/2}\phi_1\left(\sqrt{d_i + A_i^2}|0, \frac{1}{\gamma}\right) + (1-\nu)\phi_1\left(\sqrt{d_i + A_i^2}|0,1\right). \end{aligned}$$

The derivatives of d_i and A_i involves standard algebraic manipulations and are not given here. Note that since one has a closed-form expression for the observed information matrix for $\boldsymbol{\theta}$, the Newton-Raphson method can be easily applied to get the ML estimates . In the next section we discuss a technique more elaborate to find the ML estimates of the parameters vector $\boldsymbol{\theta}$ based on an EM-type algorithm.

4. Parameter estimation via the EM-algorithm

In this subsection we develop an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for maximum likelihood estimation of the parameters of SMSN-NLM. In order to do this, we first represent the SMSN-NLM in an incomplete data framework using the result presented in Lemma 2. We consider the following hierarchical representation for Y_i

$$Y_i|T_i = t_i \sim N_1(\eta(\boldsymbol{\beta}, \mathbf{x}_i) + \Delta t_i, U_i^{-1}\Gamma), \tag{8}$$

$$T_i|U_i \sim TN_1(b, u_i^{-1})I(b, \infty), \tag{9}$$

$$U_i \sim H(.; \boldsymbol{\nu}) \tag{10}$$

where

$$\Gamma = (1-\delta^2)\sigma^2, \qquad \Delta = \sigma\delta \tag{11}$$

and $TN_1(r,s)$ denotes the truncated univariate normal distribution on (r,s), with parameters values in parenthesis before truncation. An useful straightforward result is that the conditional distribution of T_i given y_i and u_i is $TN_1(\mu_{Ti}+b, u_i^{-1}M_T^2)I(b,\infty)$, with

$$M_T^2 = \frac{\Gamma}{\Delta^2+\Gamma},\ \mu_{T_i} = \frac{\Delta}{\Delta^2+\Gamma}(y_i - \eta(\boldsymbol{\beta},\mathbf{x}_i) - \Delta b)$$

Now we proceed for the E-step of the algorithm. To represent the estimator of the parameter $\xi = g(\boldsymbol{\theta})$, we will use the general notation $\widehat{\xi} = g(\widehat{\boldsymbol{\theta}})$, where $g(\cdot)$ is a generic function of $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \lambda)^\top$. Thus, let $\mathbf{y} = (y_1,\ldots,y_n)^\top$, $\mathbf{t} = (t_1,\ldots,t_n)^\top$ and $\mathbf{u} = (u_1,\ldots,u_n)^\top$. It follows that the complete log-likelihood function associated with $(\mathbf{y},\mathbf{t},\mathbf{u})$ is given by

$$\ell_c(\boldsymbol{\theta}|\mathbf{y},\mathbf{t},\mathbf{u}) = c - \frac{n}{2}\log\Gamma - \frac{1}{2\Gamma}\sum_{i=1}^{n} u_i(y_i - \eta(\boldsymbol{\beta},\mathbf{x}_i) - \Delta t_i)^2, \tag{12}$$

where c is a constant that is independent of $\boldsymbol{\theta}$. Letting $\widehat{u}_i = E[U_i|\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, y_i]$, $\widehat{ut}_i = E[U_i t_i|\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, y_i]$, $\widehat{ut_i^2} = E[U_i t_i^2|\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, y_i]$ and using known properties of conditional expectation we obtain

$$\widehat{ut}_i = \widehat{u}_i(\widehat{\mu}_{Ti}+b) + \widehat{M}_T\widehat{\tau}_{1_i}, \quad \widehat{ut_i^2} \;=\; \widehat{u}_i(\widehat{\mu}_{T_i}+b)^2 + \widehat{M}_T^2 + \widehat{M}_T(\widehat{\mu}_{T_i}+2b)\widehat{\tau}_{1_i}, \tag{13}$$

where

$$\widehat{\tau}_{1_i} = E\left[U_i^{1/2} W_\Phi(\frac{U_i^{1/2}\widehat{\mu}_{T_i}}{\widehat{M}_T})|\widehat{\boldsymbol{\theta}}, y_i\right].$$

In each step, the conditional expectations $\widehat{u}_i = \widehat{u}_{1_i}$ and $\widehat{\tau}_{1_i}$ can be easily derived from the results given in Subsection 2.3. For the skew–t and skew contaminated normal distributions we have computationally attractive expressions that can be easily implemented. However, this is not the case for the skew–slash one, where Monte Carlo integration may be employed, which yield the so–called MC–EM algorithm; see Lachos et al. (2009).

These expressions are quite useful in implementing the M-step, which consists in maximizing the expected complete data function or the $Q-$function over $\boldsymbol{\theta}$, given by

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \;&=\; E[\ell_c(\boldsymbol{\theta})|\mathbf{y},\widehat{\boldsymbol{\theta}}^{(k)}] = c - \frac{n}{2}\log(\Gamma) - \frac{1}{2\Gamma}\sum_{i=1}^{n}\Big[\widehat{u}_i^{(k)}\,(y_i - \eta(\boldsymbol{\beta},\mathbf{x}_i))^2 \\ &-\; 2\Delta(y_i - \eta(\boldsymbol{\beta},\mathbf{x}_i))\widehat{ut}_i^{(k)} + \Delta^2\widehat{ut_i^2}^{(k)}]\Big], \end{aligned}$$

where $\widehat{\boldsymbol{\theta}}^{(k)}$ is an updated value of $\widehat{\boldsymbol{\theta}}$.

When the M-step turns out to be analytically intractable, it can be replaced with a sequence of conditional maximization (CM) steps. The resulting procedure is known as *ECM algorithm* (Meng and Rubin, 1993). Next, we describe this EM-type algorithm (ECM) for maximum likelihood

estimation of the parameters of the SMSN-NLM.

E-step: Given a current estimate $\widehat{\boldsymbol{\theta}}^{(k)}$, compute $\widehat{u}_i^{(k)}$ $\widehat{ut}_i^{(k)}$, $\widehat{ut_i^2}^{(k)}$, for $i = 1, \ldots, n$.

CM-step: Update $\widehat{\boldsymbol{\theta}}^{(k)}$ by maximizing $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ over $\boldsymbol{\theta}$, which leads to the following nice expressions

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \text{argmin}_{\boldsymbol{\beta}}(\mathbf{z}^{(k)} - \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}))^\top \widehat{\mathbf{U}}^{(\mathbf{k})}(\mathbf{z}^{(k)} - \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x})), \tag{14}$$

$$\widehat{\Delta}^{(k+1)} = \frac{\sum_{i=1}^n \widehat{ut}_i^{(k)}(y_i - \eta(\boldsymbol{\beta}^{(k+1)}, \mathbf{x}_i))}{\sum_{i=1}^n \widehat{ut_i^2}^{(k)}}, \tag{15}$$

$$\widehat{\Gamma}^{(k+1)} = \frac{1}{n}\sum_{i=1}^n \Big((y_i - \eta(\boldsymbol{\beta}^{(k+1)}, \mathbf{x}_i))^2 \widehat{u}_i^{(k)} - 2\Delta^{(k+1)}(y_i - \eta(\boldsymbol{\beta}^{(k+1)}, \mathbf{x}_i))\widehat{ut}_i^{(k)} + (\Delta^2)^{(k+1)}\widehat{ut^2}_i^{(k)}\Big), \tag{16}$$

where $\widehat{\mathbf{U}}^{(k)} = diag(\widehat{u}_1^{(k)}, \ldots, \widehat{u}_n^{(k)})$, $\mathbf{z}^{(k)}$ is the corrected observed response given by $\mathbf{z}^{(k)} = \mathbf{y} - \widehat{\Delta}^{(k)}\widehat{\boldsymbol{\tau}}^{(k)}$, with $\widehat{\boldsymbol{\tau}}^{(k)} = (\widehat{\tau_1}^{(k)}, \ldots, \widehat{\tau_n}^{(k)})^\top$, $\widehat{\tau}_i^{(k)} = \widehat{ut}_i^{(k)}/\widehat{u}_i^{(k)}$ and $\boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}) = (\eta(\boldsymbol{\beta}, \mathbf{x}_1), \ldots, \eta(\boldsymbol{\beta}, \mathbf{x}_n))^\top$. An interesting observation is that the M–step to estimate $\boldsymbol{\beta}$ is equivalent to the weighted nonlinear least squares in the NLM, $\mathbf{z} = \boldsymbol{\eta}(\boldsymbol{\beta}, \mathbf{x}) + \boldsymbol{\epsilon}$, in which reliable and efficient implementation of algorithms are available in softwares as SAS, R, Ox and Matlab. Note that $\widehat{\sigma}^{2(k+1)}$ and $\widehat{\lambda}^{(k+1)}$ can be recovered using (11), that is, $\lambda = \Delta/\sqrt{\Gamma}$ and $\sigma^2 = \Delta^2 + \Gamma$.

4.1. Notes on implementation

It is well known that maximum likelihood estimation in nonlinear models may face some computational hurdles, in the sense that the method may not give maximum global solutions if the starting values are far from the real parameter values. Thus, the choice of starting values for the EM algorithm in the non-linear context plays a big role in parameter estimation. In our example we consider the following procedure for the SN-NLM

- Compute $\boldsymbol{\beta}^{(0)}$ modeling using the standard nonlinear least squares
- compute the initial values $(\sigma^2)^{(0)}$ and $\lambda^{(0)}$ using the residuals and the method of moments estimators given in Section 2.2 with $M_1 = 0$; see also Lin et al. (2007b). The range for the skewness coefficient γ_1 of the SN distribution is approximately $(-0.9953, 0.9953)$ – see Azzalini (2005). But the method of moments can produce an initial value of $\gamma_1^{(0)}$ that is not in this interval. In this case, we use as starting points the values -0.99 (if $\gamma_1^{(0)} \leq -0.9953$) or 0.99 (if $\gamma_1^{(0)} \geq 0.9953$).

Now, when modeling using the ST-NLM, SCN-NLM or the SSL-NLM we adopt the following strategy

- Obtain initial values via method of moments for the SN-NLM, as described above;

- Perform maximum likelihood estimation of the parameters of the SN-NLM via EM algorithm;
- Use the EM estimates of the regression parameter, scale and skewness parameters of the SN-NLM as initial values for the corresponding ST-NLM, SSL-NLM and SCN-NLM parameters;
- In order to estimate ν in the ST-NLM and SSL-NLM we have fixed integer values for ν from 3 to 100 and 2 to 100 by 1, respectively, choosing the value of ν that maximizes the likelihood function. A similar procedure has been adopted for the SCN-NLM.

4.2. Model selection

For each fitted model, we computed the Akaike Information Criterion (AIC) (Akaike, 1974) and the Efficient Determination Criterion (EDC) (Bai et al., 1989). AIC and EDC have the form

$$-2\ell(\widehat{\boldsymbol{\theta}}) + \gamma c_n,$$

where $\ell(\cdot)$ is the actual log-likelihood, γ is the number of free parameters that have to be estimated under the model and the penalty term c_n is a convenient sequence of positive numbers. We have $c_n = 2$. For the EDC criterion, c_n is chosen so that it satisfies the conditions $c_n/n \to 0$ and $c_n/(\log\log n) \to 0$ when $n \to \infty$. Here we use $c_n = 0.2\sqrt{n}$, a proposal that was considered in Bai et al. (1989).

4.3. Residuals

Residual analysis aims at identifying atypical observations and/or model misspecification once residuals are measures of agreement between the data and the fitted model. Most residuals are

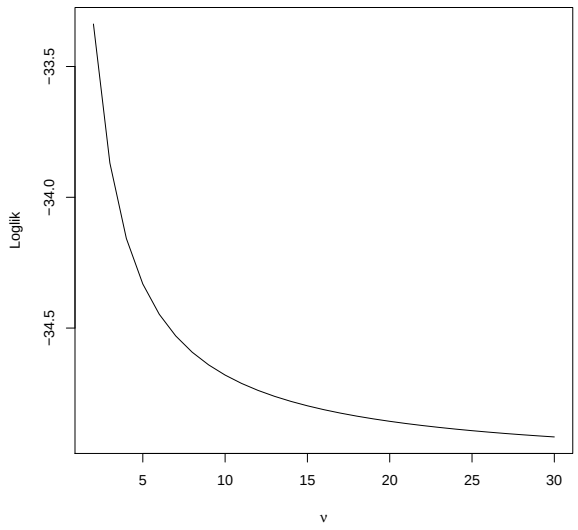


Figure 1: Oil palm data set. Plot of the profile log-likelihood of the parameter ν for fitting a ST-NLM.

based on the differences between the observed responses and the fitted conditional mean. We defined the following standardized ordinary residual (Pearson residuals):

$$r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{Var}(y_i)}},\ i = 1, \ldots, n,$$

where $\widehat{Var}(y_i) = k_2\widehat{\sigma}^2 - \frac{2}{\pi}k_1^2\widehat{\sigma}^2\widehat{\delta}^2$. Here, $\widehat{\mu}_i = \eta(\widehat{\boldsymbol{\beta}}, \mathbf{x}_i)$, and $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}^2$ and $\widehat{\delta}$ denoting the maximum likelihood estimators of $\boldsymbol{\beta}$, σ^2 and δ, respectively. We also generate envelopes, as suggested by Atkinson (1981), to detect incorrect specification of the error distribution and the systematic component $\eta(\boldsymbol{\beta}, \mathbf{x}_i)$ as well as the presence of outlying observations.

5. An Application

In this section we consider a likelihood analysis of the data set presented in Foong (1999) that describe the oil palm yield. Cancho et al. (2009) analyzed the same data set by fitting a SN-NLM. In this section, we revisit the oil palm data set with the aim of providing additional inferences by using SMSN distributions. Assuming a nonlinear growth-curve model, we fit a NLM to the data as specified by Cancho et al. (2009)

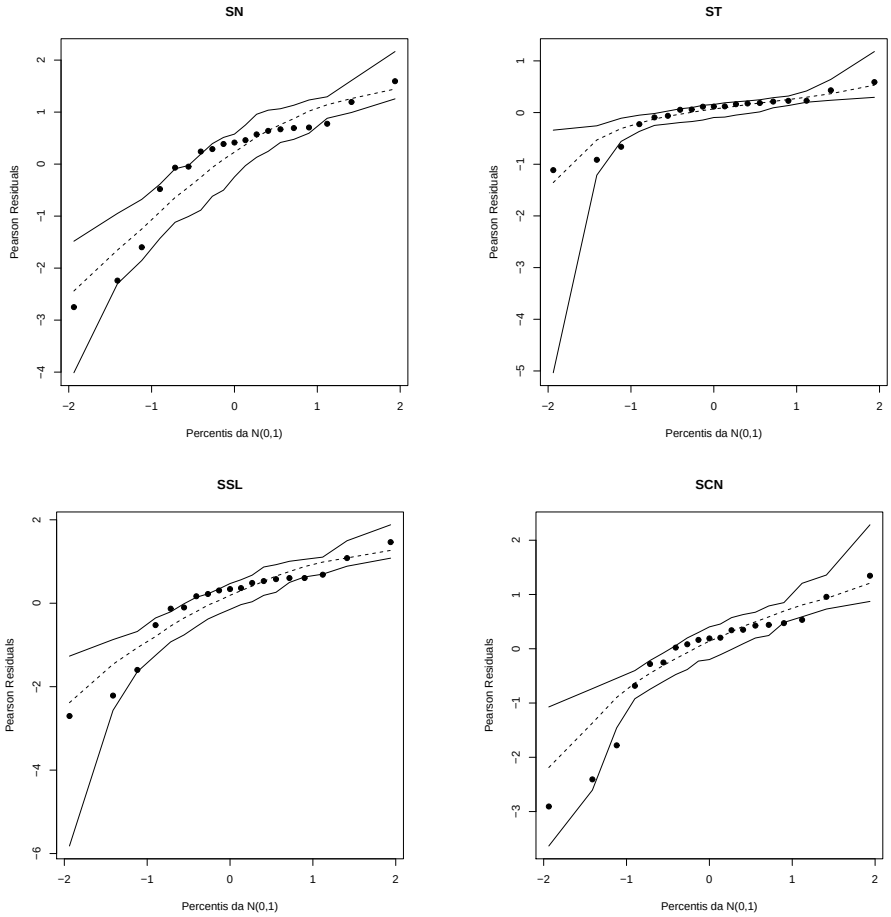
Table 1: ML estimation results for fitting various mixture models on the oil palm yield data set. SE are the asymptotic standard errors based on the observed information matrix.

	SN-NLM		ST-NLM		SCN-NLM		SSL-NLM	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	37.351	0.462	37.529	0.441	37.714	0.413	37.463	0.486
β_2	44.576	17.039	43.483	10.364	41.259	11.826	43.373	14.982
β_3	0.731	0.070	0.732	0.045	0.722	0.052	0.728	0.063
σ^2	6.919	2.655	1.644	1.152	2.077	1.343	3.105	1.708
λ	-4.453	3.125	-1.871	1.332	-2.269	1.641	-3.489	2.481
ν	-	-	3	-	0.2	-	2	-
γ	-	-	-	-	0.2	-	-	-
log-likelihood	-35.03691		-33.829		-34.132		-34.781	
AIC	80.07382		79.659		82.265		81.562	
EDC	74.43272		72.890		74.368		74.792	

$$Y_i = \frac{\beta_1}{1 + \beta_2 \exp(-\beta_3 x_i)} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} SMSN(-\sqrt{\frac{2}{\pi}}k_1\Delta, \sigma^2, \lambda; H), \tag{17}$$

for $i = 1, \ldots, 19$, where H denote the distribution function for the mixture variable U_i, for $i = 1, \ldots, 19$. In our analysis we will assume SN, ST, SSL and SCN distributions from the SMSN class for comparative purposes. We choose the value of $\boldsymbol{\nu}$ by maximizing the the likelihood function as illustrated in Figure 1. For the ST model we found $\nu = 3$, for the SSL we found $\nu = 2$ and for the SCN we found $\boldsymbol{\nu} = (0.2, 0.2)$. Table 1 contains the ML estimates of the parameters from the four models, together with their corresponding standard errors calculated via the observed information matrix. The AIC and EDC model selection criteria indicate that the ST distribution present the best fit. Although the regression estimates parameters are similar in all the four fitted models

Figure 2: Oil palm yield data set. Q–Q plots and simulated envelopes for the Pearson Residuals



(see Table 1) the standard errors of the SMSN-NLM with heavy tails are smaller than those in the SN-NLM. This suggests that the three models with longer tails than the SN model seem to produce more accurate maximum likelihood estimates. The estimates for the variance components (σ^2 and λ) are not comparable since they are on different scale.

The QQ-plots and envelopes for the Pearson residuals are shown in Figure 2. The lines in these figures represent the 5th percentile, the mean, and the 95th percentile of 100 simulated points for each observation. These Figures clearly shows once again that the ST distribution provides a better fit to the data set than the skew–normal distribution.

6. Conclusions

In this paper, we have proposed the application of a new class of asymmetric distributions, called the SMSN distribution, to nonlinear regression models. An EM–type algorithm is developed by exploring the statistical properties of the SMSN class. The observed information matrix is derived analytically which allows direct implementation of inference on this class of models. We

demonstrate our approach with a real data set and show that the ST model has better performance than the other competitors. R programs are available from the second author's homepage with website address http:// www.ime.unicamp.br/ hlachos/∼ListaPub.html.

Due to recent advances in computational technology, it is worthwhile to carry out Bayesian treatments via Markov chain Monte Carlo (MCMC) sampling methods in the context of SMSN-NLM. Other extensions of the current work include, for example, a generalization of SMSN-NLM to multivariate settings.

Acknowledgment: This research work was supported in part by grants 2008/02159-9 from FAPESP-Brazil.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Cont. 19, 716–723.

Arellano-Valle, R. B., Branco, M. D., Genton, M. G., 2006. A unified view on skewed distributions arising from selections. Canadian Journal of Statistics 34.

Arnold, B. . C., Beaver, R. J., Groeneveld, R. A., Meeker, W. Q., 1993. The nontruncated marginal of a truncated bivariate normal distribution. Psychometrika 58, 471–488.

Azzalini, A., 1985. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 12, 171–178.

Azzalini, A., Genton, M., 2008. Robust likelihood methods based on the skew-t and related distributions. International Statistical Review 76, 1490–1507.

Bai, Z. D., Krishnaiah, P. R., Zhao, L. C., 1989. On rates of convergence of efficient detection criteria in signal processing with white noise. IEEE Trans. Info. Theory 35, 380–388.

Berkane, M., Kano, Y., Bentler, P. M., 1994. Pseudo maximum likelihood estimation in elliptical theory: effects of misspecification. Computational Statistical & Data Analysis 18, :255–267.

Branco, M. D., Dey, D. K., 2001. A general class of multivariate skew-elliptical distributions. Journal of Multivariate Analysis 79, 99–113.

Cancho, V. C., Lachos, V. H., Ortega, E. M. M., 2009. A nonlinear regression model with skew-normal errors. Statistical Papers doi:10.1007/s00362-008-0139-y.

Cordeiro, G. M., Cysneiros, A. H. M. A., Cysneiros, F. J. A., 2009. Corrected maximum likelihood estimators im heteroscedastic symmetric nonlinear models. Journal of Statistical Computation and Simulation doi:10.1080/00949650802706420.

Cysneiros, F. J. A., Vanegas, L. H., 2008. Residuals and their statistical properties in symmetrical nonlinear models. Statistics & Probability Letters 78, 3269–3273.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1–38.

Foong, F. S., 1999. Impact of mixture on potential evapotranspiration, growth and yield of palm oil. PORIM Interl. Palm Oil Cong. (Agric.), 265–287.

Henze, N., 1986. A probabilistic representation of the skew-normal distribution. Scandinavian Journal of Statistics 13, 271–275.

Lachos, V. H., Ghosh, P., Arellano-Valle, R. B., 2009. Likelihood based inference for skew–normal independent linear mixed models. Statistica Sinica(in press).

Lange, K. L., Little, R., Taylor, J., 1989. Robust statistical modeling using t distribution. Journal of the American Statistical Association 84, 881–896.

Lange, K. L., Sinsheimer, J. S., 1993. Normal/independent distributions and their applications in robust regression. J. Comput. Graph. Stat 2, 175–198.

Lin, T. I., Lee, J. C., Hsieh, W. J., 2007a. Robust mixture modelling using the skew t distribution. Statistics and Computing 17, 81–92.

Lin, T. I., Lee, J. C., Yen, S. Y., 2007b. Finite mixture modelling using the skew normal distribution. Statistica Sinica 17, 909–927.

Meng, X., Rubin, D. B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 81, 633–648.

Wang, J., Genton, M. G., 2006. The multivariate skew-slash distribution. Journal of Statistical Planning and Inference 136, 209–220.

Xie, F. C., Weia, B. C., Lina, J. G., 2009. Homogeneity diagnostics for skew-normal nonlinear regression models. Statistics & Probability Letters 79, 821–827.