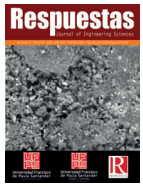

Respuestas
Journal of Engineering Sciences

RESPUESTAS

Original Article https://doi.org/10.22463/0122820X.2678

Identifying change points for linear mixed models: a solution through evolutionary algorithms

Estimación de puntos de cambio para modelos lineales mixtos: una solución usando algoritmos evolutivos

Ehidy Karime Garcia-Cruz*[1], Hugo Fernando Castro-Silva[2], Juan Carlos Salazar-Uribe[3]

[1*]*Ph.D. ekgarciac@unal.edu.co. National University of Colombia - Pedagogical and Technological University of Colombia. ORCID id: https://orcid.org/0000-0003-0034-7162. Dirección: Cra. 65 #59a-110. Teléfono: (4) 4309000. Medellín, Colombia.*

[2]*Ph.D. hugofernando.castro@uptc.edu.co. Pedagogycal and Technologycal University of Colombia. ORCID id: https://orcid.org/0000-0001-6020-402X. Dirección: Calle 4 A Sur No. 15-134. Teléfono: (8) 7707721. Sogamoso, Colombia.*

[3]*PhD. jcsalaza@unal.edu.co. National University of Colombia Medellin Headquarters. ORCID id: https://orcid.org/0000-0003-2286-3627. Dirección: Cra. 65 #59a-110. Teléfono: (4) 4309000. Medellín, Colombia.*

How to cite: E.K. Garcia-Cruz, H.F. Castro-Silva y J.C. Salazar-Uribe, "Identifying change points for linear mixed models: a solution through evolutionary algorithms". *Respuestas,* vol. 25, no. S2, pp. 69-78, 2020.

Received on September 10, 2019; Approved on December 13, 2019

ABSTRACT

Mathematical models are used to describe the relationship between two or more variables or features over the target population. Statistically, Simple Linear regression model has been extensively applied and the properties of their estimators are well known. However, this kind of model is not correctly applied in most cases, such as a longitudinal setting. Linear mixed models (LMMs) are useful when the measurements have been done over a specific interval of time. One of the most important assumptions, on both models, has been established as that the model holds for the whole data. In latter case, we could find one or several points which the function changes into. This proposal allows us to estimate the points where the model changes by minimizing a specific risk function or a loss function associated with the fitted model.

Keywords:

Linear Mixed Models;
Change points;
Evolutionary Algorithms.

RESUMEN

Los modelos matemáticos han sido usados para describir la relación entre dos o más variables o características de la población objetivo. En el campo de la estadística el modelo de regresión lineal simple han sido aplicados ampliamente y sus propiedades son bien conocidas. Sin embargo, este tipo de modelos no tiene aplicación an algunos casos como en el escenario longitudinal. Los modelos lineales mixtos (LMMs) son de utilidad cuando las medidas han sido registradas en un intervalo de tiempo. Uno de los más importantes supuestos, para los dos modelos, ha sido el establecer que el modelo se mantiene invariante sobre todo el intervalo de regresión. En caso contrario, se puede encontrar uno o varios puntos de cambio en los cuales el modelo cambia (media, varianza o simultaneamente media - varianza). Esta propuesta permite estimar los puntos de cambio sujeto-específicos que permiten la minimización de una función de riesgo o pérdida asociada al modelo en mención.

Palabras clave:

ModelosLineales mixtos; punto de cambio; Algoritmos evolutivos.

Introduction

Longitudinal data arises frequently in many applied sciences as agriculture, medicine, finances, among others. The term longitudinal suggests that data have been collected over time on the same unit which implies some structure of correlation for data. Some recognized authors have been worked on this topic Linear Mixed Models (LMMs) arises as an alternative to deal with residual correlation structures and the very usual problem of unbalanced data. A lot of theory have been widely developed for this models under some important assumptions since Harville's works [1, 2]. Fitting a model implies that it is correct for all the points of the regressor variable, however, this is not always true. Sometimes the model changes its mean or variance structure from a specific point.

Modeling and predicting is an interesting area which has been incorporated as one of the most powerful tools in data analysis. The regression analysis had been the most useful statistical technique for modeling the relationship between a dependent variable (response variable) and one or more independent variables (explanatory variables or predictors). This technique is a common method used in practice [3, 4].

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

*Corresponding author-
E-mail address- *ekgarciac@unal.edu.co* (Ehidy Karime Garcia Cruz)

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 find the parameter estimations. The general statement for a SLR model is: $Yi = \alpha + \beta Xi + \varepsilon i$, $i = 1, 2, \ldots, n$ (1) with $\varepsilon i \sim N(0, \sigma^2)$ (i.i.d.) that corresponds the usual normality, independence and homoscedasticity assumptions about the random error and n is the number of paired observations. The estimators' close expressions are well known and their properties, even when normality assumption fails,can be found at [5].

An extension from the model given by (1) is a Linear Mixed Model (LMM) due to the error component can be separated into two components the former a raw error component and the latter an explained error component. The second one contains information about the individual difference from the average model. An LMM is a parametric linear model for cross-sectional or longitudinal data. It quantifies the effect of explanatory variables on the response variable. This model allows that fixed and random effects can be estimated and predicted respectively. While fixed effects estimates give information about the relationship between a set of independent variables and the response, random effects are used in modeling the random variation in the dependent variable at different levels, clusters or subjects within a population ([6–8]). The general structure for an LMM is given by:

$$Y_i = X_i \beta + Z_i b_i + \varepsilon_i \quad (2)$$
$$b_i \sim N(0, D)$$
$$\varepsilon_i \sim N(0, \sum_i)$$
b_i and ε_i are independent,

where Y_i is a ni × 1 vector of continuous responses for the i−th subject. X_i, is a ni × p (p = k + 1) design matrix that considers all possible variables that affects the response values for the i−th subject. β, is a vector of p unknown fixed effects parameters. *Zi* is a ni × q matrix and represents the q observed values for the predicted variable in the i−th subject which affect the continuous variable that vary randomly across the subjects. *D*, is a q × q covariance matrix that consider the covariances between two random effects for each subject, here D is considered as an unstructured or a diagonal matrix. And finally, Σi, is a matrix in which the residuals association with observations on the same subject are considered under some specific assumptions so that $V(Y_i) = Z_i D Z_i' + \sum_i$. Some common choices for this matrix are: diagonal, compound symmetry , a first order autoregressive or a Toeplitz structure [6, 9, 10]. Each out of these structures could yield a better estimation of the coefficients in accordance with the data behavior.

Change point problem

Fitting a simple linear regression model to a data set in a cross-sectional setting is a common practice. It is usually assumed that the considered model holds for the whole data. However, sometimes researchers need to consider linear models where the structure of the model changes. An exploratory data analysis could allow to detect a change in the model structure in either any specific point or several points. The point in which the structure changes is called the change point. In general, change points can be known or unknown.

When a model has one or several change points we need to consider them to build a better model, in both cross-sectional and longitudinal settings. In the cross-sectional setting the change point problem has been studied extensively. For a single change point and using SLRM, the model is expressed as:

$$Y_i = \begin{cases} \beta_{01} + \beta_{11} X_i + \varepsilon_{1i} & i = 1, \ldots, s \quad \varepsilon_{1i} \sim N(0, \sigma_1^2) \\ \beta_{02} + \beta_{12} X_i + \varepsilon_{2i} & i = s+1, \ldots, n \quad \varepsilon_{2i} \sim N(0, \sigma_1^2) \end{cases} \quad (3)$$

here xs = τ, where τ is the change point of the model. If we assume τ is known then without loss of generality [11, 12], we can obtain the MLE by using the usual procedure, getting it as $\hat{\tau} = (\hat{\beta}_{02} - \hat{\beta}_{01}) / (\hat{\beta}_{1}^{1} - \hat{\beta}_{\_}^{12})$. But, If τ is unknown, the MLE cannot be found using a traditional $\hat{\beta}_{11} - \hat{\beta}_{12}$ procedure because the log-likelihood has not a solution in a closed form for estimating β and τ simultaneously, then implementing a numerical procedure is required. These numerical procedures use a grid search over a set of candidate values of τ, so an optimisation process is implicit to find the specific value of τ which maximizes the log-likelihood written as:

$$l(\beta|\tau, y) = l(\beta_1 | \tau, y_1, y_2, \ldots, y_s) + A((\beta \mid \tau, y) = l(\beta_1, \beta_2 \mid \tau, y_{s+1}, \ldots, y_n)$$

[13] remarked that an explicit form for $\hat{\tau}$ does not exist, but it is conveniently defined by variables associated with random walks. He found both the exact and the asymptotic properties associated to the change point estimator, based

on an iterative process by defining the log-likelihood increments. The estimation procedure explained by him obey to a numerical solution due to most of the integrals must be calculated numerically. [14], [15] expressed that the asymptotic distribution of the change-point MLE is quite complicated and an exact computable expression for the distribution is not available in the literature. They established an exact and yet computationally attractive form for the asymptotic distribution of the change-point MLE, at the theorem (2.1) and found the methodology quite useful, readily computable and quite robust to deviations from the Gaussianity assumption. Some R-packages available to identify these change points on a piecewise regression could be done by using Segmented created by [16] or SiZer proposed by [17] and detecting change points over time series can be worked on R by using some packages as Change point done by [18].

In longitudinal settings, authors usually work by fitting linear mixed models but to find the change points in this scenery is not an easy task. In this way, identifying change points in linear mixed models (LMMs) is an open problem and has been studied by few authors. Recent works on this topic were done by [19]. They proposed a methodology Based on a Dynamic Programming Algorithm to find common change points for the whole set of subjects using LMMs.

They proposed to get those common change points by consider an extension of the model (2) by considering a model with g change points and general random effects, so there exists g+1 groups and the extension was expressed as:

$$y_{ij}=(\beta_1 x_{ij}+\beta_2 t_{ij})1(t_{ij}\leq\tau_i)+(\beta_1 x_{ij}+\beta_2 \tau_i)1(t_{ij}>\tau_i)+b_{00i}+b_{01i} 1(t_{ij}\leq\tau_i)+b_{02i}1(t_{ij}>\tau_i)+\varepsilon_{ij}$$

And simplified as:

$y_{ij} = \beta_1 x_{ij}+ \beta_2 \{t_{ij} 1(t_{ij} \leq \tau_i)+\tau_i 1(t_{ij} > \tau_i)\}+b_{00i}+b_{01i} 1(t_{ij} \leq \tau_i)+b_{02i} 1(t_{ij} > \tau_i)+\varepsilon_{ij}$ (6) where xij is the vector of values for the fixed effects associated to each subject at time tij these can be the same all over the time or it could change along the time if it is needed. The average model given by

$$y_{ij}=\beta_1 x_{ij}+\beta_2 t_{ij} 1(t_{ij}\leq\tau_i)+\beta_2 \tau_i 1(t_{ij}>\tau_i)+\epsilon_{ij} \quad (7)$$

and its average change points can be seen graphically as:

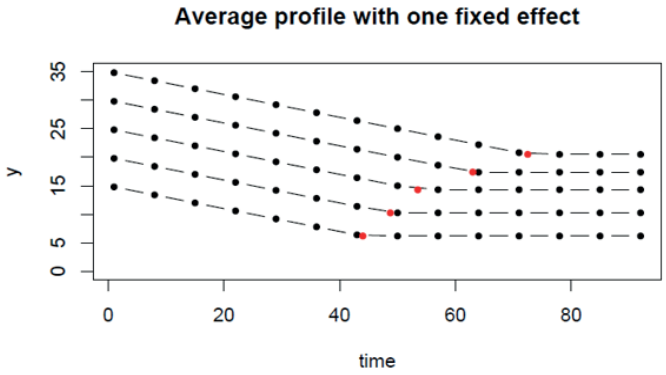

Figure 1: Change in regression regime for LMMs. Source: Author. Simulated data

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.

Evolutionary Algorithms

Evolutionary algorithms (EAs) is an evolution of genetic algorithms (GAs) and they allow to minimize an objective function over the course of successive generations ([21]). Some authors have worked and referenced the main advantages and utility of this technique in several knowledge fields and as a strategy for solving multi-objective problems [22], [23], [24], [25], [21].

Given a quality function to be maximized, we can randomly create a set of candidate solutions, i.e., elements of the function's domain [22]. Then by applying the quality function to these as an abstract fitness measure (the higher, the better). On the basis of these fitness values, some of the better candidates are chosen to seed the next generation. This is done through a recombination and/or mutation process to those values.

Recombination is an operator that is applied to two or more selected candidates (parents) producing one or more new candidates (the children). Mutation is applied to one candidate and the result is one new candi- date. Therefore, executing the operation of recombination and mutation on the parents leads to the creation of a set of new candidates (the offspring). These candidates have their fitness evaluated and then compete with the old ones for a place in the next generation. This process must be iterated until a candidate with sufficient quality is selected as the best member of the population.

Respuestas, 25 (S2), pp. 69-78 , 2020, ISSN 0122-820X - E ISSN: 2422-5053

There are two main forces that form the basis of evolutionary systems. The former is due to the variation operators (combination and mutation) because these factors create necessary diversity within the population, and thereby facilitate novelty. Selection acts as a second force, increasing the mean quality of solutions in the population.

The most important components of Evolutionary Algorithm are: Representation, evaluation function, population, parent selection mechanism, variation operators: recombination and mutation and finally, a Survivor selection mechanism. However, these components can be adapted according to the goal of the optimization process. So, there is not a unique way to execute an EA. It exists variants of the evolutionary algorithms. These methods can be found detailed at the Eiben's book [22]. Some of these methods have been imple- mented at an R package [26]. In this proposal, we worked by using a Differential Evolution Algorithm (DEA).

A Differential Evolution Algorithm follows the structure:

BEGIN
INITIALISE population with random candidate solutions;
EVALUATE each candidate;
While $f(x_i) - f(x_{i-1}) > \varepsilon$ {
1. SELECT parents;
2. RECOMBINE pair of parents;
3. MUTATE the resulting offspring;
4. EVALUATE new candidates;
5. SELECT individuals for the next generation;
}
END

Differential Evolution Algorithm (DEA), so called due to that a differential mutation process is done, to get the reproduction operators. Given a population of candidates solution vectors in R^n a new mutant vector $\mathbf{s}^*$ is generated by adding a perturbation vector **p** to an old **s** vector, so:

$$\mathbf{s}^*=\mathbf{s}+\mathbf{p}$$

and p is the scaled vector difference of two other randomly chosen population members

$$\mathbf{p}=\mathbf{F}*(\mathbf{s}_{r1,g}-\mathbf{s}_{r2,g}),$$

where F > 0 is a real number that controls the rate at which the population evolves, and r1 and r2 are randomly selected one per mutant. The flow chart for this algorithm was given by [21], as follows:

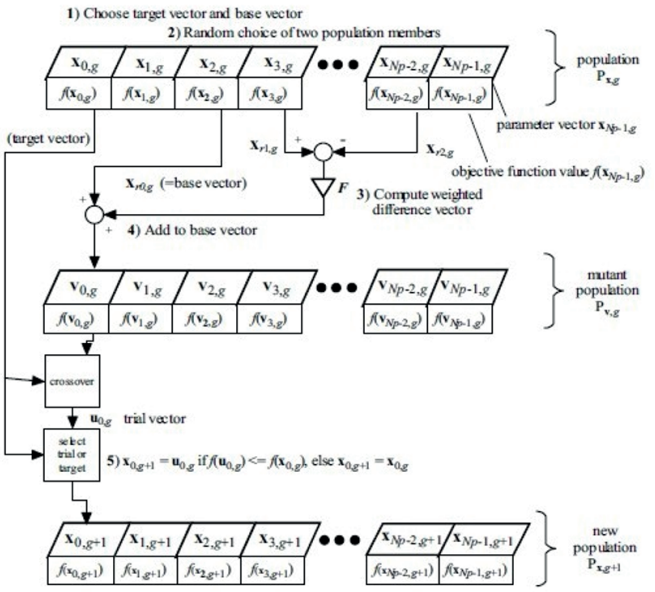

Figure 2: Flow chart to Differential evolutionary algorithms. Source: Price, et. al. (2006) [21]

Details about the algorithm are given by [25], [22] and [21]. The procedure was implemented by [27] in R-Software through a full package called DEoptim. The notation used by them is referred as: DE/rand/1/bin, which refers to a DE algorithm with a base vector randomly chosen, 1 means that a vector difference is added to it and bin means that the mutate vector closely follows a binomial distribution. To make this process faster [28] implemented DEoptim package functions inside of a new package based on a C + + code.

Methodology

The global optimization process based on DE algorithms require a target function. In this case, the main goal is to estimate a vector of values $\tau =< \tau_1, \ldots, \tau n >$, 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:

$$l(\theta) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n} log|V_i| - \frac{1}{2}\{\sum_{i=1}^{n}(Y_i - X_i\beta)'V_i^{-1}(Y_i - X_i\beta)\} \quad (8)$$

Where $\theta = (\beta, D, \sum, \tau), V_i = Z_i D Z_i^T + \sum_i$ and

$$Y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ \vdots \\ y_{in_i} \end{bmatrix}, X_i = \begin{bmatrix} x_{11} & t_{i1}1(t_{i1} \le \tau_i) \\ x_{12} & t_{i2}1(t_{i2} \le \tau_i) \\ x_{13} & t_{i3}1(t_{i3} \le \tau_i) \\ x_{1s} & \tau_i 1(t_{is} \le \tau_i) \\ . & . \\ . & . \\ x_{1n_i} & \tau_i 1(t_{i1} \le \tau_i) \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$b = \begin{bmatrix} b_0 \\ b_2 \\ b_3 \end{bmatrix}, D = \begin{bmatrix} Var(b_0) & Cov(b_0, b_1) & Cov(b_0, b_1) \\ & Var(b_1) & Cov(b_1, b_2) \\ & & Var(b_2) \end{bmatrix}$$

$$\sum_i = \frac{\sigma^2}{1-\sigma^2}\begin{bmatrix} 1 & \rho & \rho^2 \dots & \dots & \rho^{ni-1} \\ \rho & 1 & \rho \dots & \dots & \rho^{ni-2} \\ \rho^2 & \rho & 1 \dots & \dots & \rho^{ni-3} \\ \dots & \dots & \dots & \dots & . \\ . & . & . & . & . \\ \rho^{ni-1} & \rho^{ni-2} & \rho^{ni-3} \dots & & 1 \end{bmatrix} \quad (9)$$

Without losing generality, $\mathbf{X}_i$ is a simpler version for a general matrix due to in this case we assume only one explanatory variable in this study, but it can be extended to a p + 1 dimensional matrix. The last matrix $\mathbf{\Sigma}_i$ means that on the simulation process the random error follows an AR(1) process, it implies that each measurement made at each time depends on the last measure gotten one time before. It is a reasonable assumption according to the behavior of each subject on the real data set which we are working with.

As the goal of this proposal is concentrated into estimate as exact as possible the change points the target function will be changed and it will be considered the Restricted Maximum Likelihood given by:

$$lR(\theta) = -\frac{1}{2}\sum_{i=1}^{n} log[V_i] - \frac{1}{2}\left\{\sum_{i=1}^{n}\frac{1}{2}(Y_i - X_i\beta)'V_i^{-1}(Y_i - X_i\beta)\right\} - \frac{1}{2}log\left|\sum_{i=1}^{n} X_i'V_i^{-1}X_i\right| \quad (10)$$

The main idea behind REML estimation is to separate the part of the data used for estimation of $\mathbf{V}_i$ from that used of estimation of β. It is due to a procedure based on maximum likelihood, as it is well-known, produces biased estimators finite samples. Particularly it underestimates the diagonal elements of $\mathbf{V}_i$ [10]. The difference between ML and REML estimation becomes less important when n >> p

Results

Even though the change points associated to this kind of model can not be expressed in a close way, we can find a way to recursively update the specific value for τi, see appendix. Under the common assumptions for an LMM we have

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i \quad (11)$$

$$Y_i - X_i\beta \sim N_{ni}(0, ZDZ' + \sum_i) \quad (12)$$

The general expression to the model given at (6) is:

$$Y_i = \begin{bmatrix} x & t_{11} \\ x & t_{12} \\ . & . \\ . & . \\ x & \tau_i \\ . & . \\ x & \tau_i \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ : & : & : \\ . & . & . \\ 1 & 0 & 1 \\ . & & \\ . & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}\begin{bmatrix} b_{00} \\ b_{01} \\ b_{02} \end{bmatrix} + \varepsilon_i$$

This implies the log-likelihood function to be maximized is given by:

$$l(\beta_1, \beta_2, \tau) = -\frac{1}{2}log|V_i| - \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - X_i\beta(\tau_i)\right)^t V_i^{-1}(Y_i - X_i\beta(\tau_i))$$

or the REML log-likelihood function

$$l(\beta_1, \beta_2, \tau) = -\frac{1}{2}log|V_i| - \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - X_i\beta(\tau_i)\right)^t V_i^{-1}\left(Y_i - X_i\beta(\tau_i)\right) - \frac{1}{2}log|\sum_{i=1}^{n} X_i(\tau_i)'V_i^{-1}X_i(\tau_i)| \quad (13)$$

Simulation study

At the simulation study different sample sizes were studied and the subject specific change points were estimated. The first example to be shown considered 25 subjects. The results were gotten by using DEoptim R-software package. The executing time was improved by executing a parallel version of this procedure to speed up the outcome of this process. The results presented in this paper considered 200 generations and a differential weighting factor of 0.8. In the procedure, each mutant was tested against a member in the previous generation, and the best value proceeded into the next generation.

Respuestas, 25 (S2), pp. 69-78 , 2020, ISSN 0122-820X - E ISSN: 2422-5053

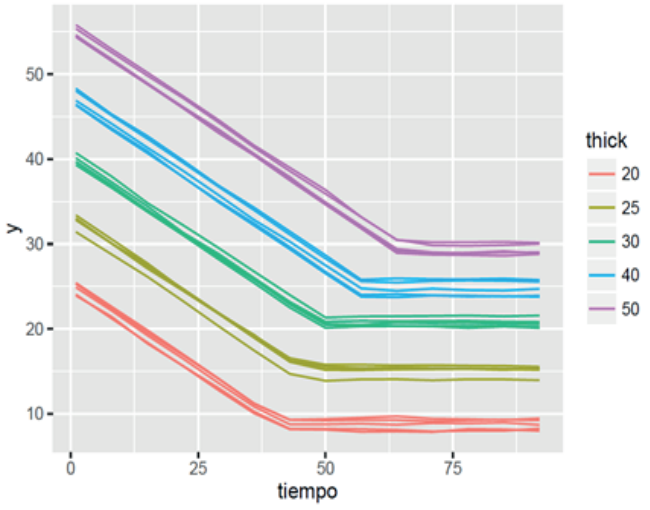

Figure 3: Profiles plot associated to 25 subjects. Source: Author

Table 1. Estimated change points using dea for 25 subjects in a lmm

Change points using DEA for a LMM
n = 25 ni = 14
Fixed effects values $\mathbf{x}_i$

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

Source: Author

Following with the simulation process we assumed, the study has 50 subjects each one with 14 the measurements made along 92 days. The simulated process can be seen on the graph.

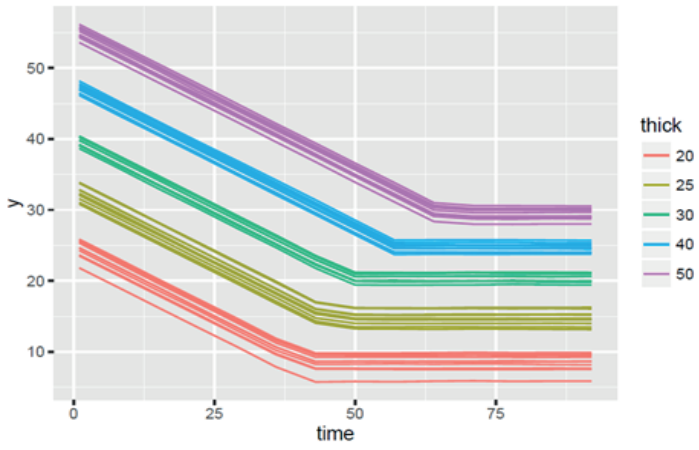

Figure 4: Profiles plot associated to 50 subjects. Source: Author

Table 2. Estimated change points using dea for 50 subjects in a lmm

Change points using DEA for a LMM
N=50, ni=14
Fixed effects values $\mathbf{x}_i$

	1	2	3	4	5
	$50mm$	$40mm$	$30mm$	$25mm$	$20mm$
	68.315	59.041	47.837	48.414	39.275
	66.515	55.484	52.807	43.835	38.350
	65.915	54.012	46.199	46.735	41.289
	64.826	55.581	51.174	46.929	39.104
	65.981	60.668	46.514	48.675	41.729
	63.724	60.761	45.487	47.636	42.543
	62.801	60.682	52.217	46.388	41.406
	65.008	57.563	49.501	44.543	37.278
	62.529	55.035	50.924	43.341	37.292
	65.783	59.779	45.020	46.812	39.284
τ	65	57	49	45	41
$\bar{\hat{\tau}}_i$	65.1397	57.8606	48.768	46.3308	39.755
$s_{\hat{\tau}i}$	1.7670	2.647	2.9163	1.8457	1.8798

Source: Author

Once we got the results and by comparing the estimated points with the profiles on the graph, they coincide with the simulated points, in some occasions the estimation even is more precise than the original value.

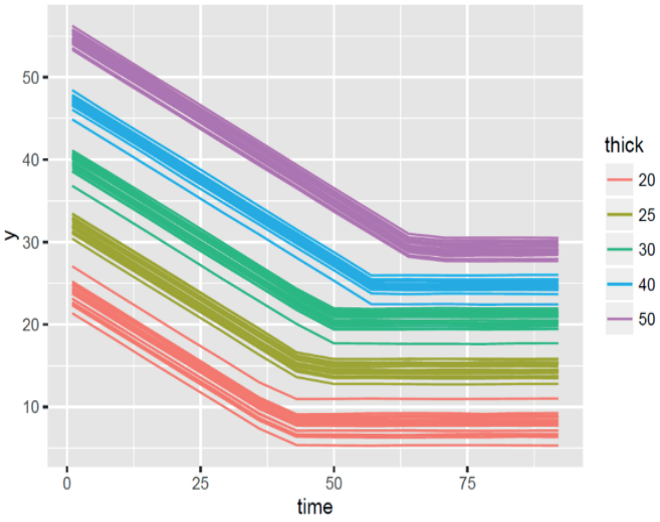

Figure 5: Profiles plot associated to 100 subjects. Source: Author

The last scenery in this simulation study was considering a whole amount of 100 subjects. In practice, it could be unattainable on longitudinal studies due to the large quantity of missing data or even censored data we should work with. However, it could help us to understand the distribution of the change points through a LMM by reaching a Global more than a local maximum.

Table 3. Estimated change points using dea for 100 subjects in a lmm change points using DEA for a LMM
n = 100, ni = 14

Fixed effects values xi

	1	2	3	4	5
	50*mm*	40*mm*	30*mm*	25*mm*	20*mm*
	65.334	55.856	48.89	43.847	41.667
	68.608	57.565	45.469	48.312	37.199
	63.589	58.679	52.208	45.092	39.460
	67.824	57.208	50.743	43.417	39.923
	64.409	53.411	46.494	45.125	37.090
	68.536	58.725	52.123	42.297	40.915
	68.052	56.093	45.757	48.185	38.252
	62.244	59.150	47.689	46.227	40.911
	65.325	59.450	49.920	47.338	38.254
	65.161	58.982	47.727	47.168	40.528
	61.181	56.609	45.713	48.734	42.168
	67.885	55.095	48.608	47.140	44.184
	66.899	58.364	50.033	43.993	37.962
	62.649	53.209	50.696	46.957	44.44
	66.985	60.408	49.613	46.492	39.519
	65.053	57.302	48.859	48.647	40.804
	66.751	53.443	49.671	46.839	38.758
	65.086	55.661	48.501	47.681	42.84
	64.179	57.861	50.896	47.226	42.944
	61.011	58.031	46.407	43.258	41.317
τ	65	57	49	45	41
$\bar{x}$	65.3381	57.055 1	48.801	46.199	40.4457
s_x	2.3659	2.0980	2.078	1.956	2.179

According to the developments made on this problem and the results obtained through the implementation of this algorithm, we propose the next theorem.

Theorem 1: Given a sequence of n, ni dimensional random vectors Y1, Y2, · · ·, Yn, such that each vector component corresponds to a data over time (individual longitudinal data set). Assuming that $yi = X_i\beta + Z_i b_i + \varepsilon_i$ and following the LMMs Gaussian
Scheme $Y_i \sim N_n (X_i \beta, Z_i DZ^T + \sum^\sum)$. Let be $\tau=(\tau_1,\tau_2,\ldots,\tau_n)'$ the subject specific change point vector, if:
i) $\varepsilon_i \sim N_{ni} (0,\sum_i)$, *where ∑i corresponds to an AR(1) process as it was given in (9)*
ii) $b_i \sim N_q (0,D)$
iii) Regularity conditions are fulfilled Then $\tau_i \sim N(x_i \gamma, S)$ where S corresponds to the variance once the DE algorithm has ended. Additionally,

$$\gamma = \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}$$

A real data set

The methodology proposed is illustrated now with a specific data set about dried cypress wood slats. The data set of data was collected by [29]. The thesis was titled: Drying of Cypress wood for industrial use: pallets, molding, and furniture. He made an experiment for his undergraduate monograph, which followed the agreement of Cartagena's joint (1982). He wanted to explain the amount of water lost in the cypress wood slats according to the thickness of the slats over time. He collected information about 20 of them per each thickness. The experiment considered slats of 50 mm, 40 mm, 30 mm, 25 mm and 20 mm and taking the measurements, in an interval of 7 days for around 92 days, about the percentage humidity of every slat.

Performing the experiment [29] highlighted that the protocol was followed correctly and he followed the recommenda- tions about that the wood can not be exposed directly to the sun ray. As, the wood win or lose water with equal easy in fiber direction and perpendicular thereto and he wanted to avoid the splits ends, frequently found on the slats, he applied an useful technique painting them using an Aluminum paint.

Respuestas, 25 (S2), pp. 69-78 , 2020, ISSN 0122-820X - E ISSN: 2422-5053

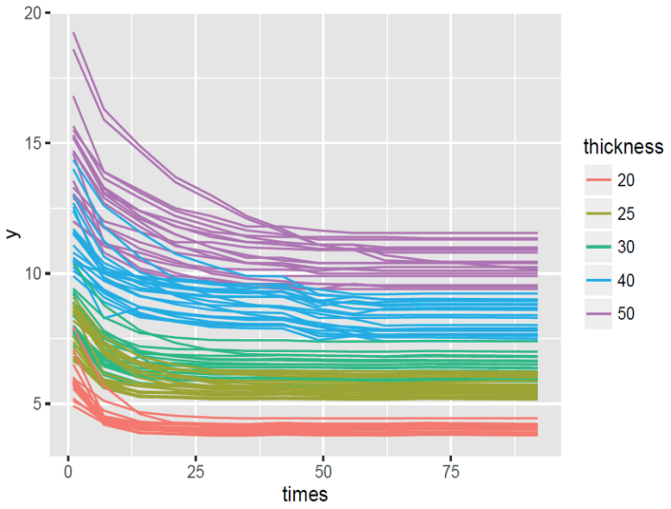

Figure 6: Profiles plot associated to Cypress wood slats. Source: Author

Table 4. Estimated change points using DEA for Dried wood slats data
Change points using EA for a LMM n = 100, ni = 14
Fixed effects values Thickness

id	50*mm*	40*mm*	30*mm*	25*mm*	20*mm*
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
$\bar{\hat{t}}_i$	43.695	34.3012	28.903	23.7081	19.4257
$s_{\hat{t}i}$	2.3495	2.693	2.3793	2.9313	3.098

As [29] describe in his thesis, the wood slats were weighted each 7 days until each wood strip reached the equilibrium humidity content. The initial humidity content was calculated using as measure the corrected humidity content. He suggested fitting an exponential model to explain the behavior of the wood dried. The change points were gotten by applying this procedure to the real data set. They are given at table 4.

The distribution associated to these change points, with n = 100 (20 subjects by each thickness), is approximately a normal distribution such as it was expressed at the theorem 1. That follows a regular media and it agrees with the results gotten by simulation with the exception that the variance and means just depend on the slat thickness which is considered as fixed value.

Remarks

The adapted methodology proposed allows estimating a global maximum more than a local maximum which is one the main advantages of Evolutionary Algorithms. The simulation results showed that the change points estimated by using this technique are quite near to the parameter value and they have been estimated quite accurately even with the sample size is small enough. The results for real data set showed that the change points estimated for the real data set are quite precise according to the subject-specific profile. The last results generate new questions about the utility of this change points and an increasing interest to model its behavior on large samples.Estimating these change points is considered a first step to build a calibration function to predict the change point given the particular values of the fixed effects. 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. Future works will incorporate the study on the asymptotic properties associated to the calibration function parameters from this subject specific change points and implementing a Bayesian methodology to obtain the calibration function associated to these change points.

Acknowledgements

The fisrt author wants to acknowledge the support given by the school of Statistics at National University of Colombia, campus Medellín. And GIRA researching group from UPTC, campus Sogamoso.

References

[1] D. A. Harville, "Bayesian inference for variance components using only error contrasts," *Biometrika,* vol. 61, no. 2, pp. 383–385, 1974.

[2] ——, "Maximum likelihood approaches to variance component estimation and to related problems," *Journal of the American Statistical Association*, pp. 320–338, 1977.

[3] G. A. Seber and A. J. Lee, Linear regression analysis. John Wiley & Sons. New Jersey, 2003, vol. 936.

[4] X. Yan, *Linear regression analysis:* theory and computing. World Scientific, 2009.

[5] C. R. Rao, *Linear statistical inference and its applications*. John Wiley & Sons, 1973, vol. 2nd. Ed.

[6] B. T. West, K. B. Welch, and A. T. Galecki, *Linear mixed models: a practical guide using statistical software*. Chapman & Hall, 2007.

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

[8] J. Correa and J. Salazar, *Introduction to mixed models*. Universidad Nacional de Colombia Campus Medellín, 2016.

[9] J. C. Pinheiro and D. M. Bates, *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media, 2000.

[10] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied longitudinal analysis*. John Wiley & Sons, 2008, vol. 998.

[11] J. Hofrichter, *Change point detection in generalized linear models*.Graz University of Technology. doi:http://www.stat.tugraz.at/dthesis/Hofrichter07.pdf. Accessed on: 2015/04/22, 2007.

[12] H. Küchenhoff, "An exact algorithm for estimating breakpoints in segmented generalized linear models," 1996.

[13] D. V. Hinkley, "Inference about the change-point in a sequence of random variables," *Biometrika*, pp. 1–17, 1970.

[14] R. Henderson, "Change-point problem with correlated observations, with an application in material accountancy," *Technometrics*, vol. 28, no. 4, pp. 381–389, 1986.

[15] S. B. Fotopoulos, V. K. Jandhyala, and E. Khapalova, "Exact asymptotic distribution of change-point mle for change in the mean of gaussian sequences," *The Annals of Applied Statistics*, pp. 1081–1104, 2010.

[16] V. M. Muggeo, "Segmented: an r package to fit regression models with broken-line relationships," *R news*, vol. 8, no. 1, pp. 20–25, 2008.

[17] D. Sonderegger and M. D. Sonderegger, "R package: Sizer," 2011.

[18] R. Killick and I. Eckley, "changepoint: An r package for changepoint analysis," J*ournal of Statistical Software,* vol. 58, no. 3, pp. 1–19, 2014.

[19] Y. Lai and P. Albert, "Identifying multiple change points in a linear mixed effects model," *Statist. Med*, vol. 33, pp. 1015 – 1028. doi: 10.1002/sim.5996, 2014.

[20] E. K. García, J. C. Correa, and J. C. Salazar, "A calibration function built from change points: a review," *Comunicaciones en Estadística*, vol. 10, no. 1, pp. 113–128, 2017.

[21] K. Price, R. M. Storn, and J. A. Lampinen, *Differential evolution: a practical approach to global optimization*. Springer Science & Business Media, 2006.

[22] A. E. Eiben and J. E. Smith, *Introduction to evolutionary computing*. Springer, 2003, vol. 53.

[23] C. C. Coello, G. B. Lamont, and D. A. Van Veldhuizen, Evolutionary algorithms for solving multi-objective problems. Springer Science & Business Media, 2007.

[24] K. C. Tan, E. F. Khor, and T. H. Lee, *Multiobjective evolutionary algorithms and applications*. Springer Science & Business Media, 2006.

[25] T. Back, *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press, 1996.

[26] E. Willighagen, M. Ballings, and M. M. Ballings, "Package genalg," 2015.

[27] K. M. Mullen, D. Ardia, D. L. Gil, D. Windover, and J. Cline, "Deoptim: An r package for global optimization by differential evolution," 2009.

[28] D. Eddelbuette, B. Peterson, J. Ulrich, and Eddelbuette, "Rcppde: extending deoptim," 2016.

[29] S. 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.