UNIVERSIDADE
ESTADUAL DE
CAMPINAS

UNICAMP

IMECC

ICSA Applied Statistics Symposium 2022

Extending multivariate Student's-t semiparametric mixed models for longitudinal data with censored responses and heavy tails

Larissa Avila Matos

Co-Authors: Thalita B. Mattos, Luis M. Castro and Victor H. Lachos

FAPESP

June 21, 2022

Contents

Introduction

Motivation

The semiparametric mixed effects model with censored responses

Simulation study

Application

Conclusions

Introduction

Linear and nonlinear mixed-effects (LME/NLME) models have been extensively studied in the literature and applied to analyze longitudinal data.

The classical LME model is often written in the following form:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i,$$

where $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$, $\quad i = 1, \ldots, n$, with $\mathbf{b}_i \perp \boldsymbol{\epsilon}_i$.

One difficulty that arises in longitudinal data analysis is when the response is censored for some of the observations.

- For example: HIV studies, where the detection of the viral load in the blood compartment is often limited by the sensitivity of a laboratory assay.

Introduction

Several statistical approaches have been developed to deal with longitudinal data with censored measurements in the LME framework:

- Hughes (1999): Monte Carlo EM (MCEM) for LME with censored responses (LMEC).
- Vaida and Liu (2009): EM algorithm for LME/NLME models with censored responses, which uses closed-form expressions at the E-step (LMEC/NLMEC).
- Matos et al. (2013): EM algorithm for LMEC/NLMEC based on the multivariate Student-t distribution, named t-LMEC/t-NLMEC.
- Lachos et al. (2019): a robust multivariate linear mixed model for multiple censored responses based on the class of SMN distributions.

Introduction

Semiparametric models:

- Zeger and Diggle (1994) proposed a semiparametric model where a nonparametric function is used to model the time effect, and a random intercept together with a Gaussian stochastic process is used to account for the within-subject correlation.
- Vock et al. (2011) developed a mixed model framework for censored longitudinal data in which the random effects are represented by the flexible seminonparametric (SNP) density.

Goal: The aim of this work is to perform a study of statistical inference in the semiparametric mixed effects models for longitudinal irregularly observed censored data (SMEC). Extend the work of Mattos et al. (2021).

Motivating example - A5055 study

The dataset:

- 44 infected patients with the human immunodeficiency virus type 1 (HIV-1).
- These patients were treated with one of two potent ARV therapies.
- The viral load ($\log_{10}$(RNA)) was quantified irregulary on days 0, 7, 14, 28, 56, 84, 112, 140, and 168 of follow-up.
- CD4 and CD8, two immunologic markers frequently used to monitor disease progression in AIDS studies, were also measured along with the viral load.

  33.5% (106 out of 316) of measurements lies below the limits (50 copies/mL) of assay quantification (left-censored).
- A more detailed description of this study and data can be found in Acosta et al. (2004)

A5055 study

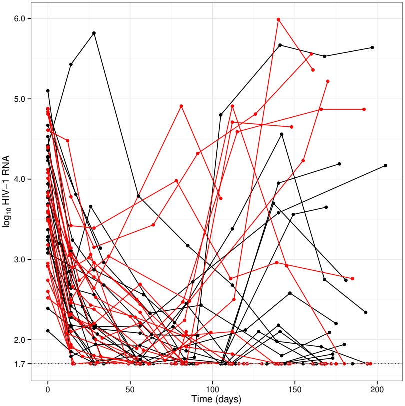

Figure: **A5055 study**. Individual profiles for HIV viral load (in log10 scale) at different follow-up times. Black lines indicate patients under treatment 1 and red lines indicate patients under treatment 2.

A5055 study

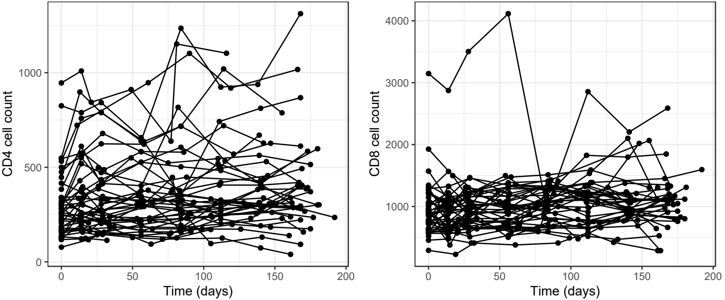

Figure: **A5055 study**. Individual profiles for CD4+ and CD8+ cell count at different follow-up times.

The model

The semiparametric mixed-effects model is specified as follows:

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

- $\mathbf{b}_i \overset{\text{iid.}}{\sim} t_q(\mathbf{0}, \mathbf{D}, \nu)$ and $\boldsymbol{\epsilon}_i \overset{\text{ind.}}{\sim} t_{n_i}(\mathbf{0}, \boldsymbol{\Omega}_i, \nu), i = 1, \ldots, n$. Note that, $\boldsymbol{\epsilon}_i$ and $\mathbf{b}_i$ are uncorrelated but not necessarily independent;

- $\mathbf{f} = (f(t_1^0), \ldots, f(t_r^0))^\top$ is an $r \times 1$ vector with $t_1^0, \ldots, t_r^0$ being the distinct and ordered values of t_{ij}, with $f(\cdot)$ a smooth function of time t_{ij};

- $\mathbf{N}_i$ is an $(n_i \times r)$ incidence matrix whose (j, s)-th element equals the indicator function $\mathbb{I}(t_{ij} = t_s^0)$ for $j = 1, \ldots, n_i$ and $s = 1, \ldots, r$;

- $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$ models between-subjects variability;

- $\boldsymbol{\Omega}_i = \sigma^2\mathbf{E}_i$ is the correlation structure of the error vector, where the $n_i \times n_i$ matrix $\mathbf{E}_i$ incorporates a time-dependence structure.

Correlation structures

DEC - Munoz et al. (1992)

Damped exponential correlation (DEC):

$$\mathbf{E}_i = \mathbf{E}_i(\phi, \mathbf{t}_i) = \left[\phi_1^{|t_{ij} - t_{ik}|^{\phi_2}} \right], \ i = 1, \ldots, n, \ \ j, k = 1, \ldots, n_i, \qquad (2)$$

For the DEC structure, we have that:

(a) if $\phi_2 = 0$, then $\mathbf{E}_i$ generates the compound symmetry correlation structure;

(b) when $0 < \phi_2 < 1$, then $\mathbf{E}_i$ presents a decay rate between the compound symmetry structure and the first-order AR (AR (1)) model;

(c) if $\phi_2 = 1$, then $\mathbf{E}_i$ generates an AR(1) structure;

(d) when $\phi_2 > 1$, $\mathbf{E}_i$ presents a decay rate faster than the AR(1) structure; and

(e) if $\phi_2 \to \infty$, then $\mathbf{E}_i$ represents the first-order moving average model, MA(1).

The model

We assume that the response y_{ij} is not fully observed for all i, j.

Let the observed data for the i-th subject be $(\mathbf{V}_i, \mathbf{C}_i)$, where

- $\mathbf{V}_i$ represents the vector of uncensored readings or censoring level,
- $\mathbf{C}_i$ is the vector of left-censoring indicators,

such that

$$\begin{aligned} y_{ij} &\leq V_{ij} \quad \text{if} \quad C_{ij} = 1, \\ y_{ij} &= V_{ij} \quad \text{if} \quad C_{ij} = 0. \end{aligned} \tag{3}$$

The model defined in (1)-(3) is henceforth called the DEC-t-SMEC model.

The log-likelihood function

Following Vaida and Liu (2009), classical inference on the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \mathbf{f}^\top, \sigma^2, \boldsymbol{\alpha}^\top, \boldsymbol{\phi}^\top)^\top$ is based on the marginal distribution of $\mathbf{y}_i$.

For complete data, we have marginally that $\mathbf{y}_i \overset{\text{ind.}}{\sim} t_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu)$, where

$$\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{N}_i\mathbf{f} \quad \text{and} \quad \boldsymbol{\Sigma}_i = \boldsymbol{\Omega}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top.$$

Let $\mathbf{y}_i^o$ be the n_i^o-vector of observed outcomes and $\mathbf{y}_i^c$ be 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 $\mathbf{y}_i^c$.

The likelihood function for subject i (using conditional probability arguments) is given by:

$$\begin{aligned} L_i(\boldsymbol{\theta}) = f(\mathbf{y}_i|\boldsymbol{\theta}) &= f(\mathbf{V}_i|\mathbf{C}_i, \boldsymbol{\theta}) = f(\mathbf{y}_i^o|\boldsymbol{\theta}) f(\mathbf{y}_i^c \leq \mathbf{V}_i^c|\mathbf{V}_i^o, \boldsymbol{\theta}) \\ &= t_{n_i^o}(\mathbf{y}_i^o; \boldsymbol{\mu}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo}) \mathrm{T}_{n_i^c}(\mathbf{V}_i^c; \boldsymbol{\mu}_{ico}, \mathbf{S}_i) = L_i. \qquad (4) \end{aligned}$$

The log-likelihood function for the observed data is thus given by $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^n \{\log L_i\}$.

The log-likelihood function

However, maximization of $\ell(\boldsymbol{\theta})$ without imposing restrictions on the function $\mathbf{f}(\cdot)$ may cause over-fitting and non-identification of $\boldsymbol{\beta}$ (Green, 1987).

A well-known procedure that is based on the idea of log-likelihood penalization consists of incorporating a penalty function in the log-likelihood, such that:

$$\ell_p(\boldsymbol{\theta}, \lambda) \quad = \quad \ell(\boldsymbol{\theta}|\mathbf{y}) - \frac{\lambda}{2} J(\mathbf{f}), \tag{5}$$

where $J(\mathbf{f})$ denotes the penalty function over $\mathbf{f}(\cdot)$, and λ is a smoothing parameter that controls the tradeoff between goodness of fit and the smoothness estimated function.

We consider the following penalty function:

$$J(\mathbf{f}) = \int_a^b [f''(t)]^2 dt = \mathbf{f}^\top \mathbf{K} \mathbf{f},$$

where $[f''(t)]$ denotes the second derivative of $f(t)$ with $[a, b]$ containing the values t_j^0, of $j = 1, \ldots, r$. By maximizing (5), one obtains the MPL estimates.

Inference

The complete-data log-likelihood function

The model can be expressed in the following hierarchical model:

$$\mathbf{y}_i|\mathbf{b}_i, u_i \overset{\text{ind.}}{\sim} \mathrm{N}_{n_i}(\boldsymbol{\mu}_i, u_i^{-1}\boldsymbol{\Omega}_i), \quad \mathbf{b}_i|u_i \overset{\text{ind.}}{\sim} \mathrm{N}_q(\mathbf{0}, u_i^{-1}\mathbf{D}), \quad u_i \overset{\text{ind.}}{\sim} \text{Gamma}(\nu/2, \nu/2).$$

Assuming that $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)$, $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_n^\top)$, and $\mathbf{u} = (u_1, \ldots, u_n)^\top$ are hypothetical missing variables. And, augmenting with the observed variables $(\mathbf{V}, \mathbf{C})$ where $\mathbf{V} = vec(\mathbf{V}_1, \ldots, \mathbf{V}_n)$, and $\mathbf{C} = vec(\mathbf{C}_1, \ldots, \mathbf{C}_n)$.

So, the penalized log-likelihood function for the model based on complete data $\mathbf{y}_c = (\mathbf{C}^\top, \mathbf{V}^\top, \mathbf{y}^\top, \mathbf{b}^\top, \mathbf{u}^\top)^\top$ is given by

$$\ell_{pc}(\boldsymbol{\theta}|\mathbf{y}_c) = \ell_c(\boldsymbol{\theta}|\mathbf{y}_c) - \frac{\lambda}{2}\mathbf{f}^\top\mathbf{K}\mathbf{f}, \tag{6}$$

with

$$\begin{aligned}\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) &= \sum_{i=1}^{n}\Bigg[-\frac{n_i}{2}\log\sigma^2 - \frac{1}{2}\log(|\mathbf{E}_i|) - \frac{u_i}{2\sigma^2}(\mathbf{y}_i - \boldsymbol{\mu}_i - \mathbf{Z}_i\mathbf{b}_i)^\top\mathbf{E}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i - \mathbf{Z}_i\mathbf{b}_i)\\ &- \frac{1}{2}\log|\mathbf{D}| - \frac{u_i}{2}\mathbf{b}_i^\top\mathbf{D}^{-1}\mathbf{b}_i + \log h(u_i|\nu) + C\Bigg].\end{aligned}$$

The EM algorithm

E-Step: Calculate the conditional expectation. Given the complete-data log-likelihood function, the Q-function can be written as:

$$\begin{aligned} Q_p(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= \mathbb{E}\left[\ell_{\mathrm{c}}(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{V},\mathbf{C},\widehat{\boldsymbol{\theta}}^{(k)}\right] - \frac{\lambda}{2}\mathbf{f}^\top\mathbf{K}\mathbf{f} \\ &= \sum_{i=1}^{n} Q_{1i}(\boldsymbol{\beta},\mathbf{f},\sigma^2,\phi|\widehat{\boldsymbol{\theta}}^{(k)}) + \sum_{i=1}^{n} Q_{2i}(\boldsymbol{\alpha}|\widehat{\boldsymbol{\theta}}^{(k)}), \end{aligned}$$

where

$$\begin{aligned} Q_{1i}(\boldsymbol{\beta},\mathbf{f},\sigma^2|\widehat{\boldsymbol{\theta}}^{(k)}) &= -\frac{1}{2\sigma^2}\left[\widehat{a}_i^{(k)} - 2\boldsymbol{\mu}_i^\top\mathbf{E}_i^{-1}\left(\widehat{u_i\mathbf{y}}_i^{(k)} - \mathbf{Z}_i\widehat{u_i\mathbf{b}}_i^{(k)}\right) + \widehat{u}_i^{(k)}\boldsymbol{\mu}_i^\top\mathbf{E}_i^{-1}\boldsymbol{\mu}_i\right] \\ &\quad -\frac{n_i}{2}\log\sigma^2 - \frac{1}{2}\log(|\mathbf{E}_i|) - \frac{\lambda}{2n}\mathbf{f}^\top\mathbf{K}\mathbf{f}, \quad \text{and} \end{aligned}$$

$$Q_{2i}(\boldsymbol{\alpha}|\widehat{\boldsymbol{\theta}}^{(k)}) = -\frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\mathrm{tr}\left(\widehat{u_i\mathbf{b}_i\mathbf{b}_i^\top}^{(k)}\mathbf{D}^{-1}\right).$$

CM-Step: Update $\widehat{\boldsymbol{\theta}}^{(k)}$ by the maximization of $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$, which leads to the closed expressions for $\widehat{\boldsymbol{\beta}}$, $\mathbf{f}$ $\widehat{\sigma^2}$ and $\widehat{\mathbf{D}}$.

Approximate standard errors

Following Segal et al. (1994) and Louis (1982), we derive the covariance matrix of $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{f}})$ by using the inverse of the penalized observed information matrix.

Thus, the approximate covariance matrix of $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{f}})$ is given as:

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{f}}) \approx \mathcal{I}_p^{-1}(\boldsymbol{\beta}, \mathbf{f})\Big|_{\widehat{\boldsymbol{\theta}}},$$

where the penalized expected information matrix $\mathcal{I}_p(\boldsymbol{\beta}, \mathbf{f})$ takes the form:

$$\mathcal{I}_p(\boldsymbol{\beta}, \mathbf{f}) = \begin{pmatrix} \mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathcal{I}_{\boldsymbol{\beta}\mathbf{f}} \\ \mathcal{I}_{\boldsymbol{\beta}\mathbf{f}}^{\top} & \mathcal{I}_{\mathbf{f}\mathbf{f}} \end{pmatrix}. \tag{7}$$

Thus, we obtain the variance of $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ estimated at convergence, respectively, as:

$$\begin{aligned} \widehat{\mathrm{Var}}_{\mathrm{approx}}(\widehat{\boldsymbol{\beta}}) &= \left(\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} - \mathcal{I}_{\boldsymbol{\beta}\mathbf{f}} \mathcal{I}_{\mathbf{f}\mathbf{f}}^{-1} \mathcal{I}_{\boldsymbol{\beta}\mathbf{f}}^{\top}\right)\Big|_{\widehat{\boldsymbol{\theta}}}, \\ \widehat{\mathrm{Var}}_{\mathrm{approx}}(\widehat{\mathbf{f}}) &= \left(\mathcal{I}_{\mathbf{f}\mathbf{f}} - \mathcal{I}_{\boldsymbol{\beta}\mathbf{f}}^{\top} \mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} \mathcal{I}_{\boldsymbol{\beta}\mathbf{f}}\right)\Big|_{\widehat{\boldsymbol{\theta}}}. \end{aligned}$$

Estimation of the smoothing parameter

Several authors have shown the connection between a smoothing spline and a linear mixed-effects model for analysis of longitudinal data (see, for instance, Speed, 1991; Wang, 1998).

Zhang et al. (1998) treated the smoothing parameter as an additional variance component. And, this parameter is estimated with other variance components simultaneously using restricted maximum likelihood (REML) estimation.

Motivated by Zhang et al. (1998) results and using the connection between the smoothing spline and LME models, we propose to estimate λ using the EM algorithm due to its simplicity of implementation and stable monotone convergence.

For more detail, see Mattos et al. (2022).

Goodness of fit

Under the assumption that $\mathbf{y}_i \overset{\text{ind.}}{\sim} t_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu)$, the Mahalanobis distance, $\delta_i^2(\boldsymbol{\theta}) = (\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$, has been considered by several authors to detect outliers in multivariate Student's-t models.

The statistics $F_i = \delta_i^2(\boldsymbol{\theta})/n_i$ is F-distributed with n_i and ν degrees of freedom, where n_i corresponds to the number of measurements associated with the ith subject.

Therefore, using the Wilson-Hilferty approximation (Johnson et al. (1994) and Galea-Rojas (1995)), we have that the transformed distance is

$$F_i^{[z]} = \frac{\left(1 - \frac{2}{9\nu}\right) F_i^{1/3} - \left(1 - \frac{2}{9n_i}\right)}{\left[\left(\frac{2}{9\nu}\right) F_i^{2/3} + \left(\frac{2}{9n_i}\right)\right]^{1/2}}, \quad i = 1, \ldots, n, \tag{8}$$

and follows approximately a standard normal distribution. Thus, a Q-Q plot of the transformed distances, $F_i^{[z]}$, can be used to assess the fit of the multivariate Student's-t distribution.

Model selection

For t-SMEC model, we define the AIC and BIC following the proposal of Taavoni et al. (2021) as follows:

$$
\begin{aligned}
AIC(\widehat{\boldsymbol{\theta}}) &= -2\ell(\widehat{\boldsymbol{\theta}}) + 2p^*, \\
BIC(\widehat{\boldsymbol{\theta}}) &= -2\ell(\widehat{\boldsymbol{\theta}}) + p^* \log N,
\end{aligned}
$$

where $\ell(\widehat{\boldsymbol{\theta}})$ corresponds to the logarithm of the observed likelihood function $\ell(\theta|\mathbf{y})$, p^* is the total number of parameters in the model, and N denotes the sample size.

Simulation study

We simulated data from the model

$$y_{ij} = \beta_1 x_{1_{ij}} + \beta_2 x_{2_{ij}} + f(t_{ij}) + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij},$$

with $i = 1, \dots, n$, $j = 1, \dots, n_i$, $(b_{0i}, b_{1i}) \overset{\text{ind.}}{\sim} t_2(\mathbf{0}, \mathbf{D}, \nu)$, and $\epsilon_{ij} \overset{\text{ind.}}{\sim} t_{n_i}(\mathbf{0}, \boldsymbol{\Omega}_i, \nu)$.

- The parameters were set at $\boldsymbol{\beta}^\top = (\beta_1, \beta_2) = (2, -1.5)$, $\sigma^2 = 0.13$, $\nu = 5$, and $\mathbf{D}$ with elements $\alpha_{11} = 0.25$, $\alpha_{12} = 0.01$, and $\alpha_{22} = 0.1$.
- We chose a smoothing function $f(t_{ij}) = \exp(\sin(0.3 t_{ij}) \cos(0.6 t_{ij}))$, with $t_{ij} = (1, 2, 3, 4, 5, 6, 7)$.
- For each sample size, we generated 500 samples of the DEC-SMEC model considering an AR(1) structure with parameter $\phi_1 = 0.8$.
- $x_1 \sim U(0, 1)$ and $x_2 \sim U(-1, 1)$, x_1 is independent of x_2.
- The censoring proportion was fixed at 10% and 20%, and sample sizes at $n = 50, 100$ and 300 were considered.

Simulation study - Asymptotic properties

Table: **Simulation study - Asymptotic properties**. Results based on 200 simulated samples.

m	Parameter	10% of censoring				20% of censoring			
		MC Mean	MC IM	MC SD	CP (%)	MC Mean	MC IM	MC SD	CP (%)
50	β_1	2.0007	0.0431	0.0386	96.4	2.0009	0.0482	0.0445	97.2
	β_2	-1.4987	0.0221	0.0199	97.2	-1.4981	0.0248	0.0219	97.6
	σ^2	0.1415				0.1165			
	ϕ_1	0.7666				0.7304			
	ν	6.1914				6.5627			
100	β_1	2.0040	0.0311	0.0287	96.8	2.0037	0.0348	0.0313	98.4
	β_2	-1.5019	0.0156	0.0141	97.2	-1.5018	0.0177	0.0162	97.2
	σ^2	0.1481				0.1132			
	ϕ_1	0.7992				0.7476			
	ν	5.5122				5.5816			
300	β_1	1.9998	0.0175	0.0145	98.8	1.9997	0.0196	0.0160	98.8
	β_2	-1.4995	0.0087	0.0078	96.0	-1.4997	0.0100	0.0091	96.8
	σ^2	0.1435				0.1109			
	ϕ_1	0.8116				0.7603			
	ν	5.1000				5.0965			

Evaluation of the parametric components

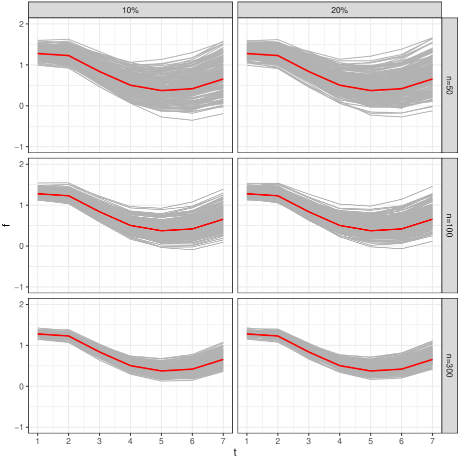

Figure: **Simulation study - Asymptotic properties**. Graphs of the non-parametric components with 200 replications. Adjusted curves (gray lines) and true curves (red lines) for all scenarios.

A5055 study

Our purpose is to investigate the relationship between the viral load and the immunological markers in AIDS clinical trials.

We considered the following model:

$$y_{ij} = \mathrm{CD4}^{+}_{ij}\beta_1 + \mathrm{CD8}^{+}_{ij}\beta_2 + f(t_{ij}) + b_{0i} + b_{1i}t_{ij} + \epsilon_{ij}, \tag{9}$$

where

- y_{ij} denotes the $\log_{10}$ transformation of the viral load for the ith subject at time t_{ij} $(i = 1, 2, \ldots, 44\ ;\ j = 1, 2, \ldots, n_i)$;
- $t_{ij} = \text{day}_{ij}/7$ (week);
- $f(t_{ij})$ is an arbitrary smoothing function;
- b_{0i}, b_{1i} are the random intercept and random slope, respectively for the i-th patient;
- ϵ_{ij} are random errors.

A5055 study

Table: **A5055 dataset.** Model selection criteria for the *t*-SMEC and N-SMEC models under different correlation structures. Bold values indicate the best model.

Model	Criteria	Correlation Structure			
		AR(1)	CS	DEC	UNC
t-SMEC	AIC	**601.5439**	633.3916	601.7918	634.2410
	BIC	**664.9556**	696.8033	668.9336	693.9226
N-SMEC	AIC	612.6097	654.4795	**610.6652**	652.1291
	BIC	**672.2913**	714.1611	674.0769	708.0806

A5055 study

Table: **A5055 study**. Parameter estimates, SE indicates the standard errors.

	t-SMEC		N-SMEC	
Parameter	Estimate	SE	Estimate	SE
β_1	-0.3854	0.1099	-0.5266	0.0969
β_2	0.0745	0.0733	0.1092	0.0706
f_1	3.6997	0.1288	3.6063	0.1361
f_2	3.0630	0.3525	3.0680	0.3867
f_3	2.5954	0.1246	2.6507	0.1348
f_4	2.3098	0.1307	2.2734	0.1414
f_5	1.8553	0.1660	1.7452	0.1703
f_6	1.7132	0.1887	1.6694	0.1865
f_7	1.6861	0.2366	1.8735	0.2209
f_8	2.0109	0.2577	2.2626	0.2446
f_9	1.7938	0.3036	1.9835	0.2888
σ^2	0.4514		0.7607	
α_{11}	0.0231		0.0157	
α_{12}	0.0020		-0.00003	
α_{22}	0.0021		0.0031	
ϕ_1	0.8604		0.8621	
ν	4.7991		-	
λ	19.8736		36.5393	

A5055 study

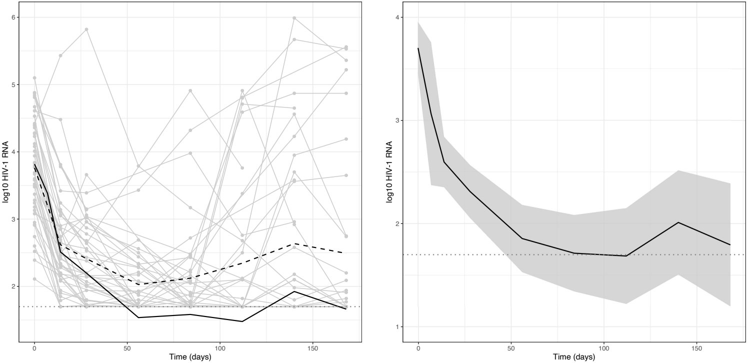

Figure: **A5055 study**. (Left panel) Viral loads in $\log_{10}$ scale (gray line) for all subjects, estimated mean trajectory (solid line) for the *t*-SMEC model under the AR structure and empirical mean trajectory (dotted line). (Right panel) Fitted curve of non-parametric part. The shaded regions denote the 95% confidence intervals obtained by $\widehat{\mathbf{f}} \pm 1.96\sqrt{\widehat{\mathrm{Var}}(\widehat{\mathbf{f}})}$.

A5055 study

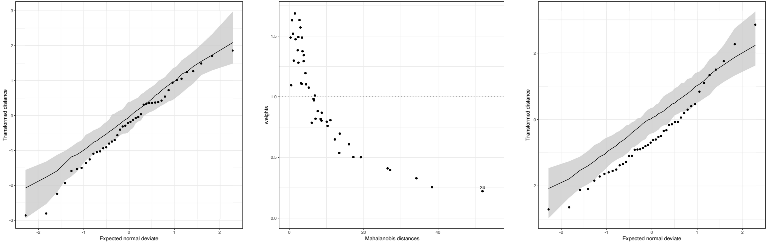

Figure: **A5055 dataset**. (Left panel) Normal probability plot for the transformed distance under the *t*-SMEC model with AR structure. (Middle panel) Estimated weights ($\widehat{u_i}$) for the estimated *t*-SMEC model with AR structure. (Right panel) Normal probability plot for the transformed distance under the N-SMEC model with AR structure. The shaded regions are the empirical envelopes obtained through bootstrap.

A5055 study

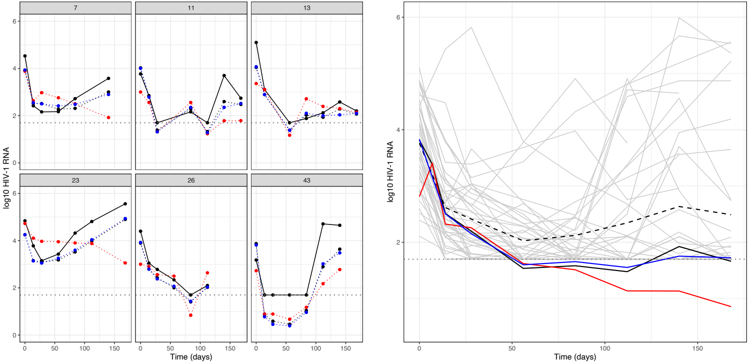

Figure: **A5055 data set.** (Left panel) Viral loads in log10 scale (solid black line) and estimated trajectories for each model for six subjects (dotted lines). (Right panel) Viral loads in log10 scale (gray line) for all subjects, empirical mean of the trajectories (dotted line) and the mean of the estimated trajectories under each model (solid lines).

Conclusions

- This work proposed a semi-parametric mixed model to analyze longitudinal censored data, assuming that the within-individual measurement errors and the random effects were distributed with Student's-t multivariate distribution.

- Simulation studies carried out suggest that the proposed method performs very well in estimating the parametric part and the nonparametric function.

- The approach was applied to analyze HIV-AIDS studies, showing the t-SMEC model's flexibility to fit those data sets in which we do not know the functional form that relates the response variable with the covariates.

- It would thus also be interesting to consider a broader family of distributions such as the multivariate skew-normal distribution (Azzalini and Valle, 1996) and the multivariate skew-t distribution (Azzalini and Genton, 2008), which could be more realistic for the random effects and error terms.

References I

Acosta, E. P., H. Wu, S. M. Hammer, S. Yu, D. R. Kuritzkes, A. Walawander, J. J. Eron, C. J. Fichtenbaum, C. Pettinelli, D. Neath, et al. (2004). Comparison of two indinavir/ritonavir regimens in the treatment of hiv-infected individuals. *JAIDS Journal of Acquired Immune Deficiency Syndromes 37*(3), 1358–1366.

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

Azzalini, A. and A. D. Valle (1996). The multivariate skew-normal distribution. *Biometrika 83*(4), 715–726.

Galea-Rojas, M. (1995). *Calibraçao Comparativa Estrutural e Funcional*. Ph. D. thesis, Tese de Doutorado em Estatıstica, Universidade de Sao Paulo, Sao Paulo.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique 55*, 245–259.

Hughes, J. (1999). Mixed effects models with censored data with application to HIV RNA levels. *Biometrics 55*, 625–629.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distributions*. Wiley New York.

Lachos, V. H., L. A. Matos, L. M. Castro, and M.-H. Chen (2019). Flexible longitudinal linear mixed models for multiple censored responses data. *Statistics in Medicine 38*(6), 1074–1102.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(2), 226–233.

Matos, L. A., L. M. Castro, and V. H. Lachos (2016, Dec). Censored mixed-effects models for irregularly observed repeated measures with applications to HIV viral loads. *TEST 25*(4), 627–653.

Matos, L. A., M. O. Prates, M.-H. Chen, and V. H. Lachos (2013). Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. *Statistica Sinica 23*, 1323–1345.

Mattos, T. B., L. Avila Matos, and V. H. Lachos (2021). A semiparametric mixed-effects model for censored longitudinal data. *Statistical Methods in Medical Research 30*(12), 2582–2603.

References II

Mattos, T. B., V. H. Lachos, L. M. Castro, and L. A. Matos (2022). Extending multivariate student's-t t semiparametric mixed models for longitudinal data with censored responses and heavy tails. *Statistics in Medicine*.

Munoz, A., V. Carey, J. P. Schouten, M. Segal, and B. Rosner (1992). A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics 48*(3), 733–742.

Segal, M. R., P. Bacchetti, and N. P. Jewell (1994). Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(2), 345–352.

Speed, T. (1991). [that blup is a good thing: The estimation of random effects]: Comment. *Statistical Science 6*(1), 42–44.

Taavoni, M., M. Arashi, W.-L. Wang, and T.-I. Lin (2021). Multivariate *t* semiparametric mixed-effects model for longitudinal data with multiple characteristics. *Journal of Statistical Computation and Simulation 91*(2), 260–281.

Vaida, F. and L. Liu (2009). Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics 18*, 797–817.

Vock, D. M., M. Davidian, A. A. Tsiatis, and A. J. Muir (2011). Mixed model analysis of censored longitudinal data with flexible random-effects density. *Biostatistics 13*(1), 61–73.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(1), 159–174.

Zeger, S. L. and P. J. Diggle (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics 50*, 689–699.

Zhang, D., X. Lin, J. Raz, and M. Sowers (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association 93*(442), 710–719.

Thank you!

www.ime.unicamp.br/~larissam

larissam@unicamp.br

Financial Support:

FAPESP