Longitudinal Models for Multiple Censored Responses Data

Larissa A. Matos, Victor H. Lachos, Luis M. Castro and Ming-Hui Chen

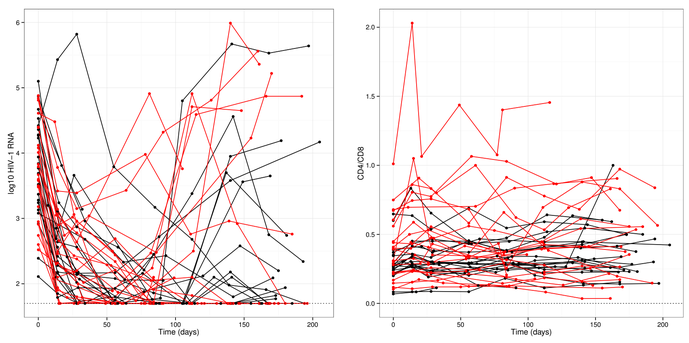
UNIVERSIDADE ESTADUAL DE CAMPINAS — UNICAMP — IMECC

Introduction

The multivariate censored linear mixed model (MLEMC) is a frequently used tool for a joint analysis of more than one series of longitudinal data. Motivated by a concern of sensitivity to potential outliers or data with longer-than-normal tails and possible serial correlation, we develop a robust generalization of the MLMEC that is constructed by using the scale mixtures of normal (SMN) distributions.

Motivation: A5055 clinical trial

- → 44 infected patients with the human immunodeficiency virus type 1 (HIV-1);
- → These patients were treated with one of two potent ARV therapies;
- → 2 response variables: the viral load ($\log_{10}$(RNA)) and the CD4/CD8, where CD4 and CD8 two immunologic markers frequently used to monitor disease progression in AIDS studies ;
- → 33.5% (106 out of 316) of measurements lies below the limits of assay quantification (left-censored).



SMN distributions $\mathbf{Y} \sim \mathbf{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{H})$

The stochastic representation is given by

$$\mathbf{y} = \boldsymbol{\mu} + \kappa(\mathrm{U})^{1/2}\mathbf{Z}, \tag{1}$$

where $\mu \in \mathbb{R}$ is a location vector, $Z \sim N(0, \boldsymbol{\Sigma})$, U is a positive random variable with cumulative distribution function (cdf) $\mathrm{H}(u|\boldsymbol{\nu})$ and probability density function (pdf) $h(u|\boldsymbol{\nu})$, $\boldsymbol{\nu}$ is a scalar or parameter vector indexing the distribution of U and $\kappa(\mathrm{U})$ is the weight function, with $Z \perp \mathrm{U}$.

Special cases distributions: $\mathbf{y} \in \mathbb{R}^p$

1. **The multivariate normal**

   If $P(U = 1) = 1$;

2. **The multivariate Student's-t**

   If $\mathrm{U} = Gama(\nu/2, \nu/2)$ and $\kappa(u) = 1/u$.

3. **The multivariate slash**

   If $\mathrm{U} = Beta(\nu, 1)$ and $\kappa(u) = 1/u$.

4. **The multivariate contaminated normal**

   If U is a discrete random variable taking one of two states and with probability function given by $h(u|\boldsymbol{\nu}) = \nu\mathbb{I}_{\{\gamma\}}(u) + (1-\nu)\mathbb{I}_{\{1\}}(u)$ and $\boldsymbol{\nu} = (\nu, \gamma)$ and $\kappa(u) = 1/u$.

References

Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions. Journal of the Royal Statistical Society.

Delyon, B., Lavielle, M. & Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. Annals of Statistics.

Muñoz, A., Carey, V., Schouten, J. P., Segal, M. & Rosner, B. (1992). A parametric family of correlation structures for the analysis of longitudinal data. Biometrics.

Samson, A., Lavielle, M. & Mentr ÌAe, F. (2006). Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: application to HIV dynamics model. Computational Statistics & Data Analysis.

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

Wang, W.-L., Lin, T.-I. & Lachos, V. H. (2015). Extending multivariate-t linear mixed models for multiple longitudinal data with censored responses and heavy tails. Statistical Methods in Medical Research.

Zhang, D., Chen, M.-H., Ibrahim, J. G., Boye, M. E., Wang, P. & Shen, W. (2014). Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials. Statistics in Medicine.

SAEM Algorithm

Let $\boldsymbol{\theta}$ be the parameter vector and $\boldsymbol{y}_c = (\mathbf{y}^\top, \boldsymbol{q}^\top)$ be the vector of complete data, i.e., the observed data $\mathbf{y}^\top$ and the missing/censored data (or the latent variables, depending on the situation) $\boldsymbol{q}^\top$.

E-Step: **Simulation:** Draw $\mathbf{q}^{(k,l)}$ $(l = 1, \ldots, m)$ from the conditional distribution $f(\mathbf{q}|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(k-1)})$; **Stochastic-approximation:** Update $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ according to

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)}) + \delta_k \left[\frac{1}{m}\sum_{l=1}^{m}\ell_c(\boldsymbol{\theta}|\mathbf{q}^{(k,l)}, \mathbf{y}) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)})\right],$$

where $\ell_c(\boldsymbol{\theta} \mid \mathbf{y}_c)$ is the complete log-likelihood function and δ_k is a smoothness parameter, *i.e.*, a decreasing sequence of positive numbers such that $\sum_{k=1}^{\infty}\delta_k = \infty$ and $\sum_{k=1}^{\infty}\delta_k^2 < \infty$.

M-Step: Update $\boldsymbol{\theta}^{(k)}$ according to $\widehat{\boldsymbol{\theta}}^{(k+1)} = \underset{\theta}{\mathrm{argmax}} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$.

The statistical model

Consider $\mathbf{y}_i = \mathrm{vec}(\mathbf{Y}_i) = (\mathbf{y}_{i1}^\top, \ldots, \mathbf{y}_{ir}^\top)^\top$, and $\boldsymbol{\epsilon}_i = \mathrm{vec}(\mathrm{E}_i) = (\boldsymbol{\epsilon}_{i1}^\top, \ldots, \boldsymbol{\epsilon}_{ir}^\top)^\top$, which are of dimension $s_i = n_i \times r$. The linear mixed effect model for the ith subject can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \;\; i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_r^\top)^\top$ is the $p \times 1$ vector of fixed effects associated with the design matrix $\mathbf{X}_i$ and $\mathbf{b}_i = (\mathbf{b}_{i1}^\top, \ldots, \mathbf{b}_{ir}^\top)^\top$ is the $q \times 1$ vector of random effects associated with the design matrix $\mathbf{Z}_i$, with $p = \sum_{j=1}^{r} p_j$ and $q = \sum_{j=1}^{r} q_j$.

Instead of the usual assumption of normality for the errors and random effects, we replace the multivariate normal distribution by the scale mixture of multivariate normal distributions, thus it follows that the model can be expressed as

$$\mathbf{y}_i \mid \mathbf{b}_i \overset{\text{ind.}}{\sim} \mathrm{SMN}_{s_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i; \mathrm{H}_1), \quad \text{and} \quad \mathbf{b}_i \overset{\text{ind.}}{\sim} \mathrm{SMN}_q(\mathbf{0}, \mathbf{D}; \mathrm{H}_2), \;\; i = 1, \ldots, n. \tag{3}$$

Using the stochastic representation, the hierarchical representation (four-stages) to the model defined is

$$\begin{aligned}
\mathbf{y}_i \mid \mathbf{b}_i, \kappa_i &\overset{\text{ind.}}{\sim} \mathrm{N}_{s_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \kappa_i^{-1}\mathbf{R}_i),\\
\mathbf{b}_i \mid \tau_i &\overset{\text{ind.}}{\sim} \mathrm{N}_q(\mathbf{0}, \tau_i^{-1}\mathbf{D}),\\
\kappa_i &\overset{\text{ind.}}{\sim} \mathrm{H}_1(\nu)\\
\tau_i &\overset{\text{ind.}}{\sim} \mathrm{H}_2(\eta), \;\; i = 1, \ldots, n,
\end{aligned} \tag{4}$$

where $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha}) = \mathbf{D}_{jj'}$ is a $q \times q$ dispersion matrix that depends on the unknown and reduced parameters $\boldsymbol{\alpha}$. We also assume that the within-subject errors for one given response at different occasions have serial correlation which is described by a $n_i \times n_i$ autocorrelation matrix $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}_i(\boldsymbol{\phi}; \mathbf{t}_i)$ that has a parsimonious dependence structure involving only parameter $\boldsymbol{\phi}$ and measurement time $\mathbf{t}_i$ of subject i, and that for the multiple responses at a particular occasion are correlated with an $r \times r$ variance-covariance matrix $\boldsymbol{\Sigma} = [\sigma_{jj'}^2]$. Accordingly, $\mathbf{R}_i = \boldsymbol{\Sigma} \otimes \boldsymbol{\Omega}_i$, where $\otimes$ denotes the Kronecker product. We adopt a DEC (damped exponential correlation) structure for $\boldsymbol{\Omega}_i$, which is defined as:

$$\boldsymbol{\Omega}_i = \boldsymbol{\Omega}_i(\boldsymbol{\phi}, \mathbf{t}_i) = \left[\phi_1^{|t_{ij} - t_{ik}|^{\phi_2}}\right], \;\; i = 1, \ldots, n, \;\; j, k = 1, \ldots, n_i, \tag{5}$$

where $\boldsymbol{\phi} = (\phi_1, \phi_2)^\top$, the parameter ϕ_1 describes the autocorrelation between observations separated by the absolute length of two time points, and the parameter ϕ_2 permits acceleration of the exponential decay of the autocorrelation function, defining a continuous-time autoregressive model.

We are interested in the case where left-censored observations can occur. That is, the observed data for the i-th subject is represented by $(\mathbf{V}_i, \mathbf{C}_i)$, where $\mathbf{V}_i$ is the vector of uncensored observation or limit of quantification and $\mathbf{C}_i$ is the vector of censoring indicator whose value equals one if censored observation and zero if uncensored observation; such that, considering the left censored case, we have that

$$y_{ijk} \leq V_{ijk} \quad \text{if} \;\; C_{ijk} = 1, \quad y_{ijk} = V_{ijk} \quad \text{if} \;\; C_{ijk} = 0.$$

Apllicattion

The model

$$\begin{aligned}
y_{i1k} &= \beta_{10} + \beta_{11}t_{ik} + \beta_{12}\mathrm{treat}_i + \beta_{13}t_{ik}^{0.5}\\
&+ \beta_{14}\mathrm{treat}_i \times t_{ik} + b_{i10} + b_{i11}t_{ik} + e_{i1k},\\
y_{i2k} &= \beta_{20} + \beta_{21}t_{ik} + \beta_{22}\mathrm{treat}_i + \beta_{23}\mathrm{treat}_i \times t_{ik}\\
&+ b_{i20} + b_{i21}t_{ik} + e_{i2k},
\end{aligned}$$

where y_{i1k} is the $\log_{10}$ (RNA) response for subject i measured at t_k and y_{i2k} is the log(CD4/CD8) response for subject i measured at t_k; 33% of all viral load measurements are below the detection limit.

Parameters	Estimate (SE)	Parameters	Estimate (SE)
β_{10}	3.743 (0.134)	d_{11}	0.1446 (0.0829)
β_{11}	0.130 (0.026)	d_{21}	0.0011 (0.0133)
β_{12}	-0.005 (0.067)	d_{22}	-0.0884 (0.1182)
β_{13}	-0.957 (0.098)	d_{31}	-0.0011 (0.0033)
β_{14}	-0.007 (0.025)	d_{32}	0.0034 (0.0027)
β_{20}	-1.284 (0.077)	d_{33}	-0.0122 (0.0116)
β_{21}	0.005 (0.005)	d_{41}	-0.0004 (0.0004)
β_{22}	0.252 (0.084)	d_{42}	0.2727 (0.0861)
β_{23}	-0.003 (0.007)	d_{43}	0.0008 (0.0015)
ν	4.737 (0.003)	d_{44}	0.0001 (0.0001)
σ_{11}	0.409 (0.076)	σ_{21}	-0.039 (0.020)
σ_{22}	0.050 (0.011)		
ϕ_1	0.704 (0.065)	ϕ_2	0.632 (0.131)
$loglik$	-344.79	AIC	739.59
		BIC	850.81

Conclusions

In this work, we have introduced a robust multivariate linear mixed model for multiple censored responses based on the class of SMN distributions. The main advantage of the proposed SMN-MLMEC model is that it can reduce the negative impact of distributional misspecification and outliers on the parameter estimation. Moreover, the SMN class allows a convenient framework for implementing the SAEM algorithm, leading to an efficient ML estimation of model parameters. An additional characteristic of our proposed model is that it allows considering different distributions for the error terms, thereby overcoming the aforementioned limitation of the MLMEC model and broadening the scope of censored mixed models.

WORKSHOP EM ANÁLISE DE SOBREVIVÊNCIA E APLICAÇÕES — WASA