Identifying change points for linear mixed models: A solution through Evolutionary Algorithms

Ehidy K. Garcia , Juan C. Correa & Juan C. Salazar

National University of Colombia Campus Medellín, Pedagogical and Technological University of Colombia, Sogamoso

ekgarciac@unal.edu.co, ehidy.garcia@uptc.edu.co — Colombia

Poster ID: 037

Abstract

The Change Point problem arises in many applied situations. The Change Point problem has been studied by several authors. It goes from the change point problem in piecewise regression through classical techniques to Change Point estimation in linear mixed models by using a dynamic programming algorithm. The objective of this proposal is estimating each subject specific change point by using Evolutionary algorithms when we consider data coming from a longitudinal setting, using linear mixed models (LMMs). The results will be based on a simulation study, varying some specific conditions on the parameters associated to the LMM. We illustrate the method with a real data set about dried Cypress wood slats in which this methodology is useful to predict the time of dried associated to a specific slat thickness. This is done as a generalization of the calibration problem. In this case, once the change points have been obtained through Evolutionary Algorithms, a calibration curve can be fitted to these change points according to their own thickness. This will allow predicting the specific change point.

Linear models and Linear Mixed Models

In a particular way, fitting a simple linear regression model (SLRM) implies to quantify the effect of the predictor (X) on the response variable (Y). This is done through the estimation of the model parameters and a posterior residual analysis. After data are collected and the model is specified, the next step is to estimate the vector of parameters $\boldsymbol{\beta}$. The general statement for a SLR model is

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n \tag{1}$$

with $\varepsilon_i \sim N\left(0, \sigma^2\right)$ $(i.i.d.)$ that corresponds the usual normality, independence and homocedasticity assumptions about the random error and n is the number of paired observations.

A Linear Mixed Model (LMM) is a parametric linear model for cross-sectional or longitudinal data. It quantifies the effect of explanatory variables on the response variable. The model given in (1), can be extended as follows:

$$\begin{aligned} \mathbf{Y}_i = & \underbrace{\mathbf{X}_i\boldsymbol{\beta}}_{\text{Fixed}} + \underbrace{\mathbf{Z}_i\mathbf{b}_i}_{\text{Random}} + \boldsymbol{\varepsilon}_i \\ & \mathbf{b}_i \sim N\left(\mathbf{0}, \mathbf{D}\right) \\ & \boldsymbol{\varepsilon}_i \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_i\right) \\ & \boldsymbol{\varepsilon}_i \text{ and } \mathbf{b}_i \text{ are independent,} \end{aligned} \tag{2}$$

where $\mathbf{Y}_i$ is a $n_i \times 1$ vector of continuous responses for the $i-$th subject.

Change point problem

For a single change point and using SLRM, the model is expressed as:

$$y_i = \begin{cases} \beta_{10} + \beta_{11} x_i + \varepsilon_{1i} & i = 1, \ldots, s \qquad \varepsilon_{1i} \sim N\left(0, \sigma_1^2\right) \\ \beta_{20} + \beta_{21} x_i + \varepsilon_{2i} & i = s+1, \ldots n \qquad \varepsilon_{2i} \sim N\left(0, \sigma_2^2\right) \end{cases} . \tag{3}$$

here $x_s = \tau$, where τ is the change point of the model.

Instead of estimating a common change point as Lai, et al. (2014) suggested, the main goal of this study is to estimate subject-specific change points, by using a linear mixed models approach we will estimate a point where the data changes.

We have considered a model with fixed effects before the change point and once the change point is found there is only a random effect that allows to explain the variation for each subject right after the change point. With only one continuous change point the model can be written as:

$$y_{ij} = \left(\beta_1 \mathbf{x}_{ij} + \beta_2 t_{ij}\right) \mathbf{1}\left(t_{ij} \leq \tau_i\right) + \left(\beta_1 \mathbf{x}_{ij} + \beta_2 \tau_i\right) \mathbf{1}\left(t_{ij} > \tau_i\right) + b_{00i} + b_{01i} 1\left(t_{ij} \leq \tau_i\right) + b_{02i} 1\left(t_{ij} > \tau_i\right) + \varepsilon_{ij}$$

And simplified as:

$$y_{ij} = \beta_1 \mathbf{x}_{ij} + \beta_2 t_{ij} \mathbf{1}\left(t_{ij} \leq \tau_i\right) + \beta_2 \tau_i \mathbf{1}\left(t_{ij} > \tau_i\right) + b_{00i} + b_{01i} 1\left(t_{ij} \leq \tau_i\right) + b_{02i} 1\left(t_{ij} > \tau_i\right) + \varepsilon_{ij} \tag{4}$$

where $\mathbf{x}_{ij}$ is the vector of values for the fixed effects associated to each subject at time t_{ij} these can be the same all over the time or it could change along the time if it is needed, and τ_i is the subject specific change point. The average model given by

$$y_{ij} = \beta_1 \mathbf{x}_{ij} + \beta_2 t_{ij} \mathbf{1}\left(t_{ij} \leq \tau_i\right) + \beta_2 \tau_i \mathbf{1}\left(t_{ij} > \tau_i\right) + \varepsilon_{ij} \tag{5}$$

and its average change points can be seen graphically as:

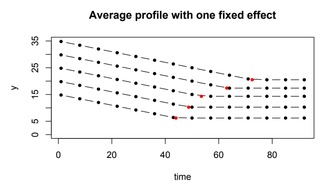

Figure 1: Change in regression regime for LMMs.

Developments have been made to identify change points in SLRMs and LMMs. This proposal considers the estimation of the subject-specific change points by using Evolutionary algorithms. (Price, et al. (2006), Storn (1997), Lai & Albert (2014))

Results

The global optimization process, based on a Differential Evolutionary Algorithm (DEA), requires a target function. In this case, the objective is to estimate a vector of values $\boldsymbol{\tau} =< \tau_1, \ldots, \tau_n >$, that is, subject specific change points by using LMMs on a longitudinal data set. The log-likelihood for an LMM, as it was written in (2) is given by:

$$\ell\left(\boldsymbol{\theta}\right) = -\frac{k}{2}\log\left(2\pi\right) - \frac{1}{2}\sum_{i=1}^{n}\log\left|\mathbf{V}_i\right| - \frac{1}{2}\left\{\sum_{i=1}^{n}\left(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)' \mathbf{V}_i^{-1}\left(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)\right\} \tag{6}$$

Once we got the results and by comparing the estimated points with the profiles on the graph, they coincide with the simulated points, with an outstanding precision. The values, found through a differential evolutionary algorithm, showed that the estimation of this change points is quite associated to the specific value given to the fixed effect, and the standard deviation is quite small. As **table 1** shows:

Change points using DEA for a LMM $n = 25, n_i = 14$

	Fixed effects values $\mathbf{x}_i$				
	1	2	3	4	5
id	$50mm$	$40mm$	$30mm$	$25mm$	$20mm$
1	65.171	56.210	49.342	47.046	40.448
2	65.359	57.573	48.738	44.213	42.088
3	67.417	60.824	47.427	47.488	37.969
4	65.148	58.623	48.109	48.139	41.413
5	65.223	60.328	47.076	48.597	42.235
τ	65	57	49	45	41
$\bar{\hat{\tau}_i}$	65.6636	58.7116	48.1384	47.0966	40.83
$S_{\hat{\tau}_i}$	0.9836	1.9129	0.9279	1.7181	1.7484

Table 1: Change points Using DEA for 25 subjects

A real data set

Dried Cypress wood slat data was collected by Botero (1993). He conducted an experiment about dried Cypress wood slats that considered 20 slats per each thickness of 50 mm, 40 mm, 30 mm, 25 mm and 20 mm and taking the measurements, every 7 days for around 92 days, about the percentage humidity of each slat.

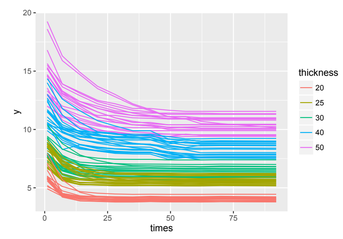

Figure 2: Humidity percentage vs time per each wood slat in the experiment
Source:Author

Change points using EA for a LMM $n = 100, n_i = 14$

	Fixed effects values *Thickness*				
id	$50mm$	$40mm$	$30mm$	$25mm$	$20mm$
1	46.493	39.055	26.763	21.515	16.699
2	41.979	35.315	27.271	20.365	16.382
3	43.871	32.047	27.904	28.515	20.715
4	45.210	35.042	28.578	25.647	15.512
5	40.508	32.608	27.016	22.924	21.096
6	41.826	33.479	27.835	25.464	23.310
7	45.576	31.917	26.739	21.836	18.351
8	44.247	37.215	31.769	20.526	15.010
9	40.318	30.310	30.394	22.584	24.067
10	41.946	31.141	27.243	23.763	24.099
11	48.875	31.757	32.623	20.091	19.911
12	42.569	32.937	29.566	21.613	18.726
13	44.158	33.631	34.190	27.331	15.682
14	44.637	36.064	29.115	23.851	19.684
15	41.280	35.524	25.407	21.632	15.633
16	44.049	35.755	26.688	28.232	24.328
17	47.245	39.618	28.011	20.029	20.721
18	42.327	30.459	30.115	29.076	17.599
19	41.159	36.307	28.222	25.708	18.730
20	45.634	35.842	32.707	23.460	22.258
$\bar{\hat{\tau}_i}$	43.695	34.3012	28.903	23.7081	19.4257
$S_{\hat{\tau}_i}$	2.3495	2.693	2.3793	2.9313	3.098

Table 2: Change Points using DEA to the wood data set

Conclusions

- The adapted algorithm allows to estimate a global maximum more than a local maximum which is one the main advantages of Evolutionary Algorithms.
- The change points associated to the real data set are quite precise according to the subject-specific profile.
- The results from the real data set generate new questions about the utility of this change points and its behaviour on large samples.
- In the illustration with real data, we observed that this approach could permit to predict, in a plausible and precise way, the change point given a specific value for the thickness. From a practical point of view, this prediction process allows to reduces both storage time and storage expenses.

Forthcoming Research

As a result of the current work on this topic we plan to do the following

- To explore the asymptotic properties associated to the calibration function parameters from this subject specific change points.
- To implement a Bayesian methodology to obtain the calibration function associated to these change points.

References

[1] Sebastian Botero. Secado de la madera cipres para uso industrial: estibas, molduras y muebles. Bachelor thesis, Tesis Universidad Nacional de Colombia. Sede Medellin. Facultad de Ciencias agropecuarias, May 1993.

[2] Juan Correa and Juan Salazar. *Introduction to mixed models*. Universidad Nacional de Colombia Campus Medellin, 2016.

[3] Ehidy Garcia, Juan Correa, and Juan Salazar. A calibration function built from change points: a review. *Comunicaciones en Estadistica*, 10(1):113–128, 2017.

[4] RG Krutchkoff. Classical and inverse regression methods of calibration in extrapolation. *Technometrics*, 11(3):605–608, 1969.

[5] Yinglei Lai and Paul Albert. Identifying multiple change points in a linear mixed effects model. *Statist. Med*, 33:1015 – 1028. doi: 10.1002/sim.5996, 2014.

[6] Katharine M Mullen, David Ardia, David L Gil, Donald Windover, and James Cline. Deoptim: An r package for global optimization by differential evolution. 2009.

[7] Kenneth Price, Rainer M Storn, and Jouni A Lampinen. *Differential evolution: a practical approach to global optimization*. Springer Science & Business Media, 2006.

[8] . R Core Team. R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria, 2015, 2015.

[9] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[10] Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2009.

Acknowledgements

The authors acknowledge to the School of Statistics, National University of Colombia, Campus Medellín and to the School of Industrial Engineering, UPTC, Campus Sogamoso, which have guaranteed the financial resources for the development of this work.