Bayesian semiparametric modeling for HIV longitudinal data with censoring and skewness

Bayesian semiparametric modeling for HIV longitudinal data with censoring and skewness

Luis M. Castro[1], Wan-Lun Wang[2], Victor H. Lachos[3], Vanda Inácio de Carvalho [4] and Cristian L. Bayes[5]

Abstract

In biomedical studies, the analysis of longitudinal data based on Gaussian assumptions is common practice. Nevertheless, more often than not, the observed responses are naturally skewed, rendering the use of symmetric mixed effects models inadequate. In addition, it is also common in clinical assays that the patient's responses are subject to some upper and/or lower quantification limit, depending on the diagnostic assays used for their detection. Furthermore, responses may also often present a nonlinear relation with some covariates, such as time. To address the aforementioned three issues, we consider a Bayesian semiparametric longitudinal censored model based on a combination of splines, wavelets, and the skew-normal distribution. Specifically, we focus on the use of splines to approximate the general mean, wavelets for modeling the individual subject trajectories, and on the skew-normal distribution for modeling the random effects. The newly developed method is illustrated through simulated data and real data concerning AIDS/HIV viral loads.

Keywords

Censored longitudinal data; HIV viral load; Mixed-effects models; Semiparametric regression; Skewness.

1 Introduction

In clinical trials of antiretroviral therapy (ARV therapy), HIV-1 RNA measures are collected longitudinally over a period of treatment with the main objective of determining the rates of change in the amount of actively replicating virus. For HIV studies, the analysts often consider longitudinal models for investigating the HIV-1 RNA level (viral load), aiming to understand the HIV pathogenesis and to assess the effectiveness of the therapy. As was mentioned by Ndembi *et al.* [1], considering the HIV-1 RNA as a key primary endpoint in AIDS studies is due to several inherent reasons: (i) the viral load monitoring during the therapy is mostly available, (ii) a failure in the treatment can be defined virologically, and (iii) a new regimen of therapy is recommended as soon as virological rebound occurs. In fact, changes in the HIV-1 RNA responses to ARV therapies are associated with clinical benefits and consequently, the frequent monitoring of this laboratory marker is rapidly becoming the standard of care in routine clinical practice [2].

From a practical viewpoint, the analysis of viral loads can be challenging due to the detection limit considered in some diagnostics assays utilized for the quantification of the HIV-1 RNA levels. The viral load responses are either left or right censored depending upon the diagnostic assays used, where the range of the limit of detection varies from 400 copies/ml for the earlier assays, to 40 copies/ml for the more sophisticated assays. To deal with such limits of detection, censored mixed-effects models are frequently used in the analysis of longitudinal AIDS data. In fact, such models are used to estimate viral load trajectories, as well as, to quantify within and between subject variations in viral load measurements [3]. For instance, in the seminal works of Hughes [4] and Jacqmin-Gadda *et al.* [5], both proposed likelihood based approaches to estimate the parameters of linear mixed-effects models for left and/or right censored Gaussian data.

Some extensions of the two works mentioned above, considering more efficient algorithms for parameter estimation in Gaussian linear mixed-effects models with censored response (LMEC), have been proposed (see, for example, Vaida *et al.* [6], and Vaida and Liu [7]). Recently, several proposals focusing their attention on the study of censored mixed-effects models under non-Gaussian data have appeared in the literature. For example, Matos *et al.* [8] proposed an EM algorithm for linear and nonlinear mixed-effects models with censored response (LMEC/NLMEC) using the multivariate Student's-t distribution

[1] Department of Statistics, Pontificia Universidad Católica de Chile, Chile
[2] Department of Statistics, Graduate Institute of Statistics and Actuarial Science, Feng Chia University, Taiwan
[3] Department of Statistics, University of Connecticut, US
[4] School of Mathematics, University of Edinburgh, UK
[5] Department of Sciences, Pontificia Universidad Católica del Perú

Corresponding author:
Address for correspondence: Luis M. Castro, Department of Statistics, Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, Chile.

The research herein was performed in part while Luis M. Castro and Victor H. Lachos were visiting the Department of Science at Pontificia Universidad Católica del Perú.
Email: mcastro@mat.puc.cl

while, in turn, Lachos *et al.*[9] adopted a Bayesian approach to carry out posterior inference for censored linear and nonlinear mixed-effects models considering a class of thick-tail distributions (the so called normal/independent proposed by Lange and Sinsheimer[10]) as the joint distribution of the error term and random effects. On the other hand, and under a Bayesian framework, Bandyopadhyay *et al.*[11] studied censored linear mixed-effects models considering both skewness and heavy tails, while Bandyopadhyay *et al.*[12] proposed a skewed censored nonlinear mixed-effect model under the presence of measurement error.

Another important feature of the HIV-1 RNA measures is that, in general, the relationship between the viral load and certain covariates, such as time, is nonlinear and thus, considering a parametric model (with a known linear or nonlinear function) can be too restrictive, possibly resulting in misleading conclusions. Moreover, even in circumstances where transformations and/or quadratic terms can be used to handle nonlinearities, their use may require a considerable expertise[13]. To address this issue, semiparametric extensions have been proposed in the context of non-censored models (see for example[14,15] and references therein). Recently, Castro *et al.*[16] proposed the study of censored models under a Bayesian semiparametric scheme but in the univariate case and considering heavy-tailed distributions. However, to the best of our knowledge, there are no works considering the study of censored mixed-effects models using, simultaneously, semiparametric techniques such as splines and wavelets, and the effect of skewness. This paper therefore proposes flexible Bayesian inference for a censored mixed-effects model, based on splines to approximate the nonlinear general mean, wavelets for modeling the individual subject trajectories, and on the the skew-normal (SN) distribution for modeling the random effects. This paper can be regarded as an extension of Ibacache *et al.*[14] and Castro *et al.*[15,16], where partially linear mixed-effects models were considered.

The rest of the paper is organized as follows. Section 2 describes the multivariate SN distribution and some of its properties, as well as, the HIV dynamics and the motivating AIDS/HIV data sets. In Section 3 we briefly discuss the penalized wavelet-based approach. Section 4 presents our semiparametric censored mixed-effects model, while in Section 5 its Bayesian formulation is developed. The application of the proposed method to two motivating data sets of HIV viral loads is presented in Section 6 and in Section 7 we present an in-depth simulation study. We conclude in Section 8 with some future research directions.

2 Preliminaries

2.1 Multivariate SN distribution

In this section we present a review of the multivariate skew-normal (SN) distribution, including some of its properties. A further discussion about this distribution can be found in Azzalini[17]. A random vector $\mathbf{Y}$ has a multivariate SN distribution with $p \times 1$ location vector $\boldsymbol{\mu}$, $p \times p$ positive definite scale matrix $\boldsymbol{\Sigma}$, and $p \times 1$ skewness parameter vector $\boldsymbol{\lambda}$, if its probability density function (*pdf*) is given by

$$f(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \tag{1}$$

where $\phi_p(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the *pdf* of the p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariate matrix $\boldsymbol{\Sigma}$, $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\Phi(\cdot)$ is the cumulative distribution function (*cdf*) of the standard normal distribution. For referring to the distribution of $\mathbf{Y}$, we use $\mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. When $\boldsymbol{\lambda} = \mathbf{0}$, the distribution of $\mathbf{Y}$ reduces to a multivariate normal (N) distribution $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

An interesting property of the multivariate SN random vector $\mathbf{Y}$ is its stochastic representation in terms of normal random quantities. It follows from Lachos *et al.* [18] that,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Delta} T + \boldsymbol{\Psi}^{1/2}\mathbf{T}_1, \tag{2}$$

where $T = |T_0|$, $T_0 \sim \mathrm{N}_1(0,1)$, and $\mathbf{T}_1 \sim \mathrm{N}_p(\mathbf{0}, \mathbf{I}_p)$ are independent, $|\cdot|$ denotes the absolute value, and

$$\boldsymbol{\Delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}, \quad \boldsymbol{\Psi} = \boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p - \boldsymbol{\delta}\boldsymbol{\delta}^\top)\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma} - \boldsymbol{\Delta}\boldsymbol{\Delta}^\top, \tag{3}$$

with $\boldsymbol{\delta} = \boldsymbol{\lambda}/\sqrt{1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda}}$. Note that $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$ can be obtained as follows:

$$\boldsymbol{\lambda} = \frac{(\boldsymbol{\Psi} + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)^{-1/2}\boldsymbol{\Delta}}{[1 - \boldsymbol{\Delta}^\top(\boldsymbol{\Psi} + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top)^{-1}\boldsymbol{\Delta}]^{1/2}} \quad \text{and} \quad \boldsymbol{\Sigma} = \boldsymbol{\Psi} + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top.$$

The next result establishes the marginal-conditional decomposition of the multivariate SN random vector $\mathbf{Y}$. Its proof can be found in Bandyopadhyay *et al.* [11].

Proposition 1. *Suppose that the random vector* $\mathbf{Y}$ *is partitioned as* $\mathbf{Y}^\top = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top)^\top$ *with dimensions* p_1 *and* p_2 *($p_1 + p_2 = p$), respectively. Let*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \ \boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top, \ \boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top, \ \boldsymbol{\upsilon} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\lambda} = (\boldsymbol{\upsilon}_1^\top, \boldsymbol{\upsilon}_2^\top)^\top$$

be the corresponding partitions of $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$, *and* $\boldsymbol{\upsilon}$. *If* $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, *then the conditional cdf of* $\mathbf{Y}_2 \mid \mathbf{Y}_1 = \mathbf{y}_1$ *is given by*

$$\Pr(\mathbf{Y}_2 \le \mathbf{y}_2 \mid \mathbf{Y}_1 = \mathbf{y}_1) = \frac{\Phi_{p_2+1}\left(((\mathbf{y}_2 - \boldsymbol{\mu}_{2.1})^\top, z_1)^\top; \mathbf{0}, \boldsymbol{\Omega}_N\right)}{\Phi(\widetilde{\boldsymbol{\upsilon}}^\top(\mathbf{y}_1 - \boldsymbol{\mu}_1))},$$

where $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$, $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, $z_1 = (\boldsymbol{\upsilon}_1 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\upsilon}_2)^\top(\mathbf{y}_1 - \boldsymbol{\mu}_1)$, $\widetilde{\boldsymbol{\upsilon}} = \dfrac{\boldsymbol{\upsilon}_1 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\upsilon}_2}{\sqrt{1 + \boldsymbol{\upsilon}_2^\top\boldsymbol{\Sigma}_{22.1}\boldsymbol{\upsilon}_2}}$, *and* $\boldsymbol{\Omega}_N = \begin{pmatrix} \boldsymbol{\Sigma}_{22.1} & -\boldsymbol{\Sigma}_{22.1}\boldsymbol{\upsilon}_2 \\ -\boldsymbol{\upsilon}_2^\top\boldsymbol{\Sigma}_{22.1} & 1 + \boldsymbol{\upsilon}_2^\top\boldsymbol{\Sigma}_{22.1}\boldsymbol{\upsilon}_2 \end{pmatrix}$.

2.2 The HIV dynamic

The main purpose of this section is to detail how HIV viral loads have been traditionally modeled in the statistical literature. In AIDS research, it is believed that the relationship between the virologic response and some immunologic covariates (e.g. $\mathrm{CD4}^+$ cells) and time, is nonlinear. As was noted by Wu and Ding [19], Wu [20], and recently by Bandyopadhyay *et al.* [12], a bi-phasic nonlinear model, associated with the $\mathrm{CD4}^+$cell counts and time, for modeling the viral load is given by

$$y_{ij} = \log_{10}(P_{1i}e^{-\psi_{1ij}t_{ij}} + P_{2i}e^{-\psi_{2ij}t_{ij}}) + \epsilon_{ij}, \tag{4}$$

$$\beta_{1ij} = \log(P_{1i}) = \beta_1 + b_{1i}, \ \ \beta_{2ij} = \psi_{1ij} = \beta_2 + b_{2i}, \tag{5}$$

$$\beta_{3ij} = \log(P_{2i}) = \beta_3 + b_{3i}, \ \ \beta_{4_{ij}} = \psi_{2ij} = \beta_4 + \beta_5\mathrm{CD4}_{ij}^+ + b_{4i}, \tag{6}$$

where y_{ij} is the $\log_{10}$-transformed viral load $V(t_{ij})$ for the ith subject at time t_{ij} $(i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n_i)$, P_{1i} and P_{2i} are the baseline viral loads, ψ_{1ij} and ψ_{2ij} are the first- and second-phases of viral decay rates representing the minimum turnover rate of productively infected cells and latently long-lived infected cells, respectively, and ϵ_{ij} is the within-subject random error. Additionally, $\boldsymbol{\beta}_{ij} = (\beta_{1ij}, \beta_{2ij}, \beta_{3ij}, \beta_{4ij})^\top$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\top$ are the subject-level, and population-level parameters, respectively; CD4^+_{ij} indicates the CD4^+ cell counts at time t_{ij}, and $\mathbf{b}_i = (b_{1i}, \ldots, b_{4i})^\top$ are the subject-level random effects. As it was already mentioned, equation (4) is derived from the bi-phasic exponential decay model $V(t) = P_1 e^{-\psi_1 t} + P_2 e^{-\psi_2 t}$ proposed by Wu and Ding [19].

The HIV dynamic is typically fitted a model framework considering the structure given in (4)–(6) and under the assumption that

$$\boldsymbol{\epsilon}_i \mid \sigma^2_\epsilon \overset{ind.}{\sim} \text{N}_{n_i}(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_{n_i}) \quad \text{and} \quad \mathbf{b}_i \mid \mathbf{D} \overset{ind.}{\sim} \text{N}_4(\mathbf{0}, \mathbf{D}), \quad i = 1, \ldots, n,$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})^T$, $\sigma^2_\epsilon > 0$, $\mathbf{D}$ is a 4×4 positive-definite matrix, and $\mathbf{I}_q$ denotes the $q \times q$ identity matrix. This model is the so-called normal NLMEC (N-NLMEC) model.

To overcome the sometimes unrealistic assumption of normality of the random effects distribution, a more robust model, named as the skew-normal nonlinear mixed-effects model with censored response (SN-NLMEC) has been proposed (see Bandyopadhyay *et al.* [12]). Under this model, it is assumed that

$$\boldsymbol{\epsilon}_i \mid \sigma^2_\epsilon \overset{ind.}{\sim} \text{N}_{n_i}(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_{n_i}) \quad \text{and} \quad \mathbf{b}_i \mid \mathbf{D}, \boldsymbol{\lambda} \overset{ind.}{\sim} \text{SN}_4(c\boldsymbol{\Delta}, \mathbf{D}, \boldsymbol{\lambda}), \quad i = 1, \ldots, n, \tag{7}$$

where $\text{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ denotes the multivariate SN distribution, $c = -\sqrt{2/\pi}$, $\boldsymbol{\Delta} = \mathbf{D}^{1/2}\boldsymbol{\delta}$ with $\boldsymbol{\delta} = \boldsymbol{\lambda}/\sqrt{1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda}}$, $\mathbf{D} = \mathbf{D}(\boldsymbol{\varphi})$ is the $q \times q$ dispersion matrix of $\mathbf{b}_i$ depending on the unknown and reduced vector parameter $\boldsymbol{\varphi}$ (e.g., the upper triangular elements in the unstructured case), σ^2_ϵ is the unknown within-subject scale parameter, and $\boldsymbol{\lambda}$ is the 4×1 skewness vector parameter corresponding to the random effects $\mathbf{b}_i$. Note that, in this setup, $E(\mathbf{b}_i \mid \mathbf{D}, \boldsymbol{\lambda}) = E(\boldsymbol{\epsilon}_i \mid \sigma^2_\epsilon) = \mathbf{0}$. Thus, the within-subject errors are symmetrically distributed around zero, whereas the random effects are asymmetric with mean zero. It can be shown that $Cov(\mathbf{b}_i, \boldsymbol{\epsilon}_i \mid \sigma^2_\epsilon, \mathbf{D}, \boldsymbol{\lambda}) = \mathbf{0}$, and thus, $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ are uncorrelated.

In order to include the censoring effect in the statistical modeling of the HIV viral load profiles, let the observed data for the ith subject to be given by $(\mathbf{Q}_i, \mathbf{C}_i)$, where $\mathbf{Q}_i = (Q_{i1}, \ldots, Q_{in_i})^\top$ represents the vector of uncensored readings or censoring levels and $\mathbf{C}_i = (C_{i1}, \ldots, C_{in_i})^\top$ is the vector of censoring indicators, where one has

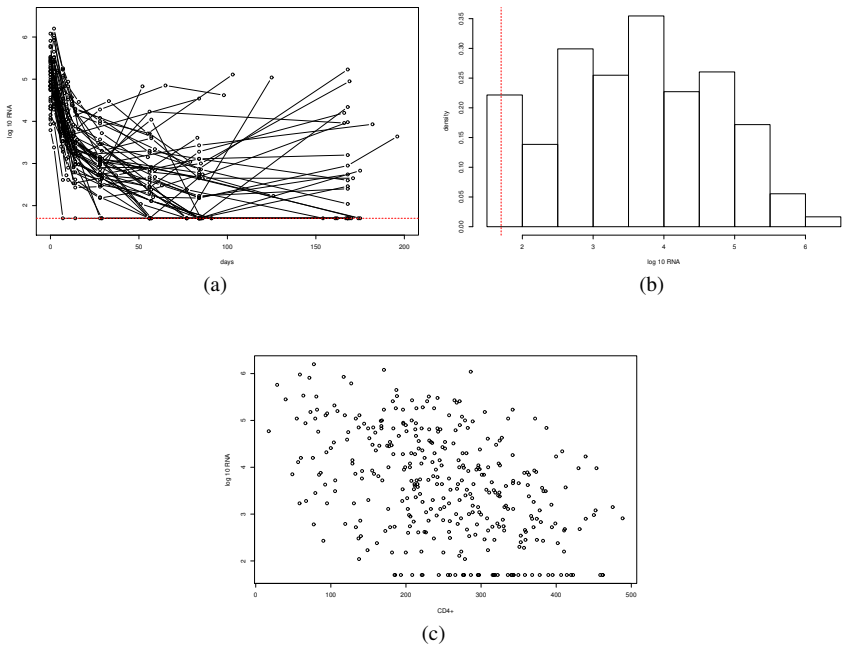
$$\begin{aligned} y_{ij} &\leq Q_{ij} \quad \text{if} \quad C_{ij} = 1, \\ y_{ij} &= Q_{ij} \quad \text{if} \quad C_{ij} = 0, \end{aligned} \tag{8}$$

i.e., $C_{ij} = 1$ if y_{ij} is left censored, case one only knows that the true observation y_{ij} is less than or equal to the observed quantity Q_{ij}, for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, n_i\}$.

2.3 The ACTG 315 clinical trial

We consider a dataset from the AIDS clinical trial group 315 (ACTG 315) protocol including 46 HIV-1 infected patients treated with a potent antiretroviral drug cocktail based on protease inhibitor ritonavir and reverse transcriptase inhibitor drugs (zidovudine and lamivudine). The aim of this antiretroviral regimen

is to show that immunity can be partially restored in people with moderately advanced HIV disease. This dataset was previously analyzed by Wu [20] in the context of a LMEC from a classic point of view and recently by Lachos *et al.* [21] in the context of a nonlinear mixed effects-model from a Bayesian perspective. From a maximum likelihood viewpoint, Lin and Wang [22] analyzed the viral load and CD4$^+$ cells simultaneously by using a multivariate SN linear mixed model, whereas Wang [23] applied a multivariate nonlinear mixed model to analyze this same dataset. Further, and still using this dataset, Matos *et al.* [24] proposed a censored nonlinear mixed-effects model using a damped exponential correlation structure for the error term. For a more detailed description of the HIV/AIDS study, we refer the interested reader to Lederman *et al.* [25], and Connick *et al.* [26].



(a) (b)

(c)

Figure 1. ACTG 315 dataset. (a) Viral load trajectories (in $\log_{10}$ scale). (b) Raw density histogram of viral loads ($\log_{10}$ scale). The vertical red dotted line indicates the censoring threshold. (c) Scatter plot of the CD4$^+$ cell counts against viral loads ($\log_{10}$ scale).

The viral loads were quantified at days $0, 2, 7, 10, 14, 21, 28, 56, 84, 168$ and 196 after start treatment, generating 361 observations. CD4$^+$ and CD8$^+$ cell counts were also measured along with viral loads. Measurements below the detectable threshold of 100 copies/mL (40 out of 361, 11%) were considered left-censored, and the censoring mechanism assumed independent of the complete data. Figure 1 (a) displays the individual profiles of the viral loads. As it can be appreciated, the HIV-1 RNA levels change over time in a nonlinear manner. Moreover, a variation in the intercept among individuals is also observed. In Figure 1 (b) a raw histogram of the viral load is shown, evidencing the presence of skewness in the

data, while in Figure 1 (c) we display a scatter plot of the viral load and CD4$^+$ cell counts, showing that the virologic and immunologic markers are negatively correlated.

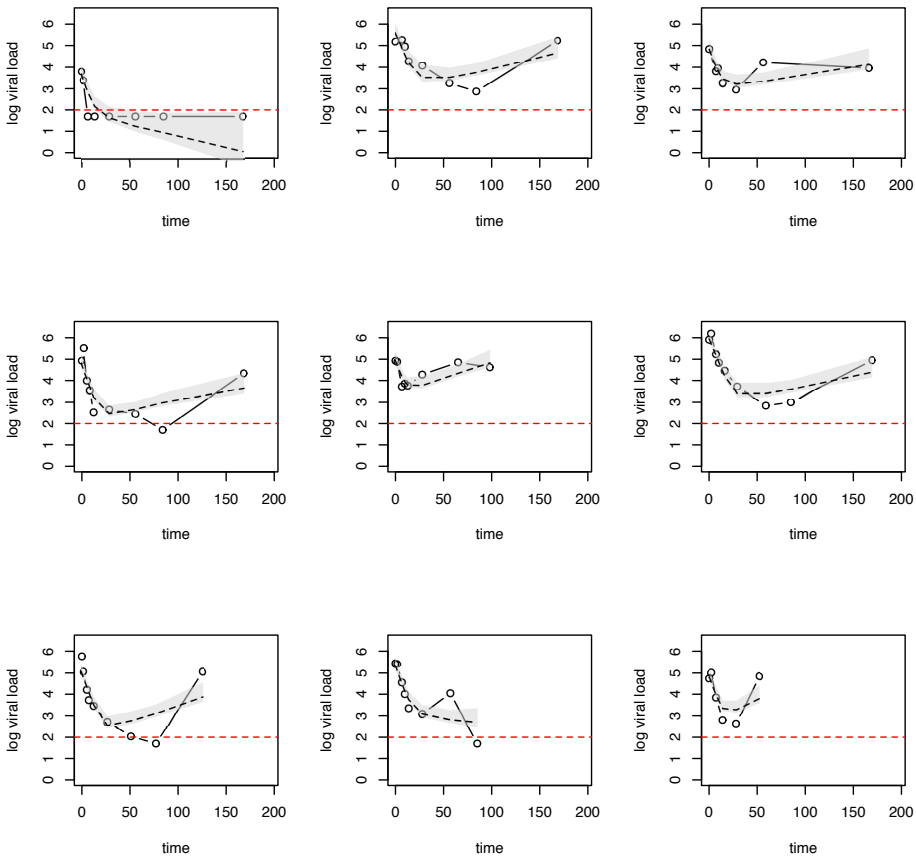


Figure 2. ACTG 315 data. Viral loads in $\log_{10}$ scale (solid line) for 9 randomly chosen patients and estimated trajectories (dotted line) under the SN-NLMEC model. The red dotted line indicates the censoring level. The grey area indicates a 95% credible band for the estimated trajectory.

In Figure 2 we present the estimated trajectories for 9 randomly chosen patients after fitting the SN-NLMEC model[12]. From this figure, it is clear that such model provides biased estimated trajectories. Moreover, the assumed bi-exponential model does not seem appropriate for the ACTG 315 data set since it assumes that the therapy works uniformly over time. However, some subjects present "U-shaped" trajectories, showing evidence about resistance to the treatment through the follow-up (see Figure 1 (a)). This fact motivates the use of a flexible semiparametric approach based on wavelets for modeling

the individual trajectories. As it was noted by Christensen *et al.*[27] (chapter 15), wavelets are useful for modeling functions whose behavior changes abruptly at different locations.

2.4 The A5055 clinical trial

The second dataset considered is obtained from the A5055 clinical trial. The study considers 44 HIV infected patients treated with one of the two potent ARV therapies, namely, indinavir 800 mg plus ritonavir 200 mg administered twice daily (arm 1) and indinavir 400 mg plus ritonavir 400 mg administered twice daily (arm 2). This dataset was previously analyzed by Wang *et al.*[28], and Lin and Wang[29] using multivariate t linear and nonlinear mixed-effects models with censoring, respectively, and recently by Lachos *et al.*[30] using scale mixtures of normal distributions in censored nonlinear mixed effects models. More details about this dataset can be found in Acosta *et al.*[31].

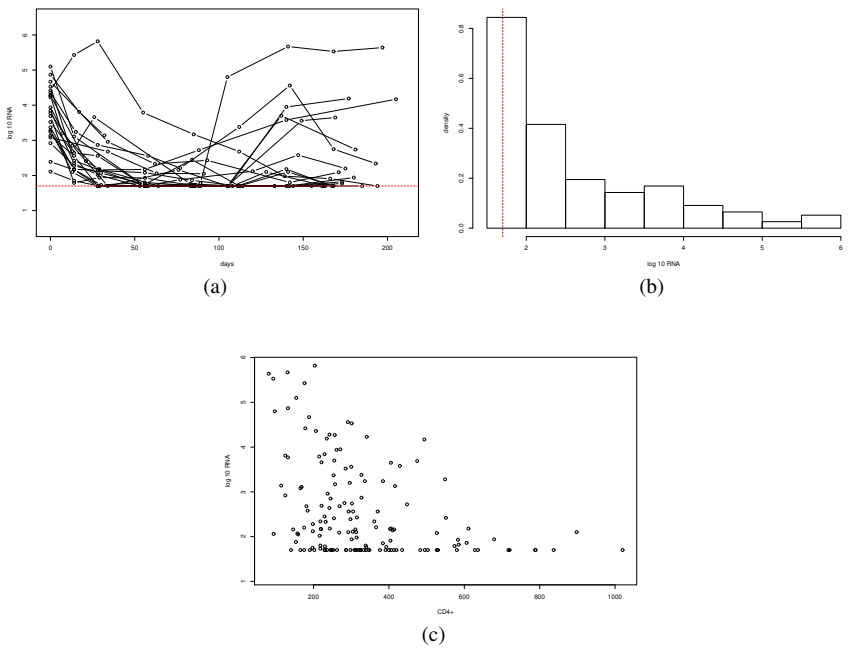


Figure 3. A5055 dataset. (a) Viral load trajectories (in $\log_{10}$ scale). (b) Raw density histogram of viral loads ($\log_{10}$ scale). The vertical red dotted line indicates the censoring threshold. (c) Scatter plot of the CD4$^+$ cell counts against viral loads ($\log_{10}$ scale).

The dataset includes information about viral load measurements (in copies per milliliter), CD4$^+$ and CD8$^+$ cell counts measured roughly at days $0, 7, 14, 28, 56, 84, 112, 140$, and 168 of follow-up for each patient. In this case the lower detection limit for the viral load is 50 copies/milliliter, and therefore 33.5% (106 out of 316) of measurements lie below the limits of assay quantification (left-censored). For the data analysis, we consider only the information corresponding to the patients treated under arm 1 (22

subjects, 154 observations and 67 of them lying below the detection limit). Figure 3 (a) and (b) show the individual trajectories and raw histogram of the viral loads, respectively. Particularly, Figure 3 (b) evidences the presence of skewness in the data. Figure 1 (c) provides a scatter plot of the viral load and $CD4^+$ cell counts, where it can be noted that the viral load and $CD4^+$ cell counts tend to exhibit negative correlation.

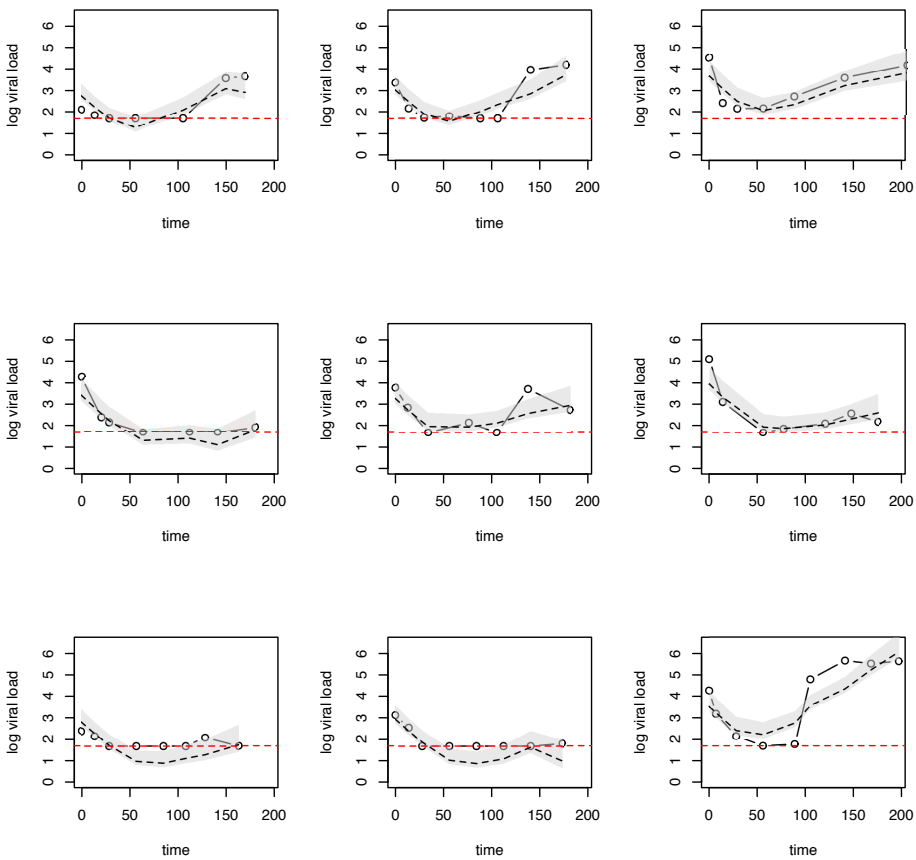


Figure 4. A5055 dataset. Viral loads in $\log_{10}$ scale (solid line) for 9 randomly chosen patients and estimated trajectories (dotted line) under the SN-NLMEC model. The red dotted line indicates the censoring level. The grey area indicates a 95% credible band for the estimated trajectory.

In Figure 4 it is presented the estimated trajectories through fitting the SN-NLMEC model proposed in (4)–(7) for 9 randomly chosen patients. As in the previous trial (ACTG 315), the SN-NLMEC model

provides biased estimated trajectories, reinforcing thus the need to consider a flexible semiparametric approach for modeling the individual viral load levels.

3 Semiparametric modeling via penalized wavelets

Wavelets are families of orthonormal basis commonly used to represent functions in a parsimonious way. According to Morris and Carroll[32], a function g can be represented through a wavelet series as follows:

$$g(t) = \sum_{j=j_0}^{\infty} \sum_{k \in \mathbb{Z}} d_{jk} \psi_{jk}(t),$$

where $t \in [0,1]$, $\psi_{jk}(t) = 2^{j/2}\, \psi(2^j\, t - k)$ is an orthogonal wavelet basis and $d_{jk} = \int g(t)\psi_{jk}(t)\, dt$ is the associated wavelet coefficient. This coefficient allows to describe particular features of the function g at locations indexed by k and frequencies indexed by j. An important characteristic of the wavelet representation of the function g is that all of the oscillating features of this function can be captured in a simply manner, making it a powerful tool for the analysis of jagged functions.

Wavelets have been successfully applied in the semiparametric regression literature (see, for instance, Silverman[33], Abramovich *et al.*[34], Morris and Carroll[32], Ko *et al.*[35], Wand and Ormerod[36], and Hu and Yuan[37], among many others). In particular, Wand and Ormerod[36] proposed the use of penalized wavelets as an alternative to spline-based strategies, particularly in the case of jagged trajectories. To make ideas clear, let us consider the univariate regression model given by

$$y_i = g(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where we do not assume any particular parametric structure for g. Following Wand and Ormerod[36], we consider

$$g(x) = \beta_0 + \sum_{w=1}^{W} u_w\, z_w(x),$$

where $\{z_w(\cdot) : 1 \leq w \leq W\}$ is an appropriate set of wavelets basis functions constructed over equally-spaced grids on $[0,1)$ of length R, with R being a power of 2. These functions take the form

$$z_w(x) = z_w^U\left(\frac{x-a}{b-a}\right), \quad 1 \leq w \leq W,$$

where a and b are the end-points of the compact inverval $[a,b]$ over which the basis functions are non-zero and

$$z_w^U(x) \approx \{1 - (x\,R - \lfloor x\,R \rfloor)\}\, z_w^U(\lfloor x\,R \rfloor / R) + (x\,R - \lfloor x\,R \rfloor) z_w^U((\lfloor x\,R \rfloor + 1)/R),$$

with $z_w^U(1) \equiv z_w^U\left(\frac{R-1}{R}\right)$. For the choice of the wavelets basis functions, the Daubechies family[38] is considered mainly because they allow for an efficient approximation of jagged trajectories[36], but also because they are freely available through the R package[39] `wavethresh`[40]. Although this family of functions do not admit explicit algebraic expressions, they can be obtained recursively. In general, R is a very large number such as $R = 2^{14} = 16384$ and $W = 2^L - 1$, where L is the level of the wavelet and commonly set to 5.

4 A semiparametric wavelet-based censored model

In this section we introduce our flexible semiparametric censored model based on wavelets for modeling the subject-specific trajectories and on splines for modeling the general mean. In what follows, the random effects follow a multivariate SN distribution while the error term remains multivariate normally distributed. It is worth noting that the use of wavelets in longitudinal models has been quite scarce (available references include Aykroyd and Mardia[41], Morris *et al.*[42], Morris and Carroll[32], Zhao and Wu[43], Wand and Ormerod[36]).

4.1 The model

In order to capture the irregular trajectories from the two AIDS studies, we consider a semiparametric strategy based on the combination of penalized splines and wavelets; our main aim is to flexible model the viral load as a function of time. Following Dúrban *et al.*[44], we consider that the jth response for the ith subject is modeled by

$$y_{ij} = f(t_{ij}) + g_i(t_{ij}) + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i. \tag{9}$$

In the spirit of Wand and Ormerod[45], we consider a linear penalized spline for modeling the general mean of the subjects across time, given by

$$f(t) = \alpha_0 + \alpha_1\, t + \sum_{s=1}^{S} u_s^{gbl}\, z_s^{gbl}(t), \quad u_s^{gbl} \mid \sigma^2_{gbl} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2_{gbl}),$$

where u_s^{gbl} and $z_s^{gbl}(t)$, $s = 1, \ldots, S$, are the spline coefficients to be estimated and the basis of the spline, respectively, with S being the number of knots.

On the other hand, a penalized wavelet is considered for approximating the $g_i(t)$ functions given by:

$$g_i(t) = b_i + \sum_{w=1}^{W} u_{iw}^{sbj}\, z_w^{sbj}(t),$$

where b_i is a subject-specific random intercept, and u_{iw}^{sbj} and z_w^{sbj} are the wavelet coefficients for each subject and the basis of the wavelet, respectively. In this case, W is the number of levels of the wavelet and

$$h(u_{iw}^{sbj} \mid \sigma_{sbj}, \gamma_{iw}) = \frac{\gamma_{iw}}{2\sigma_{sbj}} \exp\left\{ -\frac{|u_{iw}^{sbj}|}{\sigma_{sbj}} \right\} + (1 - \gamma_{iw}\delta_0(u_{iw}^{sbj})), \tag{10}$$

where γ_{iw} is a random variable over $[0, 1]$ and δ_0 is the dirac's delta.

Following Castro *et al.*[15], let $\mathbf{f}_i = (f(t_1^0), \ldots, f(t_{r_i}^0))^\top$ be a $r_i \times 1$ vector with $t_1^0, \ldots, t_{r_i}^0$ being the distinct and ordered values of t_{ij}, and $\mathbf{N}_i$ an $(n_i \times r_i)$ incidence matrix whose (j, s)-th element is equal to the indicator function $I(t_{ij} = t_s^0)$ for $j = 1, \ldots, n_i$ and $s = 1, \ldots, r_i$. Then, considering the mixed model representation of the penalized wavelet[36], we have that our proposed model (9) can be written in matrix form as:

$$\mathbf{y}_i = \mathbf{b}_i + \mathbf{N}_i\, \mathbf{f}_i + \mathbf{Z}\, \mathbf{u}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{11}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ is a $n_i \times 1$ vector of observed continuous responses for the ith subject, $\mathbf{b}_i = \mathbf{1}_{n_i} b_i$ is a $n_i \times 1$ vector of random effects, $\mathbf{Z} = (z_1^{sbj}(\mathbf{t})^\top, \ldots, z_W^{sbj}(\mathbf{t})^\top)^\top$ is a $(n_i \times W)$ design matrix corresponding to the basis wavelets with $z_k^{sbj}(\mathbf{t}) = (z_k^{sbj}(t_1), \ldots, z_k^{sbj}(t_{n_i}))^\top$, $\mathbf{u}_i = (u_{i1}^{sbj}, \ldots, u_{iW}^{sbj})^\top$ a $W \times 1$ vector, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^\top$ is the vector of random errors.

4.2 Distribution of random error and random effects

Our distributional assumptions for the random error and random effect of model (11) are as follows:

$$\boldsymbol{\epsilon}_i \mid \sigma_\epsilon^2 \overset{ind.}{\sim} \mathrm{N}_{n_i}(0, \sigma_\epsilon^2 \mathbf{I}_{n_i}), \quad \text{and} \quad b_i \mid \sigma_b^2, \lambda \overset{ind.}{\sim} \mathrm{SN}(c\Delta, \sigma_b^2, \lambda), \text{ with } c = -\sqrt{\frac{2}{\pi}}, \tag{12}$$

where $\Delta = \sigma_b \lambda / \sqrt{1+\lambda^2}$. Is important to remark that $\mathrm{E}[\mathbf{b}_i] = \mathrm{E}[\boldsymbol{\epsilon}_i] = \mathbf{0}$. Consequently, the within-subject errors $\boldsymbol{\epsilon}_i$ are symmetrically distributed, while the distribution of random effects is asymmetric (skewed) with zero-mean.

Using results from Bandyopadhyay *et al.*[11], and in the case of the complete data, the marginal distribution of $\mathbf{Y}_i$ (after integrating out the random effect $\mathbf{b_i}$), is given by:

$$\mathbf{Y}_i \mid \sigma_\epsilon^2, \lambda \overset{ind.}{\sim} \mathrm{SN}_{n_i}(\widetilde{\mathbf{X}}_i \widetilde{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_i, \bar{\boldsymbol{\lambda}}_i);$$

where $\widetilde{\mathbf{X}}_i = \begin{pmatrix} \mathbf{X}_i & c\mathbf{1}_{n_i} \end{pmatrix}$, $\mathbf{X}_i = \begin{pmatrix} \mathbf{N}_i & \mathbf{Z} \end{pmatrix}$, $\widetilde{\boldsymbol{\beta}} = \left(\mathbf{B}^\top, \Delta\right)^\top$, $\mathbf{B} = (\mathbf{f}_i^\top, \mathbf{u}_i^\top)^\top$, $\boldsymbol{\Sigma}_i = \sigma_\epsilon^2 \mathbf{I}_{n_i} + \sigma_b^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^\top$, $\bar{\boldsymbol{\lambda}}_i = \dfrac{\sigma_b \lambda}{\sqrt{1+\zeta^2 \Lambda_i}} \boldsymbol{\Sigma}_i^{-1/2} \mathbf{1}_{n_i}$, $\Lambda_i = (\sigma_b^{-2} + n_i\, \sigma_\epsilon^{-2})^{-1}$ and $\zeta = \lambda / \sigma_b$.

4.3 Censored observations

In order to include the effect of censoring in the proposed model, we denote the observed data for the i-th subject as $\mathfrak{D}_{obs,i} = \{\mathbf{Q}_i, \mathbf{C}_i\}$, where $\mathbf{Q}_i$ and $\mathbf{C}_i$ satisfies Equation 8.

In our examples the data are left-censored, however extensions to arbitrary censoring are immediate under a Bayesian framework. Further, for example, the right-censored problem can be represented by a left-censored problem by simultaneously transforming the response y_{ij} and censoring level Q_{ij} to $-y_{ij}$ and $-Q_{ij}$, respectively.

Under the censoring scheme (8), we have that $\mathbf{Y}_i$ follows a truncated multivariate SN distribution. More specifically,

$$\mathbf{Y}_i \mid \sigma_\epsilon^2, \lambda \overset{ind.}{\sim} \mathrm{TSN}_{n_i}(\widetilde{\mathbf{X}}_i \widetilde{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_i, \bar{\boldsymbol{\lambda}}_i; \mathbb{A}_i),$$

where $\mathrm{TSN}_{n_i}(.; \mathbb{A}_i)$ denotes the multivariate SN distribution truncated on the interval $\mathbb{A}_i = A_{i1} \times \ldots, \times A_{ini}$, with A_{ij} being the interval $(-\infty, \infty)$ if $C_{ij} = 0$ and $(-\infty, Q_{ij}]$ if $C_{ij} = 1$.

4.4 Likelihood function

Let $\mathbf{y}_i = (\mathbf{y}_i^o, \mathbf{y}_i^c)$, where $\mathbf{y}_i^o$ is the n_i^o-vector of observed outcomes and $\mathbf{y}_i^c$ is the n_i^c-vector of censored observations for subject i with $n_i = n_i^o + n_i^c$, such that $C_{ij} = 0$ for all elements in $\mathbf{y}_i^o$, and 1 for all elements in $\mathbf{y}_i^c$. Denoting by $\boldsymbol{\mu}_i = \widetilde{\mathbf{X}}_i \widetilde{\boldsymbol{\beta}}$ and after reordering the elements in $\mathbf{Q}_i$, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$,

these vectors and matrices can be partitioned as: $\mathbf{Q}_i = vec(\mathbf{Q}_i^o, \mathbf{Q}_i^c)$, $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_i^{o\top}, \boldsymbol{\mu}_i^{c\top})^\top$, and $\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_i^{oo} & \boldsymbol{\Sigma}_i^{oc} \\ \boldsymbol{\Sigma}_i^{co} & \boldsymbol{\Sigma}_i^{cc} \end{pmatrix}$, where $vec(\cdot)$ corresponds to the function which stacks vectors or matrices of the same number of columns. Consequently, the likelihood function for the i-th subject is given by

$$L_i(\boldsymbol{\theta}) = f(\mathbf{y}_i^o \mid \boldsymbol{\theta}) \Pr(\mathbf{y}_i^c \leq \mathbf{Q}_i^c \mid \mathbf{y}_i^o, \boldsymbol{\theta}), \tag{13}$$

where $\boldsymbol{\theta}$ is the parameter of interest. It is important to emphasize that, the multivariate SN distribution is closed under marginalization. Thus, $f(\mathbf{y}_i^o \mid \boldsymbol{\theta})$ will be a $\text{SN}_{n_i^o}(\boldsymbol{\mu}_i^o, \boldsymbol{\Sigma}_i^{oo}, \boldsymbol{\Sigma}_i^{oo\,1/2}\widetilde{\boldsymbol{\upsilon}}_i)$, with $\widetilde{\boldsymbol{\upsilon}}_i = \dfrac{\boldsymbol{\upsilon}_{1i} + \boldsymbol{\Sigma}_i^{oo\,-1}\boldsymbol{\Sigma}_i^{oc}\boldsymbol{\upsilon}_{2i}}{\sqrt{1 + \boldsymbol{\upsilon}_{2i}^\top \boldsymbol{\Sigma}_i^{cc.o}\boldsymbol{\upsilon}_{2i}}}$, $\boldsymbol{\Sigma}_i^{cc.o} = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co}\boldsymbol{\Sigma}_i^{co\,-1}\boldsymbol{\Sigma}_i^{oc}$, and $\boldsymbol{\upsilon}_i = \boldsymbol{\Sigma}_i^{-1/2}\bar{\boldsymbol{\lambda}}_i = (\boldsymbol{\upsilon}_{1i}^\top, \boldsymbol{\upsilon}_{2i}^\top)^\top$. Note that, $\Pr(\mathbf{y}_i^c \leq \mathbf{Q}_i^c \mid \mathbf{y}_i^o, \boldsymbol{\theta})$ is written using Proposition 1.

5 Bayesian approach

In this section we provide the hierarchical representation of our proposed model, including the specification of prior distributions and some details about model comparison tools used for comparing the parametric and semiparametric approaches.

5.1 Hierarchical representation

The proposed model specified in (11) can be written hierarchically; using the stochastic representation of the multivariate SN distribution provided in (2) and the parameterization (3), we have that

$$\mathbf{y}_i \mid b_i, \mathbf{C}_i, \mathbf{Q}_i, T_i = t_i \overset{ind.}{\sim} \text{TN}_{n_i}(\mathbf{b}_i + \mathbf{N}_i\,\mathbf{f}_i + \mathbf{Z}\,\mathbf{u}_i, \sigma_\epsilon^2\mathbf{I}_{n_i}; \mathbb{A}_i), \tag{14}$$

$$b_i \mid T_i = t_i \overset{ind.}{\sim} \text{N}(c\Delta + \Delta t_i, \psi), \tag{15}$$

$$T_i \overset{ind.}{\sim} \text{TN}(0, 1; (0, \infty)), \tag{16}$$

where $\text{TN}_{n_i}(.; \mathbb{A}_i)$ denotes the truncated multivariate normal distribution on the interval $\mathbb{A}_i = A_{i1} \times \ldots, \times A_{ini}$, with A_{ij} as the interval $(-\infty, \infty)$ if $C_{ij} = 0$ and $(-\infty, Q_{ij}]$ if $C_{ij} = 1$.

5.2 Prior distributions

To complete the Bayesian specification of the model proposed in (14) - (16), we need to assign prior distributions for σ_ϵ^2, ψ, and Δ. For the scale parameters σ_ϵ^2 and ψ, we consider a half-Cauchy prior, denoted by half-Cauchy(ω), where ω is the scale hyperparameter. This type of prior is recommended when a weakly informative prior is desired for the variance parameters in hierarchical models[46]. For the skewness parameter Δ, we consider a Student's-t prior, $t(0, 0.5, 2)$, as in Bandyopadhyay *et al.*[12]. In addition, we must specify prior distributions for the spline and wavelets components. These components are part of the vectors $\mathbf{f}_i$ and $\mathbf{u}_i$ respectively. We set $\alpha_k \sim \text{N}(0, \sigma_{\alpha_k}^2)$, $k = 0, 1$. Moreover, as mentioned earlier, we consider $u_s^{gbl} \mid \sigma_{gbl}^2 \overset{i.i.d.}{\sim} \text{N}(0, \sigma_{gbl}^2)$, $s = 1, \ldots, S$ with $\sigma_{gbl}^2 \sim$ half-Cauchy(ω_{gbl}). The prior distribution of u_{iw}^{sbj} for $i = 1, \ldots, n$; $w = 1, \ldots, W$ is given by Equation (10), where $\sigma_{sbj} \sim$ half-Cauchy(ω_{sbj}) and $\gamma_{iw} \mid \rho_{iw} \overset{ind.}{\sim}$ Bernoulli(ρ_{iw}) with $\rho_{iw} \sim$ Beta(α_ρ, β_ρ).

5.3 Model comparison tools

For model comparison, we use the conditional predictive ordinate[47] in order to obtain the log pseudo-marginal likelihood (LPML) statistic[48]. Larger values of LPML indicate a better fit. In addition, we also compute the expected Akaike information criterion (EAIC), the expected Bayesian (or Schwarz) information criterion (EBIC)[48], and the deviance information criterion (DIC)[49] measures. For all these criteria, the model producing the lowest value is to be preferred.

6 Application

We have applied our proposed methodology to data from the two trials, ACTG 315 and A5055, previously described in Section 2.4.

6.1 Censored longitudinal models

We compare three different censored longitudinal models: (a) the parametric SN-NLMEC model (described in Section 2.2), (b) our proposed model, i.e., the skew-normal semiparametric censored model (denoted for simplicity as SN-SPC model), and (c) a symmetric version of our proposed model in which the random effects are assumed to be normal distributed (denoted as N-SPC model). We consider a semiparametric model of the form

$$y_{ij} = \beta\, \text{CD4}^{+}_{ij} + f(t_{ij}) + g_i(t_{ij}) + \epsilon_{ij}, \tag{17}$$

where y_{ij} denotes the $\log_{10}$ transformation of the viral load for the i-th patient at time t_{ij} transformed to the unit interval ($i = 1, 2, \ldots, 46$; $j = 1, 2, \ldots, n_i$), $f(t_{ij})$ is a smooth function considered as a general mean, $g_i(t_{ij})$ is a patient-specific function of time, and ϵ_{ij} are random errors. This model relies on the fact that some investigations on HIV viral load suggested that the immunological response (CD4^{+}) is negatively correlated (in a linear way) with the virologic marker (viral load) during antiretroviral treatment[50]. For practical reasons, we have chosen $S = 15$ knots for the spline specification.

6.2 Prior distributions

In what regards prior distributions, we have considered

$$\beta \sim \text{N}(0, 10^6), \quad \sigma^2_{\epsilon} \sim \text{half-Cauchy}(25), \quad \psi \sim \text{half-Cauchy}(25) \quad \text{and} \quad \Delta \sim t(0, 0.5, 2).$$

In addition, for the splines and wavelets components, and following Wand and Ormerod[36], we consider

$$\alpha_k \sim N(0, 10^6), \quad k = 0, 1, \quad \sigma^2_{gbl} \sim \text{half-Cauchy}(25), \quad \sigma^2_{sbj} \sim \text{half-Cauchy}(25),$$

$$\gamma_{iw} \mid \rho_{iw} \sim \text{Bernoulli}(\rho_{iw}) \quad \text{and} \quad \rho_{iw} \sim \text{Beta}(1, 1).$$

It is important to emphasize that we have not assumed noninformative prior distributions for any of the parameters of interest. We have instead assumed weakly informative and proper prior distributions (such that the resulting posterior distributions are always proper). Moreover, and as suggested by a reviewer, we have conducted a sensitivity analysis for the wavelets and spline components, trying different choices of

prior parameters and by changing only one parameter at a time, keeping all the other parameters constant to their default values. We have also conducted a sensitivity analysis for the variance components prior distributions. The results, which are presented in the Supplementary Material (Section 1), suggest that the estimation of the regression coefficients, as well the skewness parameter, are robust under different hyperparameters specifications.

6.3 Convergence of the Markov chain Monte Carlo (MCMC) algorithm

Regarding the MCMC convergence, we have considered 3 Markov chains with 60000 iterations (25000 as burn-in and 1000 for the adaptive phase) and a lag of 6 observations for each chain to reach the convergence in the case of the parametric SN-NLMEC model. On the other hand, only 10000 iterations (5000 as burn-in and 1000 for the adaptive phase) and a lag of 5 observations were needed to reach convergence in the case of the semiparametric model. As can be seen, the MCMC computation of our proposed model requires a smaller number of samples than the parametric SN-NLMEC model.

6.4 Results

6.4.1 ACTG 315 clinical trial In Figure 5 we present the estimated trajectories, for 9 randomly chosen patients, produced by fitting our SN-SPC model. It is clear that, the semiparametric framework provides better subject-specific estimated trajectories than the parametric SN-NLMEC model, whose same estimated trajectories are displayed in Figure 2. The fitted global mean curves obtained using penalized splines and the SN-NLMEC model are displayed in Figure 7 (a).

In Table 1 it is shown the obtained values for the model comparison criteria described in Section 5.3. All these four criteria favor the SN-SPC model over the SN-NLMEC and N-SPC ones. Further, Table 2 shows the posterior mean, standard deviation (SD) and 95% highest posterior density (HPD) intervals for the parameter estimates of the semiparametric and SN-NLMEC models. In both cases the skewness parameter is positive. As expected, the associated coefficient to the $CD4^+$ cell count has a negative posterior mean in the case of the semiparametric model. Note that, although the estimation of this coefficient is positive for the nonlinear model, its effect is negative since the bi-phasic exponential decay model is given by $V(t) = P_1 e^{-\psi_1 t} + P_2 e^{-\psi_2 t}$.

6.4.2 A5055 clinical trial As in the previous trial, Figure 6 presents the estimated trajectories for 9 randomly chosen patients under the fitted SN-SPC model. Note that the semiparametric framework again provides better subject-specific estimated trajectories than the parametric SN-NLMEC model (see Figure 4). In Figure 7 (b) we display the fitted global mean curves using penalized splines and the SN-NLMEC model.

Using the model comparison tools described previously (LPML, EAIC, EBIC and DIC) and reported in Table 1, we conclude that the SN-SPC model outperforms the SN-NLMEC and N-SPC ones (as in the ACTG 315 clinical trial case). In addition, Table 2 shows the posterior mean, SD and 95% HPD intervals for the model parameters under the fitted semiparametric and SN-NLMEC models. Again, the skewness parameter is positive for all models considered, and the associated coefficient to the $CD4^+$ cell count has a negative posterior mean under the semiparametric model.

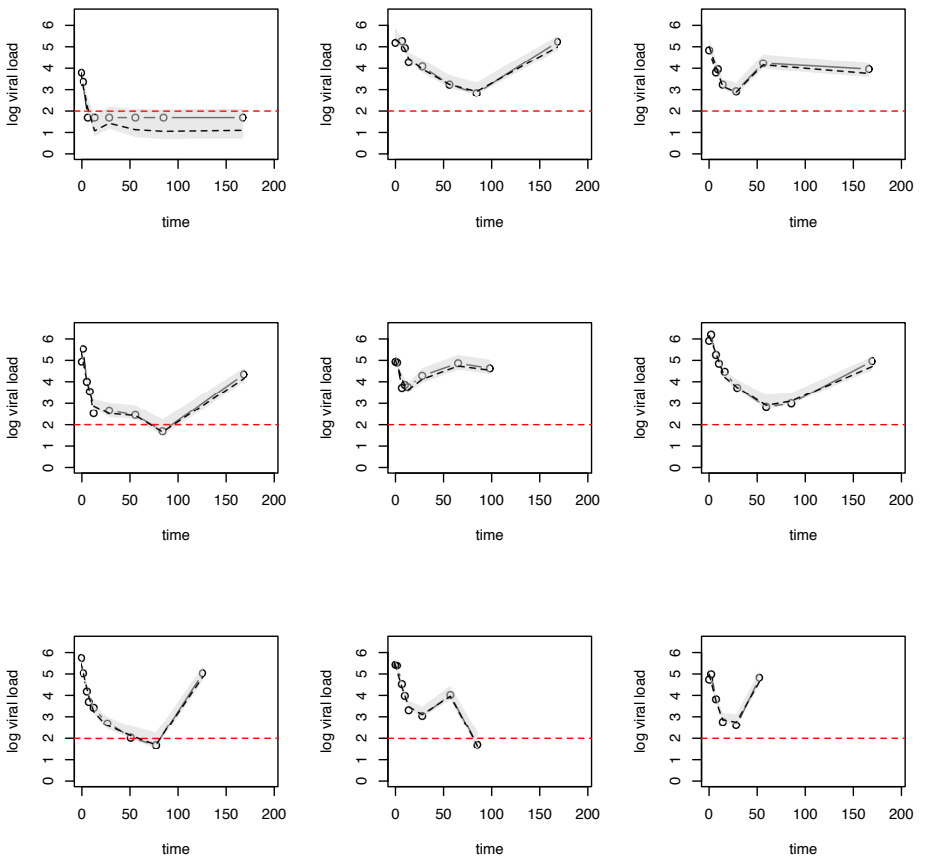


Figure 5. Application. ACTG 315 clinical trial. Viral loads in $\log_{10}$ scale (solid line) for 9 randomly chosen patients and estimated trajectories (dotted line) under the SN-SPC model. The horizontal red dotted line indicates the censoring level. The grey area represents a 95% credible band for the estimated trajectory.

7 Simulation study

In this section we perform two simulation experiments in order to show, on one hand, the ability of our proposed method for modeling skew data and, on the other hand, its capacity for recovering individual jagged trajectories.

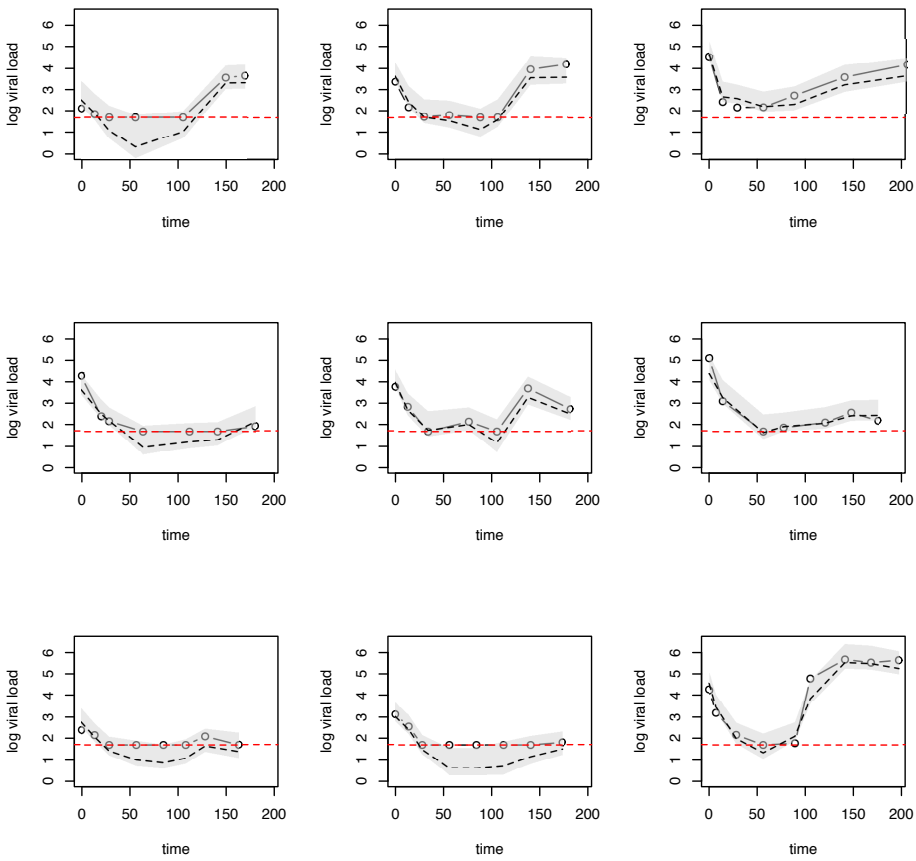


Figure 6. Application. A5055 clinical trial. Viral loads in $\log_{10}$ scale (solid line) for 9 randomly chosen patients and estimated trajectories (dotted line) under the SN-SPC model. The red dotted line indicates the censoring level. The grey area indicates a 95% credible band for the estimated trajectory.

7.1 Experiment 1: consequences of the misspecification of the skew-normal assumption for the random effects

In this first study, we consider the following nonlinear censored mixed model:

$$y_{ij} = \beta_1 \, x_{1ij} + f(t_{ij}) + b_i + \epsilon_{ij}, \quad i = 1, \ldots, 30, \quad j = 1, \ldots, n_i,$$

Table 1. Application. Model comparison criteria for the SN-NLMEC, N-SPC and SN-SPC models.

Dataset	Model	LPML	DIC	EAIC	EBIC
ACTG 315	SN-NLMEC	-353.13	1296.24	706.35	807.46
	N-SPC	-307.66	424.60	137.94	153.49
	SN-SPC	-296.38	263.39	57.49	73.05
A5055	SN-NLMEC	-184.66	666.04	390.66	469.62
	N-SPC	-140.94	451.58	199.40	211.54
	SN-SPC	-139.81	446.96	196.86	209.01

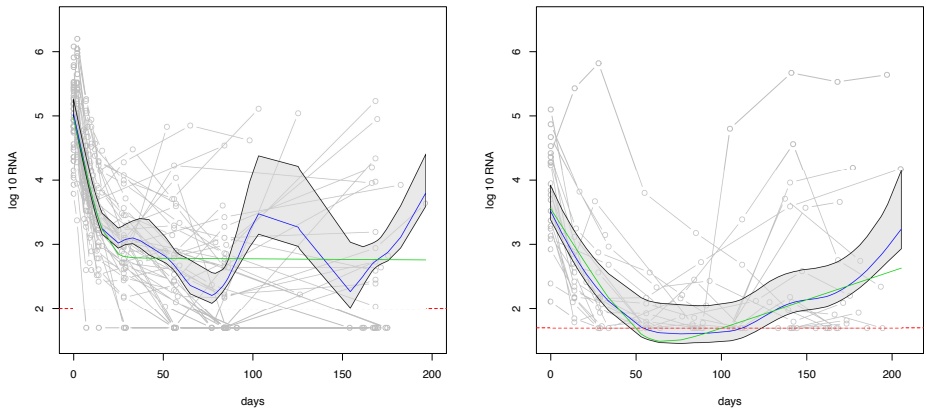
Table 2. Application. Posterior estimates under the SN-NLMEC, N-SPC and SN-SPC models.

		ACTG 315 clinical trial				A5055 clinical trial			
Model	Param.	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
SN-NLMEC	β_1	11.43	0.22	11.00	11.88	8.15	0.45	7.27	9.05
	β_2	54.23	3.52	47.14	60.96	19.78	3.45	13.67	27.08
	β_3	6.42	0.25	5.90	6.92	1.87	0.92	-0.09	3.53
	β_4	-0.01	0.84	-1.57	1.69	-4.18	1.47	-7.22	-1.43
	CD4$^+$	1.19	0.47	0.23	2.11	2.57	0.80	1.10	4.23
	σ^2_ϵ	0.16	0.01	0.13	0.20	0.46	0.07	0.33	0.63
	φ_{11}	3.55	1.21	1.67	6.30	8.30	3.92	2.87	19.19
	φ_{22}	2.97	1.42	0.78	5.95	8.52	4.28	2.80	20.21
	φ_{33}	3.23	1.11	1.48	5.79	8.33	3.95	2.85	19.25
	φ_{44}	20.47	6.36	10.83	35.49	16.65	5.70	8.09	30.43
	λ_1	5.96	3.81	1.25	16.40	19.69	14.39	5.89	61.88
	λ_2	8.08	5.08	1.61	22.13	19.48	14.26	5.68	59.32
	λ_3	7.06	4.88	1.14	20.27	19.68	14.73	6.10	64.20
	λ_4	3.06	2.15	0.44	8.89	12.76	10.44	3.35	43.13
N-SPC	CD4$^+$	-0.06	0.04	-0.14	0.03	-0.60	0.12	-0.86	-0.35
	σ^2_ϵ	0.06	0.009	0.05	0.08	0.24	0.07	0.12	0.39
	σ^2_b	0.35	0.09	0.18	0.52	0.34	0.14	0.12	0.62
SN-SPC	CD4$^+$	-0.05	0.03	-0.12	0.02	-0.59	0.13	-0.86	-0.34
	σ^2_ϵ	0.05	0.006	0.03	0.06	0.24	0.07	0.12	0.39
	σ^2_b	0.56	0.26	0.20	1.09	0.58	0.35	0.13	0.39
	λ	1.29	1.15	0.04	3.27	1.59	1.72	0.00	5.08

where

$$f(t) = 18[\sqrt{t(1-t)}\,\sin(1.6\,\pi/(t+0.2)) + 0.4\,I(t>0.13) - 0.7\,I(0.32<t<0.38) + 0.43\{(1-|(t-0.65)/0.03|)_+\}^4] + 0.42\{(1-|(x-0.91)/0.015|)_+)\}^4], \quad t\in(0,1)$$

where $|z-c|_+$ is equal to 0 for $z<c$, $b_i \overset{i.i.d.}{\sim} \text{SN}(c\Delta,\sigma_b^2,\lambda)$, and $\epsilon_{ij} \overset{i.i.d.}{\sim} \text{N}(0,\sigma_\epsilon^2)$, for $i=1,\ldots,30$ and $j=1,\ldots,n_i$. The true values of the parameters are $\beta_1=1.0$, $\sigma_\epsilon^2=1.0$, $\sigma_b^2=2.46$, $\lambda=6.34$,



(a) (b)

Figure 7. Fitted penalized spline-based global mean curve (blue line) and fitted global mean curve under SN-NLMEC model (green line). The red dotted line indicates the censoring level. The grey area represents a 95% credible band for the fitted mean curve under the SN-SPC model. (a) ACTG 315 clinical trial. (b) A5055 clinical trial.

and $\Delta = 1.2$. We consider 30 subjects with a number of observations per subject, n_i, generated from a truncated (below at 3) Poisson distribution with mean parameter equal to 5. This assures that the minimum number of observations per subject is 3, with an average of 5 observations per subject. The covariate x_{1ij} is sampled from a uniform distribution on the interval $(0, 1)$ and the values of t_{ij} are considered equally spaced over the interval $(0, 1)$. Figure 8 shows the simulated individual profiles for 9 randomly chosen datasets.

To study the effect of the level of censoring on the posterior estimates, we have considered different censoring proportions, namely, 0%, 10%, 20%, and 30%. For each censoring proportion, 100 simulated datasets are generated and two models are fitted: i) N-SPC considering a normal distribution for the random effect, i.e., $\mathbf{b}_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_b^2)$, and ii) our proposed SN-SPC that considers a skew-normal distribution for $\mathbf{b}_i$, $i = 1, \ldots, 30$. As in Bandyopadhyay *et al.* (2015)[12], we consider the normal model as a natural benchmark for comparing the behavior of the skew-normal model, with the former being a particular case of the latter.

With the aim of studying the consequences of misspecification of the skew-normal assumption for the random effects in parameter's estimation, we compute the Monte Carlo mean (MC-M), Monte Carlo standard deviation (MC-SD), and the root mean square error ($\sqrt{\mathrm{MSE}}$) for each parameter, over the 100 generated datasets, under each setting. Further, for each model we have considered (N-SPC and SN-SPC), we compute the LPML, EBIC, EAIC, and DIC over the 100 simulated datasets. In addition, we

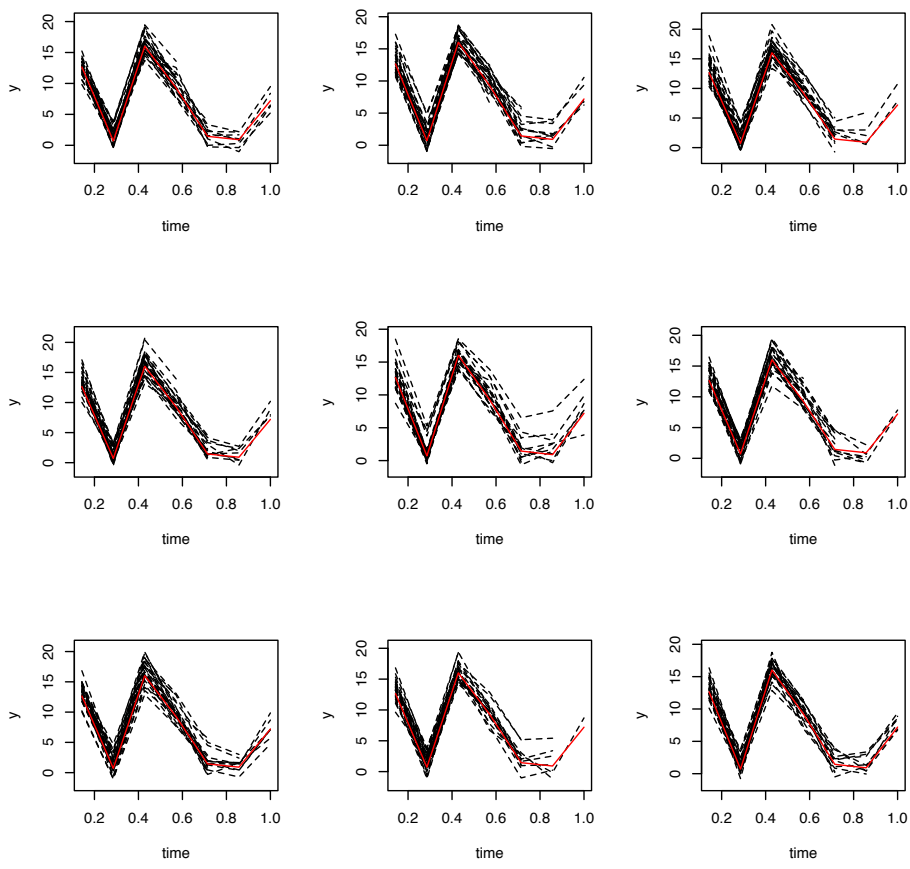


Figure 8. Experiment 1. Individual profiles for 9 randomly chosen simulated datasets. The red dotted line indicates the values of the function of $f(t)$, i.e., the global mean.

also compute the predicted mean square error, defined as $\text{PMSE} = \frac{1}{100}\sum_{k=1}^{100}\text{PMSE}^{(k)}$, with

$$\text{PMSE}^{(k)} = \frac{1}{30}\sum_{i=1}^{30}||\mathbf{y}_i^{(k)} - \widehat{\mathbf{y}}_i^{(k)}||^2,$$

where $\widehat{\mathbf{y}}_i^{(k)} = \widehat{E}^{(k)}(\mathbf{y}_i \mid \mathbf{t}_i, \mathbf{x}_i)$ is the ith predicted value in the kth sample and $\mathbf{y}_i^{(k)}$ is the vector of observed responses for subject ith in the kth sample. Results are reported in Table 3 and Table 4.

From Table 3, we observe that the SN model has the smallest MSE for the scale parameter σ_b^2, for all levels of censoring but, as expected, it increases with increasing censoring proportion. Focusing on

Table 3. **Experiment 1**. Monte Carlo simulation results, based on 100 simulated datasets, comparing the N-SPC and SN-SPC models for various levels of censoring.

			Posterior estimates of the parameters			
Level of Censoring	Model		$\beta_1 = 1.0$	$\sigma^2_\epsilon = 1.0$	$\sigma^2_b = 2.46$	$\lambda = 6.34$
0%	N-SPC	MC-M	1.031	0.860	0.929	-
		MC-SD	0.387	0.225	0.282	-
		$\sqrt{\text{MSE}}$	2.284	2.571	2.314	-
	SN-SPC	MC-M	1.024	0.849	2.377	6.134
		MC-SD	0.387	0.183	0.698	1.666
		$\sqrt{\text{MSE}}$	2.277	2.581	1.866	3.978
10%	N-SPC	MC-M	1.054	0.874	0.961	-
		MC-SD	0.395	0.250	0.296	-
		$\sqrt{\text{MSE}}$	2.280	2.558	2.289	-
	SN-SPC	MC-M	1.048	0.855	2.418	6.002
		MC-SD	0.408	0.210	0.726	1.723
		$\sqrt{\text{MSE}}$	2.272	2.568	1.861	4.017
20%	N-SPC	MC-M	1.049	0.852	0.991	-
		MC-SD	0.409	0.268	0.302	-
		$\sqrt{\text{MSE}}$	2.268	2.569	2.288	-
	SN-SPC	MC-M	1.050	0.837	2.469	6.111
		MC-SD	0.419	0.207	0.770	1.653
		$\sqrt{\text{MSE}}$	2.258	2.585	1.909	4.093
30%	N-SPC	MC-M	1.053	0.826	1.009	-
		MC-SD	0.427	0.250	0.317	-
		$\sqrt{\text{MSE}}$	2.245	2.587	2.265	-
	SN-SPC	MC-M	1.067	0.814	2.516	6.246
		MC-SD	0.435	0.214	0.767	1.676
		$\sqrt{\text{MSE}}$	2.233	2.600	1.930	4.173

the MC-M, we can see that the normal model tends to underestimate uniformly this scale parameter for all levels of censoring. It is important to remark that this situation could produce misleading conclusions about the sources of variability presented in the data, particularly, the between-subject variance component. On the other hand, the regression parameter posterior estimates are very similar under both models for all levels of censoring, indicating that the estimation of this parameter is very robust under a misspecification of the random effect distribution. Finally, we can conclude that, in general, the SN model provides parameter's estimates closer to the true values than the normal one, even when the censoring proportion is high. Table 4 presents the arithmetic averages across the 100 simulated datasets of the various model comparison measures mentioned earlier. As it can be noticed, all these criteria favored the SN model, for all censoring proportions considered in the study. Therefore, it can be concluded that,

Table 4. **Experiment 1**. Arithmetic averages, over the 100 simulated datasets, of the PMSE, LPML, DIC, EAIC, and EBIC for the N-SPC and SN-SPC models.

Level of Censoring	Criteria	N-SPC	SN-SPC
0%	PMSE	0.585	0.577
	LPML	−225.699	−224.341
	DIC	861.084	855.383
	EAIC	423.877	422.650
	EBIC	434.553	432.799
10%	PMSE	0.574	0.566
	LPML	−228.350	−227.304
	DIC	866.429	860.759
	EAIC	425.832	424.860
	EBIC	436.752	434.752
20%	PMSE	0.583	0.571
	LPML	−224.907	−224.006
	DIC	844.883	839.687
	EAIC	413.594	412.881
	EBIC	424.772	422.512
30%	PMSE	0.757	0.751
	LPML	−217.959	−217.257
	DIC	803.611	798.452
	EAIC	390.998	389.277
	EBIC	401.169	399.909

as it is to be expected, the SN model provides a better fit compared to the N model when the dataset presents an obvious departure from normality.

7.2 Experiment 2: flexibility of the semiparametric approach

The aim of this second simulation study is to assess the ability of the wavelet approach for handling different individual trajectories. In order to do that, we consider the following scenarios, all of them using the SN distribution, namely: (i) penalized wavelets and splines (SN-SPC model), (ii) only penalized splines for the global mean function (SN-Spline model), and (iii) a nonlinear model with conditional mean function given by $E(y_{ij} \mid b_i) = b_i + \beta_1\, x_{1ij} + \sin(t_{ij})$ (SN-Parametric model). Note that, the election of the $\sin$ function for the parametric model is because this function is involved in the function $f(t)$, used for simulating the data. In this experiment, we consider the same non linear model used in the previous experiment. We also compute the same model comparison criteria described in Section 5.3 and the PMSE described in Section 7.1.

Regarding parameter's estimation, Table 5 reports the posterior estimates of the model parameters under the scenarios considered in this setting. It is worth mentioning that our proposed semiparametric model, which considers both wavelets and splines at the same time (SN-SPC model), generates, in general, smaller values of $\sqrt{\text{MSE}}$. We also note that the model considering only penalized splines for the global mean function (SN-Spline model) provides, as expected, also better results when compared to the parametric one (SN-Parametric). Clearly, our proposed semiparametric approach, when compared to these other two alternatives, provides posterior estimates closer to the true values, particularly for the skewness parameter (λ) and variance components (σ^2_ϵ and σ^2_b).

Finally, the results showed in Table 6 highlight the superiority of the SN-SPC model over the other two models considered in this experiment. It is important to remark that the SN-SPC model attains the lowest PMSE values, thus demonstrating that this model is more accurate in terms of prediction of the observed individual trajectories.

8 Conclusions

In this article we have proposed a Bayesian flexible semiparametric approach to model censored longitudinal data. Under this new approach, splines are used to approximate the general mean and wavelets for modeling the individual trajectories per subject. By letting the random error of the model to be normally distributed and assuming that the random effects follow a skew normal distribution, the resulting marginal distribution of the responses follows a skew normal distribution and, therefore, we got rid of the standard assumption of normally distributed data. The newly method was applied to an HIV viral load dataset, nicely illustrating how the proposed model can produce, when compared to some of the existing alternatives, more accurate subject-specific estimated trajectories. Two simulation experiments that validate the performance of our method were conducted. Our method can be fitted using standard available software packages, e.g. `R` and `JAGS` and the computation time is viable, making our approach appealing to clinical practitioners. Sample code is provided in Section 2 of the supplementary material.

It is important to stress that although the proposed approach is able to handle the skewness commonly observed in follow-up studies of viral loads, but it cannot deal properly with extremely small and/or large viral load levels, thus making the use of heavy-tailed models possibly more suitable for such cases. Of course, a fully nonparametric alternative based, for instance, on Dirichlet processes seems to be a promising avenue for future research.

Table 5. Experiment 2. Monte Carlo simulation results based on 100 simulated datasets comparing the SN-SPC, SN-Spline, and SN-NLMEC models for various levels of censoring.

			Posterior estimates			
Level of Censoring	Model		$\beta_1 = 1.0$	$\sigma^2_\epsilon = 1.0$	$\sigma^2_b = 2.46$	$\lambda = 6.34$
0%	SN-SPC	MC M	1.024	0.849	2.377	6.134
		MC SD	0.387	0.183	0.698	1.666
		$\sqrt{\text{MSE}}$	2.277	2.581	1.866	3.978
	SN-Spline	MC M	1.032	0.994	2.332	6.147
		MC SD	0.385	0.126	0.695	1.639
		$\sqrt{\text{MSE}}$	2.697	2.809	2.229	4.450
	SN-Parametric	MC M	1.001	40.175	1.188	4.035
		MC SD	2.044	1.340	0.086	0.326
		$\sqrt{\text{MSE}}$	3.639	37.478	2.599	2.660
10%	SN-SPC	MC M	1.048	0.855	2.418	6.002
		MC SD	0.408	0.210	0.726	1.723
		$\sqrt{\text{MSE}}$	2.272	2.568	1.861	4.017
	SN-Spline	MC M	1.057	1.027	2.354	5.739
		MC SD	0.408	0.144	0.721	1.570
		$\sqrt{\text{MSE}}$	2.695	2.790	2.224	4.128
	SN-Parametric	MC M	1.231	45.911	1.274	4.067
		MC SD	2.207	1.707	0.084	0.298
		$\sqrt{\text{MSE}}$	3.661	43.201	2.549	2.866
20%	SN-SPC	MC M	1.050	0.837	2.469	6.111
		MC SD	0.419	0.207	0.770	1.653
		$\sqrt{\text{MSE}}$	2.258	2.585	1.909	4.093
	SN-Spline	MC M	1.051	1.030	2.393	6.110
		MC SD	0.418	0.165	0.765	1.742
		$\sqrt{\text{MSE}}$	2.690	2.783	2.266	4.477
	SN-Parametric	MC M	1.238	53.154	1.380	4.223
		MC SD	2.499	2.177	0.098	0.315
		$\sqrt{\text{MSE}}$	3.557	50.647	2.588	2.689
30%	SN-SPC	MC M	1.067	0.814	2.516	6.246
		MC SD	0.435	0.214	0.767	1.676
		$\sqrt{\text{MSE}}$	2.233	2.600	1.930	4.173
	SN-Spline	MC M	1.063	1.020	2.433	6.001
		MC SD	0.428	0.164	0.773	1.589
		$\sqrt{\text{MSE}}$	2.689	2.786	2.253	5.409
	SN-Parametric	MC M	1.086	60.717	1.414	4.297
		MC SD	2.541	2.814	0.072	0.331
		$\sqrt{\text{MSE}}$	3.941	58.039	2.490	2.766

Table 6. Experiment 2. Arithmetic averages, over the 100 simulated datasets, of the PMSE, LPML, DIC, EAIC, and EBIC for the three different models considered.

Level of Censoring	Criteria	SN-SPC	SN-Spline	SN-Parametric
0%	PMSE	0.577	0.854	37.427
	LPML	−224.341	−230.341	−472.461
	DIC	855.383	904.642	1886.173
	EAIC	422.650	455.378	951.616
	EBIC	432.799	467.282	966.498
10%	PMSE	0.566	0.862	37.503
	LPML	−227.304	−233.359	−455.193
	DIC	860.759	915.227	1816.849
	EAIC	424.860	460.643	916.931
	EBIC	434.752	472.539	931.813
20%	PMSE	0.571	0.871	37.773
	LPML	−224.006	−229.533	−432.766
	DIC	839.687	897.622	1726.869
	EAIC	412.881	451.739	871.916
	EBIC	422.512	463.632	886.799
30%	PMSE	0.751	0.944	38.246
	LPML	−217.257	−217.8497	−407.215
	DIC	798.452	846.061	1624.310
	EAIC	389.277	425.200	820.584
	EBIC	399.909	437.105	835.467

Acknowledgements

We thank the Editor, Associate Editor, and two anonymous referees whose constructive comments and suggestions led to an improved presentation of the paper. L. M. Castro acknowledges support from Grant FONDECYT 1170258 from the Chilean government, Programa Nacional de Innovación para la Competitividad y Productividad (Innóvate Perú) under the contract 452-PNICP-ECIP-2014 and the Department of Science of Pontificia Universidad Católica del Perú. The research of W.L. Wang was partially supported by the Ministry of Science and Technology of Taiwan under Grant no. MOST 105-2118-M-035-004-MY2. The research of V. H. Lachos was partially supported by CNPq-Brazil (Grant 305054/2011-2) and FAPESP-Brazil (Grant 2014/02938-9). V. Inácio acknowledges support from FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013. C. L. Bayes acknowledges support from Dirección de Gestión de la Investigación at PUCP through grants DGI-2014-0017/0070 and DGI-2014-0077/0065.

References

1. Ndembi N, Goodall R, Dunn D, McCormick A, Burke A, Lyagoba F, et al. Viral Rebound and Emergence of Drug Resistance in the Absence of Viral Load Testing: A Randomized Comparison between Zidovudine-Lamivudine plus Nevirapine and Zidovudine-Lamivudine plus Abacavir. Journal of Infectious Diseases. 2010;201(1):106–113.

2. Zackin R, Marschner I, Andersen J, Cowles MK, Gruttola VD, Hammer S, et al. Perspective: human immunodeficiency virus type 1 (HIV-1) RNA end points in HIV clinical trials: Issues in interim monitoring and early stopping. The Journal of Infectious Diseases. 1998;177:761–765.
3. Wu L. Mixed Effects Models for Complex Data. Boca Raton, FL: Chapman & Hall/CRC; 2010.
4. Hughes JP. Mixed effects models with censored data with application to HIV RNA levels. Biometrics. 1999;55(2):625–629.
5. Jacqmin-Gadda H, Thièbaut R, Chêne G. Analysis of left-censored longitudinal data with application to viral load in HIV infection. Biometrics. 2000;1(4):355–368.
6. Vaida F, Fitzgerald AP, DeGruttola V. Efficient hybrid EM for linear and nonlinear mixed effects models with censored response. Computational Statistics & Data Analysis. 2007;51(12):5718–5730.
7. Vaida F, Liu L. Fast Implementation for Normal Mixed Effects Models With Censored Response. Journal of Computational and Graphical Statistics. 2009;18(4):797–817.
8. Matos LA, Prates MO, Chen MH, Lachos VL. Likelihood Based Inference for Linear and Nonlinear Mixed-Effects Models with Censored Response Using the Multivariate-t Distribution. Statistica Sinica. 2013;23:1299–1322.
9. Lachos VH, Bandyopadhyay D, D DK. Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. Biometrics. 2011;67:1594–1604.
10. Lange KL, Sinsheimer JS. Normal/independent distributions and their applications in robust regression. Journal of Computational and Graphical Statistics. 1993;2:175–198.
11. Banyopadhyay D, Lachos V, Castro LM, Dey D. Skew-normal/independent linear mixed models for censored responses with applications to HIV viral loads. Biometrical Journal. 2012;54(3):405–425.
12. Bandyopadhyay D, Castro LM, Lachos VH, Pinheiro H. Robust joint non-linear mixed-effects models and diagnostics for censored HIV viral loads with CD4 measurement error. Journal of Agricultural, Biological and Environmental Statistics. 2015;20(1):121–139.
13. Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression. vol. 12. Cambridge University Press; 2003.
14. Ibacache-Pulgar G, Paula G, Cysneiros F. Semiparametric additive models under symmetric distributions. Test. 2013;22:103–121.
15. Castro LM, Lachos VH, Galvis D, Bandyopadhyay D. Bayesian semiparametric longitudinal data modeling using normal/independent densities. In: Upadhyay S, Singh U, Dey D, Loganathan A, editors. Current Trends in Bayesian Methodology with Applications. Chapman & Hall/CRC; 2015. p. 157–176.
16. Castro LM, Lachos VH, Arellano-Valle RB. Partially linear censored regression models using heavy-tailed distributions: a Bayesian approach. Statistical Methodology. 2014;18:14–31.
17. Azzalini A. The skew-normal distribution and related multivariate families. Scandinavian Journal of Statistics. 2005;32:159–188.
18. Lachos VH, Ghosh P, Arellano-Valle RB. Likelihood based inference for skew-normal independent linear mixed models. Statistica Sinica. 2010;20:303–322.
19. Wu H, Ding A. Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials. Biometrics. 1999;55:410–418.
20. Wu L. A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. Journal of the American Statistical Association. 2002;97:955–964.
21. Lachos VH, Castro LM, Dey D. Bayesian inference in nonlinear mixed-effects models using normal independent distributions. Computational Statistics and Data Analysis. 2013;64:237–252.

22. Lin TI, Wang WL. Multivariate skew-normal linear mixed models for multi-outcome longitudinal data. Statistical Modelling. 2013;13(3):199–221.
23. Wang WL. Approximate methods for maximum likelihood estimation of multivariate nonlinear mixed-effects models. Entropy. 2015;17:5353–5381.
24. Matos L, Castro LM, Lachos VH. Censored mixed-effects models for irregularly observerd repeated measures with applications to HIV viral loads. Test. 2016;25(4):627–653.
25. Lederman M, Connick E, Landay A, Kuritzkes D, Spritzler J, Clair M, et al. Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and ritonavir: results of AIDS clinical trials group protocol 315. The Journal of Infectious Diseases. 1998;178:70–79.
26. Connick E, Lederman M, Kotzin B, Spritzler J, Kuritzkes D, Clair M, et al. Immune reconstitution in the first year of potent antiretroviral therapy and its relationship to virologic response. The Journal of Infectious Disease. 2000;181:358–363.
27. Christensen R, Johnson W, Branscum A, Hanson T. Bayesian Ideas and Data Analysis. Chapman & Hall/CRC; 2011.
28. Wang WL, Lin TI, Lachos VH. Extending multivariate-t linear mixed models for multiple longitudinal data with censored responses and heavy tails. Statistical Methods in Medical Research. 2015;p. DOI:10.1177/0962280215620229.
29. Lin TI, Wang WL. Multivariate-t nonlinear mixed models with application to censored multi-outcome AIDS studies. Biostatistics. 2017;p. DOI:10.1093/biostatistics/kxx013.
30. Lachos VH, Matos L, Castro LM, Chen MH. Heavy-tailed longitudinal linear mixed models for multiple censored responses data. Universidade Estadual de Campinas; 2017.
31. Acosta E, Wu H, Walawander A, Eron J, Pettinelli C, Yu S, et al. Comparison of two indinavir/ritonavir regimens in treatment-experienced HIV-infected individuals. Journal of Acquired Immune Deficiency Syndromes. 2004;37:1358–1366.
32. Morris J, Carroll R. Wavelet-based functional mixed models. Journal of the Royal Statistical Society, Series B. 2006;68(2):179–199.
33. Silverman B. Wavelets in statistics: beyond the standard assumptions. Phiolosophical Transactions of the Royal Society of London, A. 1999;357:2459–2473.
34. Abramovich F, Bailey T, Sapatinas T. Wavelet analysis and its statistical applications. The Statistician. 2000;49(1):1–29.
35. Ko K, Qu L, Vannucci M. Wavelet-based Bayesian estimation of partially linear regression models with long memory errors. Statistica Sinica. 2009;19:1463–1478.
36. Wand M, Ormerod J. Penalized waveletes: embedding waveletes into semiparametric regression. Electronic Journal of Statistics. 2011;5:1654–1717.
37. Hongchang H, Yuan F. Asymptotic properties of wavelet estimators in a semiparametric regression model with censored data. Journal of Natural Sciences. 2012;17(4):290–296.
38. Daubechies I. Orthonormal bases of compactly supported waveletes. Communications on Pure and Applied Mathematics. 1988;41:909–996.
39. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2017. Available from: https://www.R-project.org/.
40. Nason G. `wavethresh 4.5`. Waveletes statistics and transforms; 2010.
41. Aykroyd R, Mardia K. A wavelet approach to shape analysis for splinal curves. Journal of Applied Statistics. 2003;30:605–623.

42. Morris J, Vanucci M, Brown P, Carroll R. Wavelet-based nonparametric modeling of hierarchical functions in colon carinogenesis. Journal of the American Statistical Association. 2003;98:573–597.
43. Zhao W, Wu R. Wavelet-based nonparametric functional mapping of longitudinal curves. Journal of the American Statistical Association. 2008;103:714–725.
44. Dúrban M, Harezlak J, Wand M, Carroll R. Simple fitting of subject-specific curves for longitudinal data. Statistics in Medicine. 2005;24:1153–1167.
45. Wand M, Ormerod J. On semiparametric regression with O'Sullivan penalised splines. Australian & New Zealand Journal of Statistics. 2008;50(179-198).
46. Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis. 2006;1(3):515–533.
47. Geisser S, Eddy W. A predictive approach to model selection. Journal of the American Statistical Association. 1979;74:153–160.
48. Carlin BP, Louis TA. Bayesian Methods for Data Analysis. Chapman & Hall/CRC, New York; 2008.
49. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. 2002;64:583–639.
50. Jiang R, Yang X, Qian W. Random weighting M-estimation for linear errors-in-variables models. Journal of the Korean Statistical Society. 2012;41(4):505–514.