Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/140304

The Censored Bimodal Skew-Normal-Flexible Model with an Application to Plasma HIV-1 RNA

Germán Moreno-Arenas[1], *Guillermo Martínez-Flórez*[2,3] *and Artur J. Lemonte*[4,*]

[1]Escuela de Matemáticas, Universidad Industrial de Santander, Bucaramanga, Colombia
[2]Departamento de Matemáticas y Estadística, Facultad de Ciencias, Universidad de Córdoba, Montería, Colombia
[3]Programa de Pós-Graduação em Modelagem e Métodos Quantitativos, Universidade Federal do Ceará, Fortaleza, CE, Brazil
[4]Departamento de Estatística, CCET, Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil

Received: 7 Sep. 2019, Revised: 25 Sep. 2019, Accepted: 6 Dec. 2019
Published online: 1 May 2020

Abstract: We introduce the censored bimodal skew-normal-flexible distribution which can be very useful in dealing with censored data together with bimodality and high (or low) levels of skewness and kurtosis. A simple expression for its moments is derived. The frequentist approach is considered to draw inference, and the traditional maximum likelihood method is employed to estimate the unknown parameters. We also derive the observed information matrix. An application of the new model to the plasma HIV-1 RNA measurement is presented for illustrative purposes.

Keywords: Bimodal data; Censored data, maximum likelihood estimation, skew distributions

1 Introduction

In epidemiologic studies with biomarkers as main outcomes, it is common to have limits of detection which may hinder defining the specific values. The biomarkers are useful for predictive applications in diagnostics, prognostics, or predicted response to therapy. The highly active antiretroviral therapy (HAART) is a treatment that defers the development of AIDS in HIV-positive patients and the main mechanism for achieving these results is lowering HIV-1 RNA to undetectable levels. However, undetectable levels of HIV-1 RNA in plasma do not mean that viral replication has been stopped and ultrasensitive tests generate a smaller proportion of censored data. These assays have the potential to delineate the leftmost peak of the HIV-1 RNA distribution in treated populations. The characterization of this leftmost peak is important to understand not only the effectiveness of HAART but also the biology of optimally treated HIV-1 infection.

Besides the censorship, these data have bimodal behavior. For example, [1] and [2] found that the viral load has a bimodal behavior in a study conducted on HIV-positive patients. [3] corroborated the bimodality of HIV-1 RNA viral load of individuals on HAART. The rightmost peak of the distribution reflects individuals with suboptimal virologic response to HAART, whereas the leftmost peak reflects individuals with the maximal reduction that can be achieved with medications of the therapy. In addition, bimodality of biomarkers has been shown in other contexts. For instance, [4] found the bimodality of blood glucose in populations with a very low prevalence of diabetes and obesity.

Let Y be a random variable which has a part of their probabilities at discrete points and the rest spread in some intervals. The cumulative distribution function (CDF) of Y, say $F(y)$, can be expressed as a mixture distribution in the form $F(y) = p_1F_1(y) + p_2F_2(y)$, for $y \in \mathbb{R}$, where $F_1(y)$ is a stepwise CDF, $F_2(y)$ is a continuous CDF, p_1 is the cumulative probability of all the discrete points, and $p_2 = 1 - p_1$ is the cumulative probability of the continuous portion. When data are censored to the left, the random variable Y is a mixture of a continuous latent process Y^* and a selection mechanism (censoring or truncation) modeled in binary form. This idea was popularized by [5], and the resulting model is known as Tobit model, which is defined in terms of the latent

[*] Corresponding author e-mail: arturlemonte@gmail.com

variable

$$y_i = \begin{cases} y_i^*, & y_i^* > c, \\ c, & \text{otherwise}, \end{cases}$$

where c is the point of censorship (or limit of detection), and Y^* has a certain distribution as, for example, the normal distribution studied by [5], the Student-t distribution investigated by [6], or the power-normal distribution addressed by [7].

In certain cases, in addition to censorship and bimodal behavior, the data also have high (or low) degrees of skewness and kurtosis. Therefore, data analysis is essential to make use of skewed distributions, wherein at least one of the components of the mixture or even the complete distribution has asymmetric form. An interesting alternative is appealed to the skew-normal distribution defined by [8], for example. For uncensored data, extensions for asymmetric cases have been studied by [9], [10] and [11]. In particular, [11] introduced an asymmetric distribution based on the skew-normal distribution that admits bimodality. In this paper, we propose a new distribution called *censored bimodal skew-normal-flexible model* that is quite useful in addressing censored data together with bimodality and high (or low) levels of skewness and kurtosis more than the usual normal distribution. Specifically, our approach relies on generalizing the skew-normal-flexible distribution [11] to deal with this kind of data (i.e. censored data with bimodality and high (or low) levels of skewness and kurtosis). We shall also propose the unimodal power-normal distribution [12] in such a context for the sake of comparison. Recently, distribution theory has received considerable attention. Thus, new parametric distributions have been introduced in the statistic literature as, for example, [13], [14] and [15].

The present paper is organized as follows: Section Two involves some preliminaries. In Section Three, we introduce the censored unimodal power-normal model, and maximum likelihood (ML) estimation of the model parameters presented. In Section Four, we introduce the censored bimodal skew-normal-flexible model as an extension of the skew-normal-flexible model addressed by [11]. Moments and maximum likelihood estimation for the model parameters are discussed. We also derive closed-form expressions for the observed information matrix. Appropriateness of the new censored model is illustrated using real data in Section Five. Section Six is devoted to conclusion.

2 Preliminaries

Certain parametric families of distributions are of special interest in modeling asymmetric data. A noteworthy case is the skew-normal distribution [8], because this model represents a generalization of the normal distribution with an additional parameter to regulate skewness. The probability density function (PDF) of a random variable Z skew-normal distributed and parameter $\alpha \in \mathbb{R}$ is given by

$$f(z;\alpha) = 2\phi(z)\Phi(\alpha z), \quad z \in \mathbb{R},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal PDF and CDF, respectively. The skewness is controlled by the parameter $\alpha \in \mathbb{R}$; $\alpha = 0$ yields the symmetric normal distribution. The skew-normal (SN) distribution was first introduced by [16] as a prior distribution for estimating a normal location parameter. [8] investigated its fundamental properties and proposed some generalizations and extensions. We shall use the notation $Z \sim \mathrm{SN}(\alpha)$ to refer to this distribution. The three-parameter location-scale extension of Z, say Y, is given by $Y = \xi + \eta Z$, where $\xi \in \mathbb{R}$ is a location parameter, and $\eta > 0$ is a scale parameter. In this case, we have $Y \sim \mathrm{SN}(\xi, \eta, \alpha)$.

In a hydrological context, [17] introduced the fractional order statistics distribution with PDF given by $\varphi_F(z;\beta) = \beta f(z)F(z)^{\beta-1}$ for $z \in \mathbb{R}$, where $F(\cdot)$ is an absolutely continuous CDF, $f(\cdot)$ is the corresponding PDF, and $\beta > 0$ is a shape parameter that controls the amount of asymmetry in the distribution. We refer to this model as the power distribution. Recent results by [18] revealed that power distributions can be a viable alternative for modeling asymmetric data. Following the idea of [17], [12] defined the power-normal (PN) distribution whose PDF is given by

$$f(z;\beta) = \beta\phi(z)\Phi(z)^{\beta-1}, \quad z \in \mathbb{R},$$

where $\beta > 0$, with $\beta = 1$ yielding the normal model. We shall use the notation $Z \sim \mathrm{PN}(\beta)$ to refer to this distribution. The location-scale extension is defined as the distribution of the random variable $Y = \xi + \eta Z$, where $\xi \in \mathbb{R}$ is a location parameter, and $\eta > 0$ is a scale parameter. We use the notation $Y \sim \mathrm{PN}(\xi, \eta, \beta)$. SN and PN distributions are both unimodal.

3 Censored unimodal power-normal distribution

3.1 The model

Consider a random variable $Y^* \sim \mathrm{PN}(\xi, \eta, \beta)$, and let $\{y_1^*, y_2^*, \ldots, y_n^*\}$ be a random sample of size n and point of censorship equal to c. Values of y^* greater than the constant c are recorded to themselves, whereas values of y^* less than or equal to the constant c are assigned to c. The observed random variable Y is described as

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* > c, \\ c, & \text{if } y_i^* \le c, \end{cases}$$

for $i = 1, 2, \ldots, n$. Consequently,

$$\Pr(y_i = c) = \Pr(y_i^* \le c) = \Phi\left(\frac{c-\xi}{\eta}\right)^{\beta},$$

and for $y_i^* > c$, the distribution of Y is the same as that of Y^*. The resulting sample is left-censoring, so we have the censored PN (CPN) distribution. We use the notation $Y \sim \mathrm{CPN}(\xi, \eta, \beta)$. If $\beta = 1$, the censored normal (CN) distribution arises. The inclusion of the parameter β allows for modeling the asymmetry present in the data. Therefore, the CPN distribution is more flexible than the censored normal distribution. On the other hand, the CPN model corresponds to a censored unimodal model, which represents a disadvantage in relation to the proposed censored model based on the skew-normal-flexible model. It should be noticed that the generalization to right-censoring is trivial.

3.2 Parameter estimation

Let C be the set of censored observations, and U be the set of uncensored observations. The log-likelihood function for $\boldsymbol{\theta} = (\xi, \eta, \beta)^\top$ is

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= n_1[\log(\beta) - \log(\eta)] + n_2\beta \log[\Phi(z_c)] \\ &\quad + \sum_{i\in U} \log[\phi(z_i)] + (\beta - 1)\sum_{i\in U} \log\left[\Phi(z_i)\right],\end{aligned}$$

where n_1 and n_2 denote the number of uncensored and censored observations, respectively, and $z_c = (c - \xi)/\eta$ and $z_i = (y_i - \xi)/\eta$. The ML estimates of the CPN model parameters are obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The maximization can be performed, for example, in the R [19] software using the `optim(...)` function.

The score vector $\boldsymbol{U}(\boldsymbol{\theta}) = (U_\xi, U_\eta, U_\beta)^\top$, obtained by differentiating the log-likelihood function with respect to the unknown parameters, has components

$$U_\xi = -\frac{n_1\beta w_c}{\eta} + \frac{1}{\eta}\sum_{i\in U}\left[z_i - (\beta - 1)w_i\right],$$

$$U_\eta = -\frac{n_2\beta z_c w_c}{\eta} + \frac{1}{\eta}\sum_{i\in U}\left[-1 + z_i^2 - (\beta - 1)z_i w_i\right],$$

$$U_\beta = \frac{n_1}{\beta} + n_2\log\left[\Phi(z_c)\right] + \sum_{i\in U}\log\left[\Phi(z_i)\right],$$

where $w_c = \phi(z_c)/\Phi(z_c)$ and $w_i = \phi(z_i)/\Phi(z_i)$. The ML estimates can also be obtained by solving simultaneously the nonlinear system of equations $U_\xi = 0$, $U_\eta = 0$ and $U_\beta = 0$. However, this nonlinear system of equations does not admit explicit solution. Hence, the ML estimates need to be obtained through a numerical maximization of the log-likelihood function using nonlinear optimization algorithms. In particular, we have used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) nonlinear optimization algorithm in `optim(...)` function to compute the ML estimates. To start the algorithm, we take $\beta_0 = 1$, $\xi_0 = \widehat{\xi}_0$ and $\eta_0 = \widehat{\eta}_0$, where $\widehat{\xi}_0$ and $\widehat{\eta}_0$ are the ML estimates of the Tobit model under the normal distribution.

The observed information matrix is

$$\boldsymbol{J}(\boldsymbol{\theta}) = \begin{bmatrix} j_{\xi\xi} & j_{\xi\eta} & j_{\xi\beta} \\ j_{\xi\eta} & j_{\eta\eta} & j_{\eta\beta} \\ j_{\xi\beta} & j_{\eta\beta} & j_{\beta\beta} \end{bmatrix},$$

whose elements are

$$j_{\beta\beta} = \frac{n_1}{\beta^2}, \quad j_{\xi\beta} = \frac{n_2 w_c}{\eta} + \frac{1}{\eta}\sum_{i\in U} w_i,$$

$$j_{\eta\beta} = \frac{n_2 z_c w_c}{\eta} + \frac{1}{\eta}\sum_{i\in U} z_i w_i,$$

$$j_{\xi\xi} = \frac{n_2\beta}{\eta^2}[z_c w_c + w_c^2] + \frac{1}{\eta^2}\sum_{i\in U}\{1 + (\beta - 1)[z_i w_i + w_i^2]\},$$

$$\begin{aligned} j_{\xi\eta} &= \frac{\beta}{\eta^2}\sum_{i\in C}[-w_c + z_c^2 w_c + z_c w_c^2] \\ &\quad + \frac{1}{\eta^2}\sum_{i\in U}\{2z_i + (\beta - 1)[-w_i + z_i^2 w_i + z_i w_i^2]\},\end{aligned}$$

$$\begin{aligned} j_{\eta\eta} &= \frac{\beta}{\eta^2}\sum_{i\in C}[-2z_c w_c + z_c^2 w_c^2 + z_c^3 w_c] \\ &\quad + \frac{1}{\eta^2}\sum_{i\in U}\{-1 + 3z_i^2 + (\beta - 1)[-2z_i w_i + z_i^2 w_i^2 + z_i^3 w_i]\}.\end{aligned}$$

The observed information matrix is used for computing asymptotic standard errors for the ML estimates, as well as asymptotic confidence intervals for the model parameters. The censored bimodal model based on the skew-normal-flexible model will be introduced in the next section.

4 Censored bimodal skew-normal-flexible model

4.1 The model

Under uncensored settings, [11] introduced the skew-normal-flexible (SNF) distribution. Its PDF is given by

$$f(z;\alpha,\delta) = \kappa_\delta \phi(|z| + \delta)\Phi(\alpha z), \quad z \in \mathbb{R}, \qquad (1)$$

where $\delta \in \mathbb{R}$, and $\kappa_\delta = [1 - \Phi(\delta)]^{-1}$ is the normalization constant. We shall use the notation $Z \sim \mathrm{SNF}(\alpha, \delta)$ to refer to this distribution. If $\delta = 0$, the SNF model reduces to the SN model. When $\alpha = \delta = 0$ we obtain the normal distribution. In addition, $\delta < 0$ implies in an asymmetric bimodal PDF, whereas $\alpha = 0$ and $\delta < 0$ yields a bimodal symmetric PDF. The reader is referred to [11] for a detailed description of the SNF distribution.

Let us assume now that $Y^* \sim \mathrm{SNF}(\alpha, \delta)$. Let $\{y_1^*, y_2^*, \ldots, y_n^*\}$ be a random sample of size n and point of censorship equal to c. As previously defined, the observed random variable Y is described as

$$y_i = \begin{cases} y_i^*, & y_i^* > c, \\ c, & y_i^* \leq c, \end{cases}$$

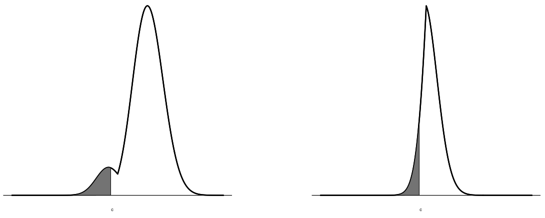
for $i = 1, 2, \ldots, n$. The resulting sample is a SNF with left censoring. Hence, we say that Y is a censored SNF (CSNF) random variable. We use the notation $Y \sim \text{CSNF}(\alpha, \delta)$. We have that the CSNF PDF is bimodal for $\delta < 0$, and unimodal for $\delta > 0$. The case $\delta = 0$ and $\alpha = 0$ corresponds to the CN distribution, while $\delta = 0$ and $\alpha \neq 0$ corresponds to the censored SN (CSN) distribution. Thus, the proposed censored model generalizes some known models. In addition, it is more flexible than these models because it describes the bimodal behavior in the data which cannot be described by the CN and CSN models. The PDF of the location-scale extension of the SNF model takes the form ($y \in \mathbb{R}$)

$$f(y; \xi, \eta, \alpha, \delta) = \frac{\kappa_\delta}{\eta} \phi\left(\left|\frac{y-\xi}{\eta}\right| + \delta\right) \Phi\left(\alpha \frac{y-\xi}{\eta}\right),$$

where $\xi \in \mathbb{R}$ is the location parameter, and $\eta > 0$ is the scale parameter. In this case, we have $Y^* \sim \text{SNF}(\xi, \eta, \alpha, \delta)$. Putting

$$y_i = \begin{cases} y_i^*, & y_i^* > c, \\ c, & y_i^* \leq c, \end{cases}$$

we obtain the location-scale extension of the CSNF distribution. In this case, we use the notation $Y \sim \text{CSNF}(\xi, \eta, \alpha, \delta)$. Figure 1 illustrates the CSNF distribution for $\delta < 0$ and $\delta > 0$.



(a) $\delta < 0$ (b) $\delta > 0$

Fig. 1: CSNF model with left censoring.

4.2 Moments

The moments of the random variable $Z \sim \text{CSNF}(\alpha, \delta)$ censored at the point c are functions of the quantities

$$\mu_r(x) = \int_{-\infty}^{x} z^r \phi(z) dz,$$

$$\mu_r(x, \alpha, \delta) = \int_{-\infty}^{x} z^r \phi(z) \Phi(\alpha(z+\delta)) dz.$$

Note that $\mu_r(x)$ and $\mu_r(x, \alpha, 0)$ are the incomplete moments of the standard normal and SN distribution, respectively. In particular, for $r = 0$ and $\delta = 0$, we have that $\mu_0(x, \alpha, 0) = \Phi(x) - T(x, \alpha)$, where $T(h, k)$ is the function known as Owen function [20]. The r-th moment of $Z \sim \text{CSNF}(\alpha, \delta)$ can be reduced to

$$\begin{aligned} \mathbb{E}(Z^r) = c^r \psi(c) + \kappa_\delta (-1)^r \sum_{k=0}^{r} \binom{k}{r} \delta^{r-k} \Big[& \mu_k(-(c+\delta)) \\ & - \mu_k(-(c+\delta), \alpha, \delta) \Big], \end{aligned}$$

where $\psi(c) = \kappa_\delta \int_{-\infty}^{c} \phi(|z| + \delta) \Phi(\alpha z) dz$. The mean is given by

$$\begin{aligned} \mathbb{E}(Z) = {} & \kappa_\delta \big[\delta (\Phi(c+\delta)(\Phi(\lambda\delta) + 1) - 1) \\ & + \phi(c+\delta)\Phi(\alpha\delta) \big] + c\psi(c) \\ & + \delta\kappa_\delta \left[T\left(c+\delta, \frac{c\alpha}{c+\delta}\right) - T\left(c+\delta, \frac{\lambda\delta}{c+\delta}\right) \right] \\ & + \kappa_\delta \lambda \phi(\lambda\delta) \Phi\left(\alpha\lambda\delta - \frac{\alpha(c+\delta)}{\lambda}\right) \\ & + \delta\kappa_\delta \left[T\left(\lambda\delta, \frac{c(1+\alpha^2)+\delta}{\alpha\delta}\right) - T\left(\lambda\delta, \frac{c+\delta}{\lambda\delta}\right) \right], \end{aligned}$$

where $\lambda = \alpha/\sqrt{1+\alpha^2}$. The variance is $\mathbb{VAR}(Z) = \mathbb{E}(Z^2) - \mathbb{E}(Z)^2$, where

$$\begin{aligned} \mathbb{E}(Z^2) = {} & c^2 \psi(c) + \kappa_\delta \big[(1+\delta^2)(1 - \Phi(c+\delta)) \\ & + \delta(\delta\Phi(c\alpha) - 2\phi(c+\delta) + \Phi(c+\delta)\Phi(\lambda\delta)) \big] \\ & - \lambda\delta(2+\alpha\delta)\kappa_\delta \phi(\lambda\delta) \big[\phi(\sqrt{1+\alpha^2}(\delta\lambda^2 - \delta - c)) \\ & + \alpha\lambda\delta\Phi(\sqrt{1+\alpha^2}(\delta\lambda^2 - \delta - c)) \big] \\ & + \delta(1+\delta)\kappa_\delta \Bigg[T\left(c+\delta, \frac{c\alpha}{c+\delta}\right) \\ & \quad - T\left(c+\delta, \frac{\lambda\delta}{c+\delta}\right) \Bigg] \\ & + \delta(1+\delta)\kappa_\delta \Bigg[T\left(\lambda\delta, \frac{c(1+\alpha^2)+\delta}{\alpha\delta}\right) \\ & \quad - T\left(\lambda\delta, \frac{c+\delta}{\lambda\delta}\right) \Bigg]. \end{aligned}$$

For the location-scale extension $Y = \xi + \eta Z \sim \text{CSNF}(\xi, \eta, \alpha, \delta)$, it then follows that

$$\mathbb{E}(Y^r) = \sum_{k=0}^{r} \binom{k}{r} \xi^{r-k} \eta^k \mathbb{E}(Z^r).$$

Next section covers parameter estimation of the CSNF model parameters. Similar to the CPN model, we will consider the ML method, because the method of moments is quite complicated and creates a complex system of equations.

4.3 Parameter estimation

To estimate the CSNF model parameters, we consider the ML method. We can assume $c = 0$ to estimate the CSNF model parameters without loss of generality, i.e.

$$y_i = \begin{cases} y_i^*, & y_i^* > 0, \\ 0, & y_i^* \leq 0. \end{cases}$$

Consequently, the contribution of censored and uncensored observations to the log-likelihood function is as follows: if $y_i = 0$, $\Pr(y_i \leq 0) = \psi(0) = \kappa_\delta \rho(0)$, where

$$\begin{aligned}\rho(0) = {} & \Phi(\lambda\delta)\left(1 - \Phi\left(\frac{\xi + \eta\delta}{\eta}\right)\right) \\ & + T\left(\lambda\delta, \frac{\xi(1+\alpha^2) + \eta\delta}{\alpha\eta\delta}\right) \\ & - \left[T\left(\frac{\xi+\eta\delta}{\eta}, \frac{\lambda\xi}{\xi+\eta\delta}\right) + T\left(\frac{\xi+\eta\delta}{\eta}, \frac{\lambda\eta\delta}{\xi+\eta\delta}\right) \right. \\ & \left. + T\left(\lambda\delta, \frac{\xi+\eta\delta}{\lambda\eta\delta}\right)\right],\end{aligned}$$

and if $y_i > 0$, the distribution of Y_i is equal to that of Y_i^*. Then, for a random sample $\{y_1, y_2, \ldots, y_n\}$ of size n, the log-likelihood function for $\boldsymbol{\theta} = (\xi, \eta, \alpha, \delta)^\top$ is

$$\begin{aligned}\ell(\boldsymbol{\theta}) = {} & n_1[\log(\kappa_\delta) - \log(\eta)] + n_2 \log[\psi(0)] \\ & + \sum_{i \in U} \{\log[\phi(|z_i| + \delta)] + \log[\Phi(\alpha z_i)]\},\end{aligned}$$

where $z_i = (y_i - \xi)/\eta$, and n_1 and n_2 denote the number of uncensored and censored observations, respectively. The ML estimates of the CSNF model parameters are obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, and this maximization can be performed in the R program [19] using the `optim(...)` function. Similar to the CPN model, we have used the BFGS nonlinear optimization algorithm in `optim(...)` function to compute the ML estimates. To start the algorithm, we take $\delta_0 = 0$, $\xi_0 = \widehat{\xi}_0$, $\eta_0 = \widehat{\eta}_0$ and $\alpha_0 = \widehat{\alpha}_0$, where $\widehat{\xi}_0$, $\widehat{\eta}_0$ and $\widehat{\alpha}_0$ are the ML estimates of the Tobit model under the SN distribution.

The elements of the score vector $\boldsymbol{U}(\boldsymbol{\theta}) = (U_\xi, U_\eta, U_\alpha, U_\delta)^\top$ are

$$U_\xi = n_2 \Lambda_\xi + \frac{1}{\eta}\sum_{i \in U} z_i + \frac{\delta}{\eta}\sum_{i \in U} \operatorname{sign}(y_i - \xi) - \frac{\alpha}{\eta}\sum_{i \in U} w_i,$$

$$U_\eta = n_2 \Lambda_\eta - \frac{n_1}{\eta} + \frac{1}{\eta}\sum_{i \in U} z_i^2 + \frac{\delta}{\eta}\sum_{i \in U} |z_i| - \frac{\alpha}{\eta}\sum_{i \in U} w_i z_i,$$

$$U_\alpha = n_2 \Lambda_\alpha + \sum_{i \in U} w_i z_i,$$

$$U_\delta = n_2 \Lambda_\delta - n_1 \delta + \frac{n\phi(\delta)}{1 - \Phi(\delta)} - \sum_{i \in U} |z_i|,$$

where $w_i = \phi(\alpha z_i)/\Phi(\alpha z_i)$, $n = n_1 + n_2$ and Λ_ξ, Λ_η, Λ_α and Λ_δ correspond to the partial derivatives of $\log[\psi(0)]$ with respect to ξ, η, α and δ, respectively. These quantities are provided in the Appendix. We can also obtain the ML estimates by simultaneously solving the nonlinear system of equations $U_\xi = 0$, $U_\eta = 0$, $U_\alpha = 0$ and $U_\delta = 0$, which has no closed-form solution. As a result, it is necessary to consider nonlinear optimization algorithms to obtain the ML estimates numerically. Finally, for $c \neq 0$, we make $y_i = y_i^* - c$. Hence, the ML estimates obtained by considering $c = 0$ can be used for this more general setting.

Next, we make some assumptions on the behavior of $\ell(\boldsymbol{\theta})$ as the sample size n approaches infinity, such as the regularity of the first two derivatives of $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and the existence and uniqueness of the ML estimate of $\boldsymbol{\theta}$; see, for example, [21]. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\xi}, \widehat{\eta}, \widehat{\alpha}, \widehat{\delta})^\top$ be the ML estimator of $\boldsymbol{\theta} = (\xi, \eta, \alpha, \delta)^\top$. When n is large and under standard regularity conditions, the ML estimators of the CSNF model parameters are asymptotically normal, asymptotically unbiased and have asymptotic variance-covariance matrix given by the inverse of the expected information matrix: $\widehat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}_4(\boldsymbol{\theta}, \boldsymbol{K}(\boldsymbol{\theta})^{-1})$, where "$\overset{a}{\sim}$" means approximately distributed, $\boldsymbol{K}(\boldsymbol{\theta}) = -\mathbb{E}(\boldsymbol{H}(\boldsymbol{\theta}))$ is the 4×4 expected information matrix, and $\boldsymbol{H}(\boldsymbol{\theta}) = \partial^2 \ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ is the Hessian matrix. There is no closed-form expression for the matrix $\boldsymbol{K}(\boldsymbol{\theta})$. However, the asymptotic behavior remains valid if $\boldsymbol{K}(\boldsymbol{\theta})$ is approximated by $\boldsymbol{J}_n(\widehat{\boldsymbol{\theta}}) = -\boldsymbol{H}(\widehat{\boldsymbol{\theta}})$, which is the 4×4 observed information matrix evaluated at $\widehat{\boldsymbol{\theta}}$. The matrix $\boldsymbol{J}_n(\boldsymbol{\theta})$ has the form

$$\boldsymbol{J}_n(\boldsymbol{\theta}) = \begin{bmatrix} j_{\xi\xi} & j_{\xi\eta} & j_{\xi\alpha} & j_{\xi\delta} \\ j_{\xi\eta} & j_{\eta\eta} & j_{\eta\alpha} & j_{\eta\delta} \\ j_{\xi\alpha} & j_{\eta\alpha} & j_{\alpha\alpha} & j_{\alpha\delta} \\ j_{\xi\delta} & j_{\eta\delta} & j_{\alpha\delta} & j_{\delta\delta} \end{bmatrix},$$

and the elements are provided in the Appendix. The above-mentioned asymptotic normal distribution can be used to construct approximate confidence intervals for the CSNF model parameters, as well as to compute asymptotic standard errors for the ML estimates.

5 Real data application

In this section, we present an application of the proposed CSNF distribution to real data for illustrative purposes. The R code used in the real data application can be obtained from the authors upon request. We also consider the CSN and CPN models for comparison. We shall consider the real dataset presented by [3], which represents the plasma HIV-1 RNA measured in 306 samples, collected from 273 men in highly active antiretroviral therapy, with both Roche [22] (whose limit of detection is 20 copies per millilitre), and Roche [23]

Table 1: ML estimates and standard errors (in parenthesis).

CSN(ξ,η,α)
$\widehat{\xi}=4.355\,(0.379)$
$\widehat{\eta}=11.121\,(1.371)$
$\widehat{\alpha}=-9.637\,(3.274)$
CPN(ξ,η,β)
$\widehat{\xi}=0.001\,(0.025)$
$\widehat{\eta}=1.901\,(0.149)$
$\widehat{\beta}=0.492\,(0.052)$
CSNF(ξ,η,α,δ)
$\widehat{\xi}=1.519\,(0.090)$
$\widehat{\eta}=0.963\,(0.127)$
$\widehat{\alpha}=-0.667\,(0.097)$
$\widehat{\delta}=-2.208\,(0.220)$

Table 2: AIC values.

Model	AIC
CSN(ξ,η,α)	685.37
CPN(ξ,η,β)	611.04
CSNF(ξ,η,α,δ)	584.75

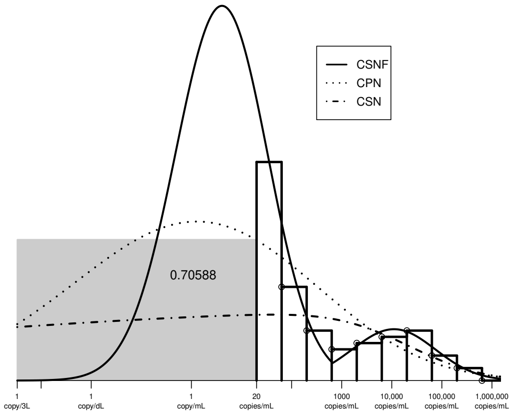


Fig. 2: Histogram and fitted models: CSNF (solid line), CPN (dotted line) and CSN (dashed line).

(whose limit of detection is 50 copies per millilitre) assays; see [3] for details. The data used in this paper are only measurements made with the Roche COBAS® TaqMan assay. The histogram in Figure 2 shows a bimodal behavior of the HIV-1 RNA measurements.

Table 1 lists the ML estimates of the CSN, CPN and CSNF parameters. ML estimate of δ is negative ($\widehat{\delta}=-2.208$) suggesting that the CSNF model can describe the bimodal behavior of the data evidenced by the histogram in Figure 2. Now, we shall compute the Akaike information criterion (AIC) to verify which model fits these real data better. AIC is the most common parametric statistic for model selection. The values of AIC for the CSN, CPN and CSNF models fitted to the data are presented in Table 2. It is evident that the CSNF model has the smallest value of AIC in comparison to the other ones, so it might be chosen as the best model. More information is provided by a visual comparison of the three fitted models (see Figure 2). Note that the CSN and CPN models fail to explain the bimodality of the data. On the other hand, the CSNF model is capable of describing this bimodal behavior adequately.

Next, we perform a parametric test to probe the bimodality hypothesis, that is, we want to test

$$H_0:\ \delta=0\ \text{ against }\ H_1:\ \delta<0.$$

In other words, we are comparing the CSN model (i.e. $\delta=0$ in the CSNF model) with the CSNF model. Using the likelihood ratio statistic, we have

$$\omega=\frac{\ell_{\mathrm{CSN}}(\hat{\boldsymbol{\theta}})}{\ell_{\mathrm{CSNF}}(\hat{\boldsymbol{\theta}})},$$

leading to

$$-2\log(\omega)=-2(-339.6866+288.3771)=102.619,$$

which is greater than the value of the chi-square critical point to a 5% significance level ($\chi^2_{1;0.95}=3.8414$). So, it confirms that the CSN model fails to describe the bimodality of data at hand and, in addition, the CSNF model outperforms the CSN model on the basis of the likelihood ratio test.

In general, some authors consider the two-component mixture normal (MN) distribution for modeling bimodal dada. The MN PDF can be expressed as

$$f(x)=pf_1(x;\mu_1,\sigma_1)+(1-p)f_2(x;\mu_2,\sigma_2),\quad x\in\mathbb{R},$$

where $f_j(\cdot)$ is the PDF of the normal distribution with parameters (μ_j,σ_j) for $j=1,2$, and $0<p<1$ the probability of mixture. For censored data, we then have the censored MN (CMN) model, denoted by CMN$(\mu_1,\sigma_1,\mu_2,\sigma_2,p)$. In particular, the current data were analyzed in [3] by considering the CMN model. The ML estimates of the CMN model parameters are $\widehat{\mu}_1=0.577$, $\widehat{\sigma}_1=0.903$, $\widehat{\mu}_2=4.15$, $\widehat{\sigma}_2=0.706$ and $\widehat{p}=0.897$. The AIC of the CMN model fitted to the data is 585.27, which is slightly greater than that of the CSNF model (see Table 2). Therefore, the proposed four-parameter CSNF model fits the data better than the five-parameter CMN model. Figure 3 displays the estimated CSNF and CMN PDFs. Note that the models are very similar, but the CSNF model has fewer parameters to be estimated and so should be preferable. Figure 4 presents the QQ-plot regarding the uncensored data, which indicate the appropriateness of the proposed

CSNF distribution for modeling the HIV-1 RNA data. Finally, the proportion of censored data corresponds to 70.58% of the population, and the estimated proportion of censoring based on the CSNF model is 70.56%, which definitely demonstrates the good performance of the CSNF distribution in modeling the HIV-1 RNA data.

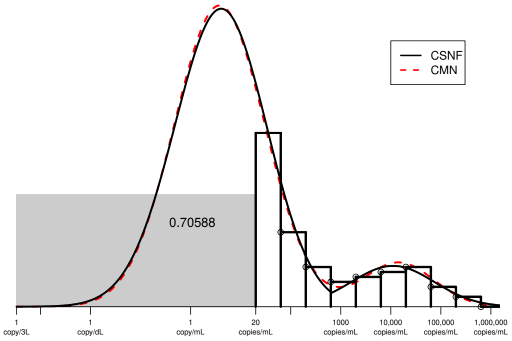


Fig. 3: Histogram and fitted models: CSNF (solid line) and CMN (dashed line).

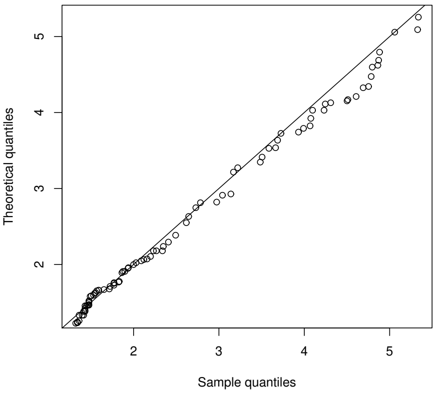


Fig. 4: QQ-plot of the uncensored data.

6 Conclusion

In this paper, we investigated a new unimodal censored model called the censored power-normal (CNP) distribution, as well as a new bimodal censored model called the censored skew-normal-flexible (CSNF) distribution. In particular, the CSNF model is quite flexible and it can deal with censored data together with bimodality and high (or low) levels of skewness and kurtosis. A simple expression for the moments of the CSNF model was derived. The estimation of the CNP and CSNF parameters was performed using the method of maximum likelihood. We also provided closed-form expressions for the observed information matrix of the CPN and CSNF models. In addition, we illustrated the methodology developed in this paper by means of an application to real data. We verify through the real data application that the CSNF model surpassed the well-known censored skew-normal model, and was very similar to the censored two-component mixture normal (CMN) model, which might be the most used model to deal with censored data together with bimodality. The advantage of the proposed CSNF model in relation to the CMN model is that the CSNF model has fewer parameters to be estimated than the CMN model. Hence, the proposed CSNF model is simpler (i.e. more parsimonious) than the CMN model in practical situations. However, the CMN model parameters have a direct interpretation, while the CSNF model parameters lack it. Finally, formulas related with the CSNF model are manageable (such as log-likelihood function, score function, and observed information matrix). Also, using modern computer resources and its numerical capabilities, this censored model may prove to be a useful addition to the arsenal of applied statisticians.

Acknowledgement

Artur Lemonte acknowledges the financial support of the Brazilian agency CNPq (grant 301808/2016–3). Germán Moreno–Arenas acknowledges the Mobility Program of the Universidad Industrial de Santander (UIS), Bucaramanga, Colombia. The research of G. Martínez-Flórez was supported by Grant Proyecto Universidad de Córdoba: Distribuciones de probabilidad asimétrica bimodal con soporte positivo (Colombia), FCB-03-18.

Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS) and the Women's Interagency HIV Study (WIHS) Collaborative Study Group with centers (Principal Investigators) located at The Johns Hopkins Bloomberg School of Public Health (Joseph Margolick); Howard Brown Health Center and Northwestern University Medical School (John Phair, Steven Wolinsky); University of California, Los Angeles (Roger Detels, Oto Martinez-Maza); University of Pittsburgh (Charles Rinaldo); and Data Analysis Center (Lisa Jacobson). The MACS is funded by the National Institute of Allergy and Infectious Diseases, with additional supplemental funding from the National Cancer Institute; and the National Heart, Lung, and Blood

Institute: UO1-AI-35042, UL1-RR025005, UO1-AI-35043, UO1-AI-35039, UO1-AI-35040, and UO1-AI-35041.

The authors are grateful to the anonymous referees for a careful checking of the details and for helpful comments that improved this paper.

Appendix

Let $\rho(0) = \int_{-\infty}^{-\frac{\xi+\eta\delta}{\eta}} \phi(z)\Phi(\alpha z + \alpha\delta)dz$. We have that

$$\Lambda_\xi = -\frac{w_{01}}{\eta}, \quad \Lambda_\eta = \frac{\xi w_{01}}{\eta^2},$$

$$\Lambda_\alpha = \alpha^{-2}\lambda\left[\alpha^{-1}\lambda\delta w_{03} - w_{04}\right], \quad \Lambda_\delta = \lambda w_{03} - w_{01},$$

where $\lambda = \alpha/\sqrt{1+\alpha^2}$, and

$$w_{01} = \frac{1}{\rho(0)}\phi\left(\frac{\xi+\eta\delta}{\eta}\right)\Phi\left(\frac{-\alpha\xi}{\eta}\right),$$

$$w_{02} = \frac{1}{\rho(0)}\phi\left(\frac{\xi+\eta\delta}{\eta}\right)\phi\left(\frac{-\alpha\xi}{\eta}\right),$$

$$w_{03} = \frac{1}{\rho(0)}\phi(\lambda\delta)\Phi\left(\frac{(\xi+\eta\delta)+\xi\alpha^2}{\eta\sqrt{1+\alpha^2}}\right),$$

$$w_{04} = \frac{1}{\rho(0)}\phi(\lambda\delta)\phi\left(\frac{(\xi+\eta\delta)+\xi\alpha^2}{\eta\sqrt{1+\alpha^2}}\right).$$

The elements of the observed information matrix $\boldsymbol{J}_n(\boldsymbol{\theta})$ are

$$j_{\xi\xi} = \sum_{i\in C}\Lambda_{\xi\xi} + \frac{1}{\eta^2}\sum_{i\in U}\left[1 + \alpha^3 w_i z_i + \alpha^2 w_i^2\right],$$

$$j_{\xi\delta} = \sum_{i\in C}\Lambda_{\xi\delta} - \frac{1}{\eta}\sum_{i\in U}\text{sign}(y_i - \xi),$$

$$j_{\xi\eta} = \sum_{i\in C}\Lambda_{\xi\eta} + \frac{2}{\eta^2}\sum_{i\in U}\left[2z_i - \delta\text{sign}(y_i - \xi)\right. \\ \left. - \alpha w_i + \alpha^3 z_i^2 w_i - \alpha^2 z_i w_i^2\right],$$

$$j_{\xi\alpha} = \sum_{i\in C}\Lambda_{\xi\alpha} + \frac{1}{\eta}\sum_{i\in U}\left[w_i - \alpha^2 z_i^2 w_i - \alpha z_i w_i^2\right],$$

$$j_{\alpha\delta} = \sum_{i\in C}\Lambda_{\alpha\delta},$$

$$j_{\eta\eta} = \sum_{i\in C}\Lambda_{\eta\eta} + \frac{1}{\eta^2}\sum_{i\in U}\left[-1 + 3z_i^2 + 2\delta\left|z_i\right|\right. \\ \left. - 2\alpha z_i w_i + \alpha^3 z_i^3 w_i + \alpha^2 z_i^2 w_i^2\right],$$

$$j_{\eta\alpha} = \sum_{i\in C}\Lambda_{\eta\alpha} + \frac{1}{\eta^2}\sum_{i\in U}\left[z_i w_i - \alpha^2 z_i^3 w_i - \alpha z_i^2 w_i^2\right],$$

$$j_{\eta\delta} = \sum_{i\in C}\Lambda_{\eta\delta} - \frac{1}{\eta}\sum_{i\in U}\left|z_i\right|,$$

$$j_{\alpha\alpha} = \sum_{i\in C}\Lambda_{\alpha\alpha} + \sum_{i\in U}\left[\alpha z_i^3 w_i + z_i^2 w_i^2\right],$$

$$j_{\delta\delta} = \sum_{i\in C}\Lambda_{\delta\delta} + n\left(\delta - \frac{\phi(\delta)}{1-\Phi(\delta)}\right)\frac{\phi(\delta)}{1-\Phi(\delta)} + n_1,$$

where

$$\Lambda_{\xi\xi} = \frac{1}{\eta^2}\left[w_{01}^2 - \frac{\xi+\eta\delta}{\eta}w_{01} - \alpha w_{02}\right],$$

$$\Lambda_{\xi\eta} = -\frac{1}{\eta^3}\left[\xi w_{01}^2 - (\xi + \eta(1+\delta))w_{01} - \alpha w_{02}\right],$$

$$\Lambda_{\xi\alpha} = -\frac{1}{\eta^2}\left[\xi w_{02} + \alpha^{-2}\eta\lambda^2 w_{01}(\alpha^{-1}\lambda\delta w_{03} - w_{04})\right],$$

$$\Lambda_{\xi\delta} = -\frac{w_{01}}{\eta^2}\left[(\xi+\eta\delta) + \eta(\lambda w_{03} - w_{01})\right],$$

$$\Lambda_{\eta\eta} = \frac{\xi}{\eta^4}\left[\xi w_{01}^2 + \left(2\xi\eta - \frac{\xi+\eta\delta}{\eta}\right)w_{01} - \alpha\xi w_{02}\right],$$

$$\Lambda_{\eta\alpha} = \frac{\xi}{\eta^3}\left[\xi w_{02} + \alpha^{-2}\eta\lambda^2 w_{01}(\alpha^{-1}\lambda\delta w_{03} - w_{04})\right],$$

$$\Lambda_{\eta\delta} = \frac{w_{01}}{\eta^3}\left[(\xi+\eta\delta) + \eta(\lambda w_{03} - w_{01})\right],$$

$$\Lambda_{\delta\delta} = \lambda\left[\lambda^2\delta w_{03} - w_{04} + \lambda w_{03}^2 - 2w_{01}w_{03}\right] \\ + w_{01}\left[w_{01} - \frac{\xi+\eta\delta}{\eta}\right],$$

$$\Lambda_{\alpha\delta} = \alpha^{-3}\lambda^3 w_{03}(\lambda^2\delta^2 - 1) \\ + \alpha^{-1}\lambda^2 w_{03}(\alpha^{-1}\lambda\delta w_{03} - w_{04}) - \frac{\xi}{\eta}w_{02} \\ - \frac{\lambda^2}{\eta}\left(\frac{\xi - \eta\delta + \alpha^2\xi}{1+\alpha^2}\right)w_{04} \\ - \alpha^{-2}\lambda^2 w_{01}(\alpha^{-1}\lambda\delta w_{03} - w_{04}),$$

$$\Lambda_{\alpha\alpha} = \alpha^{-4}\lambda^4 w_{04}\left[w_{04} - \alpha^{-1}\lambda\delta(\lambda\delta + w_{03})\right. \\ \left. - 2\alpha - \frac{\lambda}{\eta}\left((1+\alpha^2)\xi - \eta\delta\right)\right], \\ + \alpha^{-3}\lambda^3\delta\left[-\alpha^{-2}\lambda^2 w_{04}\left(\frac{\lambda}{\eta}((1+\alpha^2)\xi - \eta\delta)\right.\right. \\ \left.\left. + w_{03}\right) + \alpha^{-3}\lambda^3\delta w_{03}(\lambda\delta + w_{03}) + 3\alpha w_{03}\right].$$

References

[1] H. Chu, S. Gange, X. Li, D. Hoover, C. Liu, J. Chmiel and L. Jacobson. The effect of HAART on HIV RNA trajectory among treatment-naive men and women: a segmental Bernoulli/lognormal random effects model with left censoring, *Epidemiology*, **21**, 25–34, (2010).

[2] X. Li, H. Chu, J.E. Gallant, D.R. Hoover, W.J. Mack, J.S. Chmiel and A. Muñoz. Bimodal virological response to antiretroviral therapy for HIV infection: an application using a mixture model with left censoring, *Journal Epidemiol Community Health*, **60**, 811–818, (2006).

[3] M.F. Schneider, J.B. Margolick, L.P. Jacobson, S. Reddy, O. Martinez-Maza and A. Muõz. Improved estimation of the distribution of suppressed plasma HIV-1 RNA in men receiving effective antiretroviral therapy, *Journal of Acquired Immune Deficiency Syndromes*, **59**, 389–392, (2012).

[4] T.O. Lim, R. Bakri, Z. Morad and M.A. Hamid. Bimodality in blood glucose distribution: is it universal?, *Diabetes Care*, **25**, 2212–2217, (2002).

[5] J. Tobin. Estimation of relationships for limited dependent variables, *Econometrica*, **26**, 24–36, (1958).

[6] R.B. Arellano-Valle, L.M. Castro, G. González-Farías and K.A. Muñoz-Gajardo. Student-t censored regression model: properties and inference, *Statistical Methods & Applications*, **21**, 453–473, (2012).

[7] G. Martínez-Flórez, H. Bolfarine and H. Gómez. The alpha-power Tobit model, *Communications in Statistics-Theory and Methods*, **42**, 633–643, (2013).

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

[9] H. Kim. On a class of two-piece skew-normal distribution, *Statistic*, **39**, 537–553, (2005).

[10] B. Arnold, H. Gómez and H. Salinas. On multiple constraint skewed models, *Statistics*, **43**, 279–293, (2009).

[11] H. Gómez, D. Elal-Olivero, H.S. Salinas and H. Bolfarine. Bimodal extension based on the skew-normal distribution with application to pollen data, *Environmetrics*, **22**, 50–62, (2011).

[12] R.D. Gupta and R.C. Gupta. Analyzing skewed data by power normal model, *Test*, **17**, 197–210, (2008).

[13] G.M. Cordeiro and A.J. Lemonte. The exponentiated generalized Birnbaum-Saunders distribution, *Applied Mathematics and Computation*, **247**, 762–779, (2014).

[14] S.M.T.K. MirMostafaee, M. Mahdizadeh and S. Nadarajah. The beta Lindley distribution, *Journal of Data Science*, **13**, 603–626, (2015).

[15] S.M.T.K. MirMostafaee, M. Mahdizadeh and A.J. Lemonte. The Marshall-Olkin extended generalized Rayleigh distribution: Properties and applications, *Communications in Statistics: Theory and Methods*, **46**, 653–671, (2017).

[16] A. O'Hagan and T. Leonard. Bayes estimation subject to uncertainty about parameter constraints, *Biometrika*, **63**, 201–203, (1976).

[17] S. Durrans. Distributions of fractional order statistics in hydrology, *Water Resources Research*, **28**, 1649–1655, (1992).

[18] A. Pewsey, H.W. Gómez and H. Bolfarine. Likelihood-based inference for power distributions, *Test*, **21**, 775–789, (2012).

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018.

[20] D.B. Owen. Tables for computing bivariate normal probabilities, *The Annals of Mathematical Statistics*, **27**, 1075–1090, (1956).

[21] D.R. Cox and D.V. Hinkley, *Theoretical Statistics*, Chapman and Hall, London UK, (1974).

[22] COBAS® AmpliPrep/COBAS® TaqMan® HIV-1 Test, *Branchburg, NJ: Roche Molecular Systems, Inc.*, version 2.0, (2010).

[23] Amplicor® HIV-1 MONITOR® Test. *Branchburg, NJ: Roche Molecular Systems, Inc.*, version 1.5, (2002).

Germán Moreno-Arenas received the PhD degree in Statistics, at University of São Paulo, located in São Paulo (Brazil). His research interests were in the areas of applied mathematics and statistics including parametric distributions. Unfortunately, during the development of this work, Prof. Germán Moreno-Arenas passed away. RIP.

Guillermo Martínez-Flórez is full professor at the Universidad de Córdoba, located in Montería (Colombia). He is also a visiting professor at the Federal University of Ceara (Brazil). He received the PhD degree in Statistics, at University of São Paulo, located in São Paulo (Brazil). His research interests are in the areas of applied mathematics and statistics including regression models, parametric inference, and distribution theory.

Artur Lemonte is an assistant professor at the Federal University of Rio Grande do Norte, located in Natal (Brazil). He received the PhD degree in Statistics, at University of São Paulo. He has published research articles in reputed international journals of statistics and applied mathematics. His main research interests are: higher order asymptotics, mathematical statistics, regression models, parametric inference, and distribution theory.