skewlmm: An R package for robust estimation of linear mixed models based on scale mixture of skew-normal distributions

Fernanda L. Schumacher, Larissa A. Matos, Victor H. Lachos

fernandalschumacher@gmail.com

Introduction

Linear mixed models are frequently used to analyze repeated measures data, because they model flexibly the within-subject correlation often present in this type of data. Usually for mathematical convenience, it is assumed that both random effect and error term follow normal distributions. These restrictive assumptions, however, may result in a lack of robustness against departures from the normal distribution and invalid statistical inferences, especially when the data show heavy tails and skewness.

Another common feature of these classes of LMMs is that the error terms are conditionally independent. However, in longitudinal studies, repeated measures are collected over time and hence the error term tends to be serially correlated. Extending the proposal of Lachos et al. [4], Schumacher et al. [7] proposed a full likelihood approach via an EM-type algorithm for fitting scale mixture of skew-normal linear mixed models (SMSN-LMM) with serially correlated errors, considering some useful correlation structures, such as autoregressive correlation of order p (AR(p)) [3] and damped exponential correlation (DEC) [6].

The methods developed in Schumacher et al. [7] were implemented in the R package *skewlmm* [8], available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=skewlmm`. This work aims to briefly introduced the proposed model and to describe the main features of the *skewlmm* package, which offers an automatic fit of SMSN-LMM and contain some tools for model evaluation, allowing users to make robust inferences in practical longitudinal data analysis.

Motivation

To motivate the need of a more flexible model, a real dataset regarding reaction times in a sleep deprivation study, which is available at the R package *lme4*, will be introduced. The average reaction time per day for subjects was evaluated by Belenky et al. [2] in a sleep deprivation study, where on day 0 the subjects had their normal amount of sleep and starting that night they were restricted to 3 hours of sleep per night for 9 days, and the reaction time based on a series of tests was measured on each day for each subject.

Figure 1 presents individual reaction time trajectories evolved over time along with their mean profile, and results of the empirical Bayes estimates of random effects obtained from fitting a normal LMM, where it can be seem that the normal model does not seem to be appropriate, since the quantile plots indicate heavy tails.

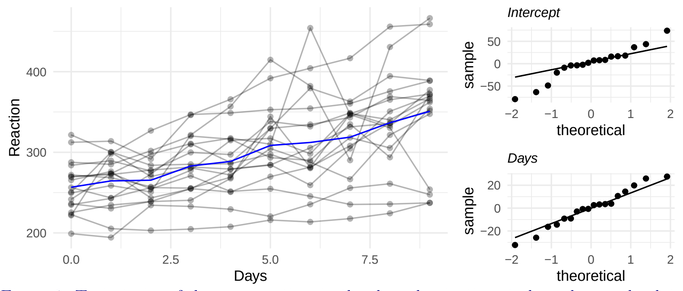

Figure 1: Trajectories of the reaction time in the sleep deprivation study and quantile plots of the empirical Bayes estimates of random effects obtained from fitting a normal LMM.

Model formulation

The skew-normal (SN$_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$) distribution [1] can be defined from

$$f(\mathbf{y}) = 2\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(A), \quad \mathbf{y} \in \mathbb{R}^p,$$

where $A = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$. If $\mathbf{W} \sim \text{SN}_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, the scale mixture of skew-normal (SMSN$_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}; H)$) class of distributions is the distribution of

$$\mathbf{Y} = \boldsymbol{\mu} + \kappa(U)^{1/2}\mathbf{W},$$

where U is a positive random variable with cdf $H(\cdot; \boldsymbol{\nu})$, independent of $\mathbf{W}$, and $\kappa(u)$ is a positive weight function. Thus, the pdf of $\mathbf{Y}$ is

$$f(\mathbf{y}) = 2\int_0^\infty \phi_p(\mathbf{y}; \boldsymbol{\mu}, \kappa(u)\boldsymbol{\Sigma})\Phi(\kappa(u)^{-1/2}A)dH(u; \boldsymbol{\nu}),$$

$\mathbf{y} \in \mathbb{R}^p$. Considering $\kappa(u) = u^{-1}$, we get the skew-normal/independent (SNI) class of distributions, the depending of the distribution of U we can derived the skew-t (ST), skew-slash (SSL), and the skew-contaminated normal (SCN) distribution. Furthermore, the symmetric versions of the distributions are attained when $\boldsymbol{\lambda} = 0$, when the SMSN reduces to the scale mixture of normal (SMN$_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; H)$) [5] class.

Now, when a variable of interest together with several covariates are repeatedly measured for each of n subjects at certain occasions over a period of time. For the ith subject, $i = 1, \ldots, n$, let $\mathbf{Y}_i$ be a $n_i \times 1$ vector of observed continuous responses. The SMSN-LMM can be defined by considering

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

and

$$\begin{pmatrix} \mathbf{b}_i \\ \boldsymbol{\epsilon}_i \end{pmatrix} \overset{\text{ind.}}{\sim} \text{SMSN}_{q+n_i}\left(\begin{pmatrix} c\boldsymbol{\Delta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_i \end{pmatrix}, \begin{pmatrix} \boldsymbol{\lambda} \\ \mathbf{0} \end{pmatrix}; H\right), \tag{2}$$

where $\mathbf{X}_i$ of dimension $n_i \times l$ is the design matrix corresponding to the fixed effects, $\boldsymbol{\beta}$ of dimension $l \times 1$ is a vector of fixed effects, $\mathbf{Z}_i$ of dimension $n_i \times q$ is the design matrix corresponding to the $q \times 1$ random effects vector $\mathbf{b}_i$, $\boldsymbol{\epsilon}_i$ of dimension $n_i \times 1$ is the vector of random errors, $c = c(\boldsymbol{\nu}) = -\sqrt{2/\pi}k_1$, with $k_1 = \text{E}\{U^{-1/2}\}$, $\boldsymbol{\Delta} = \mathbf{D}^{1/2}\boldsymbol{\delta}$, and $\boldsymbol{\delta} = \boldsymbol{\lambda}/\sqrt{1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda}}$.

The $q \times q$ random effects scale matrix $\mathbf{D}$ can be unstructured or diagonal, and we consider the $n_i \times n_i$ error scale matrix as $\boldsymbol{\Sigma}_i = \sigma_e^2\mathbf{R}_i$, with $\mathbf{R}_i = \mathbf{R}_i(\boldsymbol{\phi})$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$, being one of the following:

- Uncorrelated (UNC): $\mathbf{R}_i = \mathbf{I}_{n_i}$.
- Autoregressive dependence of order p (AR(p)):

$$\mathbf{R}_i = \mathbf{R}_i(\boldsymbol{\phi}) = \frac{1}{1 - \phi_1\rho_1 - \ldots - \phi_p\rho_p}[\rho_{|r-s|}],$$

  where $\rho_1, \ldots, \rho_p$ are the theoretical autocorrelations of the process and functions of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$, and they satisfy the Yule-Walker equations.
- Damped exponential correlation (DEC):

$$\mathbf{R}_i = \mathbf{R}_i(\phi_1, \phi_2, \mathbf{t}_i) = \left[\phi_1^{|t_{ij} - t_{ik}|^{\phi_2}}\right], \quad 0 < \phi_1 < 1, \ \phi_2 > 0.$$

The SMSN-LMM has a convenient hierarchical representation, which is useful for the implementation of an EM-type algorithm, and therefore the package *skewlmm* uses the ECME algorithm for parameter estimation through the functions `smsn.lmm` and `smn.lmm`, where the latter refers to the special case of $\boldsymbol{\lambda} = \mathbf{0}$. An introduction to the package and its use to fit a SMSN-LMM to the sleep study data is given next.

The R package *skewlmm*

The package *skewlmm* provides tools for fitting and evaluating the SMSN-LMM given in (1)-(2) in R using S3 class, with a user-friendly interface.

The basic syntax of the main functions is as follows:

```
smsn.lmm(data, formFixed, groupVar, formRandom,
         depStruct, distr, covRandom, ...)
smn.lmm(data, formFixed, groupVar, formRandom,
        depStruct, distr, covRandom, ...)
```

where

- `data` is a data frame containing all the variables to be used in the model.
- `formFixed` is a two-sided linear formula object describing the fixed effects part of the model.
- `groupVar` is a character containing the name of the variable which represents the subjects or groups in data.
- `formRandom` is an one-sided linear formula object describing the random effects part of the model.
- `depStruct` is a character indicating which dependence structure should be used.
- `distr` is a character indicating which distribution should be used.
- `covRandom` is either `"pdSymm"` or `"pdDiag"`.

The functions return an object of the class `SMSN` and `SMN`, respectively, containing a list of elements. Additionally, some estimation options can be controlled using the argument `control` with the function `lmmControl`.

For example, a SL-LMM and a SSL-LMM, respectively, can be fitted given below. Additionally, a likelihood ratio test for testing $\text{H}_0 : \boldsymbol{\lambda} = \mathbf{0}$ can be performed using the `lr.test` function.

```
fit1<-smn.lmm(data = sleepstudy, formFixed = Reaction~Dayst,distr = 'sl',
              formRandom = ~Dayst, groupVar = "Subject",
              control = lmmControl(quiet = TRUE))
fitskew1<-smsn.lmm(data = sleepstudy, formFixed = Reaction~Dayst,distr = 'ssl',
                   formRandom = ~Dayst, groupVar = "Subject",
                   control = lmmControl(quiet = TRUE))
lr.test(fit1,fitskew1)
```

```
## Model selection criteria:
##             logLik      AIC      BIC
## fit1     -861.088 1736.175 1758.526
## fitskew1 -860.745 1739.490 1768.226
##
##     Likelihood-ratio Test
## chi-square statistics =  0.6854122
## df =  2
## p-value =  0.7098468
## The null hypothesis that both models represent the
## data equally well is not rejected at level  0.05
```

To evaluate the adequacy of the distributional assumption, a Healy-type plot can be used as illustrated next, where the gain in considering a heavy-tailed distribution for modeling this dataset is evidenced.

```
grid.arrange(healy.plot(fit0,calcCI = TRUE),
             healy.plot(fit1,calcCI = TRUE),nrow=1)
```

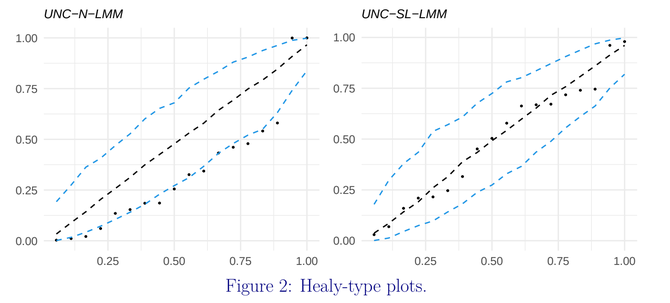

Figure 2: Healy-type plots.

Furthermore, to check if the uncorrelation assumption that is used by default is appropriate, a possible approach is to use the function `update` to refit the model considering different correlation structures and then compare AIC and BIC for selecting the most appropriate model. Since the data are equally spaced and sorted by time, the use of `timeVar` in here is optional (the function will use the position if `timeVar` is not provided).

```
fitar1<-update(fit1, depStruct = "ARp",
               pAR = 1)
fitar2<-update(fit1, depStruct = "ARp",
               pAR = 2)
fitDEC<-update(fit1, depStruct = "DEC",
               timeVar = "Days")
```

depStruct	AIC	BIC
UNC	1736.2	1758.5
AR(1)	**1716.8**	**1742.3**
AR(2)	1717.3	1746.0
DEC	1718.6	1747.3

Additionally, we can compute the empirical autocorrelation function (ACF) for standardized marginal residuals, which at lag l can be defined as

$$\hat{\rho}(l) = \frac{\sum_{i=1}^n \sum\{(j,k)|t_k - t_j = l\} r_{it_j} r_{it_k}/N(l)}{\sum_{i=1}^n \sum_{j=1}^{n_i} r_{it_j}^2/N(0)},$$

where $\mathbf{r}_i = \hat{\boldsymbol{\Upsilon}}_i^{-1/2}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)$ is the standardized marginal residual vector for subject i, with $\boldsymbol{\Upsilon}_i = \text{Var}(\mathbf{Y}_i)$, and $N(\cdot)$ is the number of pairs used in the respective numerator summation.

```
grid.arrange(plot(acfresid(fit1, calcCI =  TRUE, maxLag =  6))+
               ggtitle("UNC-SL-LMM"),
             plot(acfresid(fitar1, calcCI =  TRUE, maxLag =  6))+
               ggtitle("AR(1)-SL-LMM"), nrow=1)
```

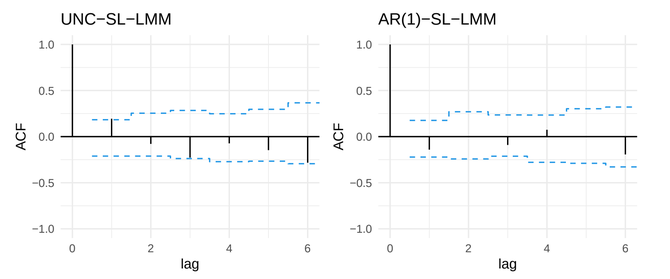

Figure 3: ACF plots.

Finally, methods such as `print`, `summary`, `plot`, `fitted`, `residuals` and `predict` are implemented, and an example of its use is given below.

```
plot(fitar1, type =  "normalized") + ggtitle('AR(1)-SL-LMM')
```

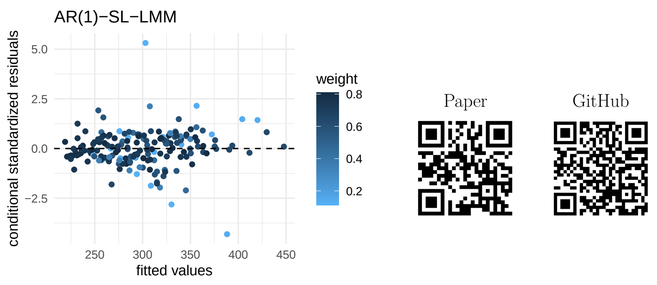

Figure 4: Plotting a fitted object.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPq).

References

[1] A. Azzalini and A. Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

[2] G. Belenky, N. J. Wesensten, D. R. Thorne, M. L. Thomas, H. C. Sing, D. P. Redmond, M. B. Russo, and T. J. Balkin. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of Sleep Research*, 12(1):1–12, 2003.

[3] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, CA, 1976.

[4] V. H. Lachos, P. Ghosh, and R. B. Arellano-Valle. Likelihood based inference for skew–normal independent linear mixed models. *Statistica Sinica*, 20:303–322, 2010.

[5] K. L. Lange and J. S. Sinsheimer. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2:175–198, 1993.

[6] A. Muñoz, V. Carey, J. P Schouten, M. Segal, and B. Rosner. A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics*, 48:733–742, 1992.

[7] F. L. Schumacher, V. H. Lachos, and L. A. Matos. Scale mixture of skew-normal linear mixed models with within-subject serial dependence. *Statistics in Medicine*, 40(7):1790–1810, 2021.

[8] F. L. Schumacher, L. A. Matos, and V. H. Lachos. *skewlmm: Scale Mixture of Skew-Normal Linear Mixed Models*, 2021. R package version 0.2.3.