Revista Colombiana de Estadística - Applied Statistics
July 2023, volume 46, issue 2, pp. 121–143
http://doi.org/10.15446/rce.v46n2.104019

An Adaptive Method for Likelihood Optimization in Linear Mixed Models Under Constrained Search Spaces

Un método adaptativo para optimizar la función de verosimilitud en modelos lineales mixtos bajo espacios de búsqueda restringidos

Mauricio A. Mazo-Lopera[a], Juan C. Salazar-Uribe[b], Juan C. Correa-Morales[c]

School of Statistics, Faculty of Sciences, Universidad Nacional de Colombia, Medellín, Colombia

Abstract

Linear mixed effects models are highly flexible in handling correlated data by considering covariance matrices that explain variation patterns between and within clusters. For these covariance matrices, there exist a wide list of possible structures proposed by researchers in multiple scientific areas. Maximum likelihood is the most common estimation method in linear mixed models and it depends on the structured covariance matrices for random effects and errors. Classical methods used to optimize the likelihood function, such as Newton-Raphson or Fisher's scoring, require analytical procedures to obtain parametrical restrictions to guarantee positive definiteness for the structured matrices and it is not, in general, an easy task. To avoid dealing with complex restrictions, we propose an adaptive method that incorporates the so-called *Hybrid Genetic Algorithms* with a penalization technique based on minimum eigenvalues to guarantee positive definiteness in an evolutionary process which discards non-viable cases. The proposed method is evaluated through simulations and its performance is compared with that of Newton-Raphson algorithm implemented in SAS® `PROC MIXED` V9.4.

***Key words*:** Hybrid genetic algorithm; Linear mixed model; Optimization; Positive definite matrices.

[a]Ph.D. E-mail: mamazol@unal.edu.co
[b]Ph.D. E-mail: jcsalaza@unal.edu.co
[c]Ph.D. E-mail: jccorrea@unal.edu.co

Resumen

Los modelos lineales mixtos son muy flexibles cuando se trabaja con datos correlacionados ya que estos consideran matrices de covarianza que explican los patrones de variación entre individuos y dentro de sus observaciones. Para estas matrices de covarianza existe una amplia lista de posibles estructuras propuestas por investigadores en múltiples áreas científicas. El método de máxima verosimilitud es el más común para la estimación de los parámetros en modelos lineales mixtos y depende de las matrices de covarianza estructuradas para efectos aleatorios y errores. Los métodos clásicos utilizados para optimizar la función de verosimilitud, como Newton-Raphson o Fisher's scoring, requieren desarrollos analíticos para obtener restricciones sobre los parámetros que garanticen matrices estructuradas y definidas positivas, y en general, esto no es una tarea fácil. Para evitar lidiar con restricciones complejas, proponemos un método adaptativo que incorpora los llamados *Algoritmos Genéticos Híbridos* con una técnica de penalización basada en valores propios mínimos con el fin de garantizar matrices positivas definidas en un proceso evolutivo que descarta casos no viables. El método propuesto se evalúa a través de simulaciones y se compara su desempeño con el algoritmo de Newton-Raphson implementado en SAS® `PROC MIXED` V9.4.

***Palabras clave*:** Algoritmo genético híbrido; Modelo lineal mixto; Optimización; Matrices definidas positivas.

1. Introduction

Linear mixed models (LMM) are an extension of simple linear models to take into account fixed and random effects and are widely applied when there is dependency between observations. The most common case of this dependency appears when several measures are taken for the same individual inducing a within-subjects variation which is included in the LMM through an error term. Between-subjects variation is represented in the LMM by means of a random effect term. The theoretical assumption for these two sources of variation is that they are multivariate independent and normally distributed random variables with zero-mean vectors and structured covariance matrices. Laird & Ware (1982) proposed both *Maximum Likelihood* (ML) and *Restricted Maximum Likelihood* (REML) to estimate the parameters in these covariance matrices and they claim that these methods are quite sensitive to the selection of the covariance structures.

For independent data, the covariance structure is a multiple of the identity and there is only one parameter to be estimated. A simple way to extend this structure for taking into account the dependence on the data is to assume homogeneous covariances which implies including an additional parameter to be estimated. However, this assumption may not be realistic because in general, the correlations are dynamic, as for example in longitudinal data.

Conversely, an unstructured covariance matrix, where all parameters are assumed to be different, could lead to overparameterization problems. Therefore, it is desirable to select an intermediate structure to describe the variation patterns and, at the same time, to reduce the number of parameters to be estimated

(Pinheiro, 1994).

Once the structures are selected, the optimization process of the likelihood function must ensure that these structures are maintained in addition to the positive-definiteness condition.

Some convergence problems arise in the optimization process of the likelihood function when using algorithms based on derivatives. Verbeke & Molenberghs (1997), for instance, identify positive-definiteness problems with respect to the estimated covariance matrices when the Newton-Raphson algorithm is used. This is because these matrices are structured and some parametric restrictions are required to ensure that they are always positive definite (PD) during the optimization process, which is not an easy task. The `PROC MIXED` of `SAS`®, for example, imposes that diagonal elements in the covariance matrices are nonnegative, but this is not a sufficient condition and Verbeke & Molenberghs (1997) recommend that one should check the obtained maximum likelihood estimators for the covariance matrices to verify positive-definiteness.

To avoid dealing with these restrictions in the function domain, we propose an optimization method, for the ML and REML functions, based on a penalty technique and we use hybrid genetic algorithms (HGA) as optimizer (Scrucca, 2017). The main advantage of this proposed method is that it penalizes the parameters in the space of all feasible solutions or simply *search space* (see subsection 3.1 and Appendix B for more details) that leads to non-PD covariance matrices and also it combines the power of genetic algorithms as a global optimization technique with the speed of a local optimizer such as Newton-Raphson. Mebane & Sekhon (2011) implemented this method to solve difficult optimization problems when the function to be optimized is discontinuous or nonlinear in their parameters.

This paper is organized as follows. In Section 2, we describe the theoretical aspects of linear mixed models as well as the likelihood functions. Conditions for the covariance matrices are exposed in Section 3. The proposed optimization method of the likelihood functions is presented in Section 4. Section 5 includes some simulations for different structures. In Section 6 the conclusions are given. In Appendix A there is an additional proof and a general description for the Newton-Raphson algorithm. Finally, Appendix B gives some general aspects of hybrid genetic algorithms.

2. Linear Mixed Models

Let us assume we have a random sample of N independent subjects from a target population, where the i-th member has n_i observations. Let Y_{ij} denotes the j-th ($j = 1, \ldots, n_i$) continuous response from the i-th ($i = 1, \ldots, N$) subject. By denoting $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, the individual-LMM developed by Laird & Ware (1982), is given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \text{ for } i = 1, 2, \ldots, N, \tag{1}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are the design matrices for both fixed and random effects, $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$, respectively, and $\boldsymbol{\epsilon}_i$ is an error vector. In addition, let us assume $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are

independent random vectors with $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$.

The covariance matrix of $\boldsymbol{Y}_i$, denoted by $\boldsymbol{V}_i = Var(\boldsymbol{Y}_i)$, is given by

$$\boldsymbol{V}_i \quad = \quad \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \boldsymbol{\Sigma}_i. \tag{2}$$

By denoting $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_N')'$, the design matrix by $\boldsymbol{X} = [\boldsymbol{X}_1 \ \ldots \ \boldsymbol{X}_N]'$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \ldots, \boldsymbol{\epsilon}_N')'$, $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_N')'$, $\boldsymbol{G} = \text{diag}\{\boldsymbol{D}, \ldots, \boldsymbol{D}\}$, $\boldsymbol{\Sigma} = \text{diag}\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N\}$, and $\boldsymbol{Z} = \text{diag}\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N\}$, the general LMM can be written as

$$\boldsymbol{Y} \quad = \quad \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are independent, $\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and from (2)

$$\boldsymbol{V} \quad = \quad \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{\Sigma}. \tag{4}$$

2.1. Log-Likelihood Function

The log-likelihood function for model (3) is given by

$$\log L(\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\Sigma} | \boldsymbol{Y}) \quad = \quad -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{N}{2}\log(2\pi).$$

To simplify the optimization process of the previous log-likelihood function, Henderson (1984) proposes to assume that $\boldsymbol{V}$ is known, derivate with respect to $\boldsymbol{\beta}$ and equating by zero to obtain

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{5}$$

Then, the matrices $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ can be estimated by minimizing the function

$$\ell\left(\boldsymbol{G}, \boldsymbol{\Sigma} | \boldsymbol{Y}, \widehat{\boldsymbol{\beta}}\right) = -2\log L(\boldsymbol{G}, \boldsymbol{\Sigma} | \boldsymbol{Y}) = \log|\boldsymbol{V}| + \boldsymbol{r}'\boldsymbol{V}^{-1}\boldsymbol{r} + N\log(2\pi). \tag{6}$$

where $|\cdot|$ denotes the determinant function, and

$$\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{7}$$

To correct the bias from estimating $\boldsymbol{\beta}$, Patterson & Thompson (1971) proposed to minimize the REML function

$$\ell_R\left(\boldsymbol{G}, \boldsymbol{\Sigma} | \boldsymbol{Y}, \widehat{\boldsymbol{\beta}}\right) = \log|\boldsymbol{V}| + \boldsymbol{r}'\boldsymbol{V}^{-1}\boldsymbol{r} + \log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| + (N - p)\log(2\pi), \tag{8}$$

where $\boldsymbol{r}$ is given in (7) and p is the rank of $\boldsymbol{X}$.

To save notation in the rest of this work, we will denote $\ell(\boldsymbol{G}, \boldsymbol{\Sigma})$ and $\ell_R(\boldsymbol{G}, \boldsymbol{\Sigma})$ or simply ℓ and ℓ_R, instead of $\ell\left(\boldsymbol{G}, \boldsymbol{\Sigma} | \boldsymbol{Y}, \widehat{\boldsymbol{\beta}}\right)$ and $\ell_R\left(\boldsymbol{G}, \boldsymbol{\Sigma} | \boldsymbol{Y}, \widehat{\boldsymbol{\beta}}\right)$, respectively.

TABLE 1: Examples of covariance structures.

Name	**Abbreviation**	**Example** (3×3)
Compound Symmetry	CS	$\begin{pmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{pmatrix}$
First-Order Autoreg. Moving Average	ARMA(1, 1)	$\sigma^2 \begin{pmatrix} 1 & \gamma & \gamma\rho \\ \gamma & 1 & \gamma \\ \gamma\rho & \gamma & 1 \end{pmatrix}$
Toeplitz	TOEP	$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}$

Source: SAS Institute Inc. (2008)

3. Covariance Structures for G and Σ

Since $\boldsymbol{G} = \text{diag}\{\boldsymbol{D}, \boldsymbol{D}, \ldots, \boldsymbol{D}\}$ and $\boldsymbol{\Sigma} = \text{diag}\{\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_N\}$ are block diagonal matrices, conditions for them depend on conditions for $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_N$, respectively. For example, symmetry and positive definiteness can be ensured by assuming that their diagonals matrices are symmetric and PD. In addition, covariance structures can be assigned to $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ through $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_i$. Regarding to these structures, different authors have proposed several structures for covariance matrices to model variability patterns between correlated observations (Wolfinger, 1993). `PROC MIXED` of SAS® (SAS Institute Inc., 2008) includes an extensive list of structures providing a detailed description with respect to the number of parameters and variability patterns. Some useful structures in practice include Compound Symmetry (CS), First-Order Autoregressive Moving Average (ARMA(1, 1)), Toeplitz (TOEP) and Multiple of the identity (MI) given by $\sigma^2\boldsymbol{I}$ (where $\boldsymbol{I}$ is the identity matrix). Table 1 shows examples of (3×3) dimensions for these structures.

3.1. Naive Search Spaces for the Covariance Parameters and Positive-Definiteness

The covariance structures depend on a set of parameters with certain particular restrictions. A "*naive*" approximation for the parameters search space is the cross product of the intervals that result after considering these particular restrictions. In this context, "*naive*" means that this first approximation does not take into account the positive definiteness condition of the covariance matrices. Let Ω denotes the *naive* search space for the covariance parameters. As an example of Ω, consider the CS structure, which depends on $\boldsymbol{\theta} = (\sigma, \sigma_1)'$ and the particular restrictions for these parameters are $\sigma > 0$ and $\sigma^2 + \sigma_1 > 0$, so the *naive* search space for $\boldsymbol{\theta}$ is

$\Omega = (0, \infty) \times \left(-\sigma^2, \infty\right)$. However, CS is not PD for all parameter vectors in Ω. To see this, let us consider the CS-(3×3) matrix with $(\sigma, \sigma_1) = (1.2, -1)$. It can be seen that $\sigma = 1.2 > 0$ and $\sigma^2 + \sigma_1 = 0.44 > 0$, but the eigenvalues of this matrix are equal to 1.44, 1.44 and -1.56, respectively. The Cholesky decomposition is a method to guarantee PD matrices (Pinheiro & Bates, 1996), but it not an easy task to find the parameterization when the matrix is not unstructured. To see this, let us consider the Cholesky decomposition for the CS-(3×3) structure,

$$\begin{pmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ 0 & a_{22} & a_{32} \\ 0 & 0 & a_{33} \end{pmatrix}.$$

It can be proved that the solutions are given by $a_{11} = \sqrt{\sigma^2 + \sigma_1}$, $a_{21} = a_{31} = \sigma_1/\sqrt{\sigma^2 + \sigma_1}$, $a_{22} = \sqrt{\sigma^4 + 2\sigma^2\sigma_1}/\sqrt{\sigma^2 + \sigma_1}$, $a_{32} = \sigma_1\sigma/\left(\sqrt{\sigma^2 + \sigma_1}\sqrt{\sigma^2 + 2\sigma_1}\right)$ and $a_{33} = \sigma\sqrt{\sigma^2 + 3\sigma_1}/\sqrt{\sigma^2 + 2\sigma_1}$, with the restrictions $\sigma^2 + \sigma_1 > 0$, $\sigma^2 + 2\sigma_1 > 0$ and $\sigma^2 + 3\sigma_1 > 0$. The intersection of these restrictions is $\sigma_1 > -\sigma^2/3$ and therefore the space where the CS-(3×3) structure is PD is given by $\boldsymbol{\theta} \in (0, \infty) \times (-\sigma^2/3, \infty)$. Figure 1 (a) shows this PD region (vertical lines) compared with the *naive* search space Ω (horizontal lines) [1].

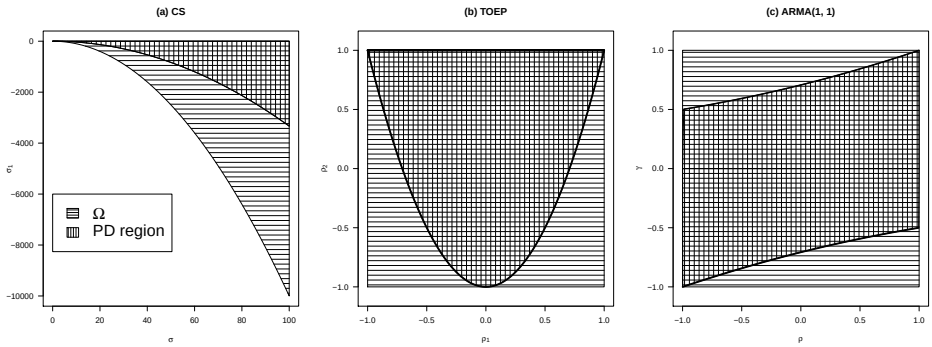

FIGURE 1: *Naive* search space Ω (horizontal lines) and PD region (vertical lines) for the (3×3) matrices, considering the structures: (a) CS, (b) TOEP and (c) ARMA(1, 1).
Source: Author.

The ARMA(1, 1) depends on $\boldsymbol{\theta} = (\sigma, \rho, \gamma)'$ with $\sigma > 0$, $|\rho| < 1$ and $|\gamma| < 1$ for stationarity and invertibility, respectively (Box et al., 1970).Therefore, the *naive* search space is $\Omega = (0, \infty) \times (-1, 1) \times (-1, 1)$. Following the method given by Madar (2015), we obtained the PD region (for all $\sigma > 0$) of Figure 1 (c) for the ARMA(1, 1)-(3×3) in contrast to the *naive* search space Ω. The border curves were obtained with the relationship $\rho = (2\gamma^2 - 1)/\gamma$. This condition can be extended for higher dimensions using the method of Madar (2015).

The AR(1) structure (a particular case of ARMA(1, 1) with $\gamma = \rho$) depends on $\boldsymbol{\theta} = (\sigma, \rho)'$, with particular restrictions $\sigma > 0$ and $|\rho| < 1$ for stationarity

[1] We selected $0 < \sigma < 100$ to obtain the figure.

(Box et al., 1970). Madar (2015) showed that the elements of the lower triangular matrix, $\boldsymbol{L}$, of the Cholesky decomposition $\boldsymbol{LL}'$ for the AR(1) covariance structure are given by

$$l_{ji} = \begin{cases} \rho^{j-1}, & j \geq i = 1 \\ \rho^{j-1}\sqrt{1-\rho^2}, & j \geq i \geq 2, \end{cases}$$

and therefore the condition for positive definiteness of AR(1) structure is $|\rho| < 1$, which matches with the stationarity condition. So, the *naive* search space is $\Omega = (0, \infty) \times (-1, 1)$ and the AR(1) structure is PD for all parameter vectors in this space.

The TOEP-(3×3) structure depends on $\boldsymbol{\theta} = (\sigma, \sigma_1, \sigma_2)'$, with standard deviation $\sigma > 0$ and covariances $-\infty < \sigma_1, \sigma_2 < \infty$. This structure can be rewritten as

$$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix} = \begin{pmatrix} \sigma & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & \sigma \end{pmatrix} \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{pmatrix} \begin{pmatrix} \sigma & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & \sigma \end{pmatrix}, \tag{9}$$

where ρ_1, ρ_2 are correlations. So, the new restrictions are $\sigma > 0$ and $-1 < \rho_1, \rho_2 < 1$ and the *naive* search space is $\Omega = (0, \infty) \times (-1, 1) \times (-1, 1)$. However, the correlation matrix in (9) is not PD for all $(\sigma, \rho_1, \rho_2) \in (0, \infty) \times (-1, 1) \times (-1, 1)$. To see this, let us consider the following Cholesky decomposition

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ 0 & a_{22} & a_{32} \\ 0 & 0 & a_{33} \end{pmatrix}.$$

It can be proved that solutions for the lower triangular matrix are $a_{11} = 1$, $a_{21} = \rho_1$, $a_{31} = \rho_2$, $a_{33} = \sqrt{(1 - \rho_2^2 - 2\rho_1^2 + 2\rho_1^2\rho_2)/(1 - \rho_1^2)}$, $a_{22} = \sqrt{1 - \rho_1^2}$, and $a_{32} = \rho_1(1 - \rho_2)/\sqrt{1 - \rho_1^2}$, with conditions $1 - \rho_1^2 > 0$ and $1 - \rho_2^2 - 2\rho_1^2 + 2\rho_1^2\rho_2 > 0$. Therefore, the restrictions for ensuring positive definiteness for the TOEP-(3×3) matrix in the equation (9), are $\sigma > 0$, $-1 < \rho_1 < 1$ and $\rho_2 > 2\rho_1^2 - 1$. Figure 1 (b) shows this PD region (vertical lines) compared with the *naive* search space Ω (horizontal lines). This regions are given for all $\sigma > 0$. For higher matricial dimensions, Madar (2015) describes the structure for the lower triangular matrix of the Cholesky decomposition and this can be used to obtain the restrictions that lead to PD TOEP matrices. Finallly, the MI structure is PD for all $\sigma > 0$.

The above mentioned restrictions for positive definiteness show that, in general, to obtain the parametrical restrictions to find the PD region is not an easy task and we propose in Subsection 4.2 to use a penalization method for non-PD matrices based on HGAs to optimize the ML and REML functions in the LMM. This method uses the *naive* search space Ω which is easy to obtain in contrast to the region of positive definiteness.

4. Proposed Optimization Method of the Likelihood Function

One of the most common iterative algorithms used to optimize the log-likelihood functions (6) and (8) is the so called Newton-Raphson (NR). This methods is based on derivatives and for each iteration it is necessary to ensure positive definiteness for matrices $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$. Unfortunately, it is difficult to meet these requirements and some software procedures establish simpler conditions that are necessary, but not sufficient, to ensure positive definiteness constraints (West et al., 2006). For example, the `PROC MIXED` of SAS® restricts the diagonal elements of $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ to be positive in the NR algorithm during the entire iteration process but this is not sufficient to ensure positive definiteness of covariance matrices and some warning messages may appear during the optimization process. Other optimization methods are Fisher's Scoring which is based on the expected information matrix instead of the observed information matrix (Wolfinger et al., 1994) and Expectation Maximization (EM) algorithm proposed by Dempster et al. (1977) which is mainly used to find starting values for other algorithms or to optimize complicated likelihood functions. Lindstrom & Bates (1988) provide reasons for preferring NR to the EM based on time consuming and consistent convergence. In this section, we propose a heuristic method based on HGAs (see Appendix B) to optimize (6) and (8) by designing penalty functions to ensure that matrices $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ are PD. We also describe the NR algorithm in the context of LMM as well as the theoretical framework of the proposed method.

4.1. Newton-Raphson Algorithm

Let us suppose that matrices $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ depend on the parameter vectors $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_\Sigma$, respectively, and denote this dependence by $\boldsymbol{G}(\boldsymbol{\theta}_G)$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_\Sigma)$. So, the matrix $\boldsymbol{V}$ depends on the parameter vector $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_G', \boldsymbol{\theta}_\Sigma'\right)'$ and from equation (5), $\widehat{\boldsymbol{\beta}}$ also depends on $\boldsymbol{\theta}$. NR algorithm follows the general steps (Demidenko, 2004):

1. With an initial parameter vector $\boldsymbol{\theta}_0$ compute the gradient vector $\boldsymbol{g}$ and the Hessian matrix $\boldsymbol{H}$.
2. Let $\lambda = 1$.
3. Compute new estimates for $\boldsymbol{\theta}$, iteratively, using $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} - \lambda\left(\boldsymbol{H}^{-1}\boldsymbol{g}\right)$.
4. If $\boldsymbol{\theta}_i$ is a valid vector of covariance parameters and improves the likelihood, continue to Step (5). Otherwise, reduce λ by half and return to Step (3).
5. Check for convergence and stop if some convergence criterion is reached. Otherwise, compute $\boldsymbol{g}$ and $\boldsymbol{H}$ with the new estimates for $\boldsymbol{\theta}$ and go back to Step (2).

Wolfinger et al. (1994) recommended the Minimum Variance Quadratic Unbiased Estimators (MIVQUE) method proposed by Rao (1972) for computing the

initial parameter vector, $\boldsymbol{\theta}_0$. The gradient vector and the Hessian matrix for both ML and REML are described in Appendix A.2. Since they depend on matrix $\boldsymbol{V}$, it is clear that for each iteration, NR method requires to ensure positive definiteness for matrices $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$. A partial solution to reach this condition is to restrict the diagonal elements of $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ to be positive. Another alternative is to apply Cholesky decomposition (or other parameterizations) to these matrices to simplify the optimization problem. Pinheiro & Bates (1996) described some parametrizations that enforce the positive definiteness and compared their computational efficiency and statistical interpretability using ML and REML estimation. Those parameterizations are easy to use when covariance matrices $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ are not structured, that is, when it is assumed that the singular variances and covariances do not follow a pattern. However, the general assumption is that $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ are structured.

4.2. Proposed Method

Let us suppose that $\boldsymbol{G}(\boldsymbol{\theta}_G)$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_\Sigma)$ are continuous matricial functions and that the applications

$$\begin{aligned} T_G : \Omega_G &\to \mathcal{A}_G \\ \boldsymbol{\theta}_G &\mapsto \boldsymbol{G}(\boldsymbol{\theta}_G) \end{aligned} \qquad \text{and} \qquad \begin{aligned} T_\Sigma : \Omega_\Sigma &\to \mathcal{A}_\Sigma \\ \boldsymbol{\theta}_\Sigma &\mapsto \boldsymbol{\Sigma}(\boldsymbol{\theta}_\Sigma) \end{aligned}$$

are continuous, where Ω_G and Ω_Σ are Euclidean subspaces and $\mathcal{A}_G$, $\mathcal{A}_\Sigma$ are the sets of symmetric matrices with the corresponding dimensions.

Since $\boldsymbol{G}(\boldsymbol{\theta}_G)$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_\Sigma)$ are symmetric matrices, their eigenvalues are real numbers (see Schott (1997), Theorem 3.9 page 106 for details) and therefore, we propose to calculate their minimum eigenvalues, $\lambda_{\min}$, and compare them with zero for checking positive definiteness. Let us define

$$\begin{aligned} \Lambda_G : \mathcal{A}_G &\to \mathbb{R} \\ \boldsymbol{G} &\mapsto \lambda_{\min}(\boldsymbol{G}) \end{aligned} \qquad \text{and} \qquad \begin{aligned} \Lambda_\Sigma : \mathcal{A}_\Sigma &\to \mathbb{R} \\ \boldsymbol{\Sigma} &\mapsto \lambda_{\min}(\boldsymbol{\Sigma}) \end{aligned}$$

where $\lambda_{\min}(\boldsymbol{G})$ and $\lambda_{\min}(\boldsymbol{\Sigma})$ are the minimum eigenvalues of $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$, respectively. Since Λ_G and Λ_Σ are continuous (see Schott (1997), Theorem 3.14 page 115 for details), the composite functions,

$$\begin{aligned} \Lambda_G \circ T_G : \Omega_G &\to \mathbb{R} \\ \boldsymbol{\theta}_G &\mapsto \lambda_{\min}[\boldsymbol{G}(\boldsymbol{\theta}_G)] \end{aligned} \qquad \text{and} \qquad \begin{aligned} \Lambda_\Sigma \circ T_\Sigma : \Omega_\Sigma &\to \mathbb{R} \\ \boldsymbol{\theta}_\Sigma &\mapsto \lambda_{\min}[\boldsymbol{\Sigma}(\boldsymbol{\theta}_\Sigma)] \end{aligned}$$

are also continuous (see Munkres (2000), Theorem 18.2 page 107). So, if we have that $\lambda_{\min}[\boldsymbol{G}(\boldsymbol{\theta}_G^0)] > 0$, for some $\boldsymbol{\theta}_G^0 \in \Omega_G$, there is an open interval $I_{\boldsymbol{\theta}_G^0} \in \mathbb{R}^+$ and, by definition of continuity functions, $(\Lambda_G \circ T_G)^{-1}(I_{\boldsymbol{\theta}_G^0}) \in \Omega_G$ is also open, i.e., there is an open set around $\boldsymbol{\theta}_G^0$ for which all vectors in it yield PD matrices. An equivalent result is given for $(\Lambda_\Sigma \circ T_\Sigma)$. As we will see later, these results are important to guarantee that local optimization for HGAs is possible.

4.2.1. Penalized Fitness Function

To guarantee positive definiteness as well as to preserve the specific structures for $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$, we propose to optimize the functions (6) and (8) by applying HGAs and penalizing the parameters vectors, $\boldsymbol{\theta_G}$ and $\boldsymbol{\theta_\Sigma}$, when they do not lead to PD matrices. Coello (2002), Kuri-Morales & Gutiérrez-García (2002), Lin (2013) and Chehouri et al. (2016) explored different penalization methods with GAs in the context of constrained optimization and we are using an equivalent idea to ensure the positive definiteness of the covariance matrices.

Before describing those penalizations, note that both functions ℓ and ℓ_R include the logarithm of the determinants $|\boldsymbol{V}|$ and $|\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}|$, so we take the absolute value to these terms to guarantee their existence for all $(\boldsymbol{\theta_G}, \boldsymbol{\theta_\Sigma}) \in \Omega_{\boldsymbol{G}} \times \Omega_{\boldsymbol{\Sigma}}$. So, instead of ℓ and ℓ_R let us consider the modified functions

$$\ell^*(\boldsymbol{G}, \boldsymbol{\Sigma}) = \log[\text{abs}(|\boldsymbol{V}|)] + \boldsymbol{r}'\boldsymbol{V}^{-1}\boldsymbol{r} + N\log(2\pi) \tag{10}$$

and

$$\ell_R^*(\boldsymbol{G}, \boldsymbol{\Sigma}) = \log[\text{abs}(|\boldsymbol{V}|)] + \boldsymbol{r}'\boldsymbol{V}^{-1}\boldsymbol{r} + \log[\text{abs}(|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|)] + (N-p)\log(2\pi), \tag{11}$$

where abs$(\cdot)$ is the absolute value function. The purpose of defining the functions ℓ^* and ℓ_R^* is to evaluate and identify the "undesired" cases when $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ are non-PD. However, for PD matrices this functions are equivalent to the functions given in (6) and (8), respectively. Therefore, this method does not induce bias in the maximum likelihood estimators. Our strategy consists in including a "naive" search space (as explained in subsection 3.1), even when the covariance matrices are not PD and then penalizing those cases by using the following penalized fitness function (for REML use ℓ_R^* instead of ℓ):

$$f_{\epsilon,K}(\boldsymbol{\theta_G}, \boldsymbol{\theta_\Sigma}) = \begin{cases} \ell^*[\boldsymbol{G}(\boldsymbol{\theta}_G) + \epsilon\boldsymbol{I}_G, \boldsymbol{\Sigma}(\boldsymbol{\theta}_\Sigma) + \epsilon\boldsymbol{I}_\Sigma], & \text{if } m_{\min} > 0 \\ \text{abs}(\ell^*[\boldsymbol{G}(\boldsymbol{\theta}_G) + \epsilon\boldsymbol{I}_G, \boldsymbol{\Sigma}(\boldsymbol{\theta}_\Sigma) + \epsilon\boldsymbol{I}_\Sigma]) + K, & \text{otherwise} \end{cases} \tag{12}$$

where $m_{\min} = \min\{\Lambda_G \circ T_G(\boldsymbol{\theta}_G), \Lambda_\Sigma \circ T_\Sigma(\boldsymbol{\theta}_\Sigma)\}$ and $K > 0$ is a penalty term. Kuri-Morales & Gutiérrez-García (2002) provide several methods for selecting this penalty term and the simplest one is to define K as a large constant ($O\left(10^9\right)$). We propose a dynamic selection for K by taking into account the number of digits in the integer part of either ℓ^* or ℓ_R^*. Denote this integer value by *ndigs* and take $K = 10^{ndigs+9}$. We selected this penalization term because it has a simple structure and it also provides a high penalty for the parameter vectors that do not lead to PD matrices. However, as mentioned above, there are other penalization methods in the literature (Kuri-Morales & Gutiérrez-García, 2002), but we decided to use K for its simple form and because it is dynamic by giving a high penalization depending on the order of magnitude of the likelihood function. As an example, suppose that ℓ_R is equal to -345.1982 for some $(\boldsymbol{\theta}_G, \boldsymbol{\theta}_\Sigma) \in \Omega_G \times \Omega_\Sigma$, with $\min\{\Lambda_G \circ T_G(\boldsymbol{\theta}_G), \Lambda_\Sigma \circ T_\Sigma(\boldsymbol{\theta}_\Sigma)\} \leq 0$, so the integer part of this number is -345 ($ndigs = 3$) and therefore $K = 10^{3+9} = 10^{12}$. Thus, the penalized value is $\text{abs}(-365.1982) + 10^{12} = 365.1982 + 10^{12}$.

The matrices $\epsilon \boldsymbol{I}_G$ and $\epsilon \boldsymbol{I}_\Sigma$ were included in the function f to avoid the estimation of $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ to be computationally singular in the optimization process (for all $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_\Sigma$ in the search space), where $\boldsymbol{I}_G$ and $\boldsymbol{I}_\Sigma$ are the identity matrices with the corresponding dimensions of $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$, respectively. The value $\epsilon > 0$ is used as a pivot and we choice $\epsilon = 10^{-10}$. This selection is based on the tolerance or limit value in the determinant to declare a matrix as singular, the software `R` version 4.1.0, for example, considers a tolerance with order $O(10^{-16})$.

The function (12) is defined such that the parameter vectors that lead to non-PD covariance matrices have always the highest values in the minimization process. This implies that the HGA discard those cases in a evolutionary way because they are the worst candidates for optimizing the log-likelihood function compared to the parameter vectors that lead to PD matrices.

4.2.2. Minimizing the Penalized Fitness Function

Given specific structures for $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$, we propose to apply the HGAs to the function f to obtain local optimum candidates and also carrying out local optimization iteratively until it reaches an optimum. This local optimization could be done with Newton-Raphson or another optimization method based on derivatives. An overview of the HGAs is available in Appendix B.

Since functions $T_G \circ \Lambda_G(\boldsymbol{\theta}_G)$ and $T_\Sigma \circ \Lambda_\Sigma(\boldsymbol{\theta}_\Sigma)$ are continuous for all $\boldsymbol{\theta}_G \in \Omega_G$ and $\boldsymbol{\theta}_\Sigma \in \Omega_\Sigma$, the function $\min\{T_G \circ \Lambda_G(\boldsymbol{\theta}_G), T_\Sigma \circ \Lambda_\Sigma(\boldsymbol{\theta}_\Sigma)\}$ is also continuous, for all $(\boldsymbol{\theta}_G, \boldsymbol{\theta}_\Sigma) \in \Omega_G \times \Omega_\Sigma$. This means that if the GA finds a couple $(\boldsymbol{\theta}_{G_1}, \boldsymbol{\theta}_{\Sigma_1})$ which produces PD matrices $\boldsymbol{G}_1$ and $\boldsymbol{\Sigma}_1$, it is possible to conduct a local optimization search around it. This is because the log-likelihood functions ℓ^* and ℓ_R^* are differentiable for all PD matrices $\boldsymbol{G}$, $\boldsymbol{\Sigma}$ and, as we discussed before, the continuity of $\min\{T_G \circ \Lambda_G(\boldsymbol{\theta}_G), T_\Sigma \circ \Lambda_\Sigma(\boldsymbol{\theta}_\Sigma)\}$ implies that there exists an open set $C \subseteq \Omega_G \times \Omega_\Sigma$, with $(\boldsymbol{\theta}_{G_1}, \boldsymbol{\theta}_{\Sigma_1}) \in C$, such that for all $(\boldsymbol{\theta}_G, \boldsymbol{\theta}_\Sigma) \in C$, their corresponding matrices are PD.

4.2.3. Algorithm

Given the fitness function f, HGA (see Appendix B for details) is applied according to the following pseudocode:

(1) Initialize population $P(0)$ of size n_{pop} in the search space Ω;
(2) set $g = 0$;
(3) **for** $i = 1$ to n_{iter};
(4) evaluate fitness (f) of population $P(g)$;
(5) select the best-fit individuals from $P(g)$ to produce $P'(g)$, according to a probability $p_{elitism}$;
(6) encode the individuals of $P'(g)$;
(7) apply the crossover method to produce the population $P''(g)$ from $P'(g)$;
(8) perform the mutation method on $P''(g)$ to produce the decoded $P'''(g)$;
(9) select the best-fit individual of $P'''(g)$;
(10) according to a probability p_{local}, take the decision of carrying out a local optimization;

(11) **if** the decision of local optimization is affirmative **then**;
(12) carry out local optimization with the best-fit individual of $P'''(g)$ as initial value;
(13) replace best-fit individual of $P'''(g)$ with the obtained local optimum to produce the next generation $P(g+1)$;
(14) **else**;
(15) $P(g+1) = P'''(g)$;
(16) **end if**;
(17) $g = g+1$;
(18) **end for**;

where g is the iteration number, $P(g)$ is the initial population at iteration g, $P'(g)$ is the resulting population after best-fit individuals selection at iteration g, $P''(g)$ is the resulting population after crossover at iteration g and $P'''(g)$ is the resulting population after mutation at iteration g.

5. Simulation Study

A simulation study were conducted to evaluate the proposed methodology for optimizing the likelihood functions in equations (6) and (8) given in Subsection 4.2. The proposed methodologies based on HGAs were compared with the Newton-Raphson algorithm incorporated in `PROC MIXED` of `SAS`® version 9.4. For the HGAs, we used the package "GA" (Scrucca, 2017), version 3.2.2 of the sofware `R`, version 4.1.0. The simulations were run on Windows 10, Intel(R) Core(TM) i5-4590 3.30 GHz and 8.00 GB of RAM.

To carry out the simulation study, we use a dental growth measurements data for 11 girls and 16 boys at ages 8, 10, 12 and 14 from Pothoff & Roy (1964) to obtain the reference parameters in the simulations. Let us denote Y_{ij} the dental growth measurement for i-th subject ($i = 1, \ldots, N$) at jth time ($j = 1, \ldots, s$). We consider the model:

$$Y_{ij} = \beta_0 + \beta_1 Age_{ij} + \beta_2 Gender_i + \beta_3 (Age_{ij} * Gender_i) + b_{0i} + b_{1i} Age_{ij} + \varepsilon_{ij}, \quad (13)$$

where $Age_{ij} * Gender_i$ denotes the interaction between age and gender and $\boldsymbol{\epsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{is})^T$ are independent random vectors following a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_\bullet$, for all i. The random effects vector $\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$ in the model (13) is independent of $\boldsymbol{\epsilon}_i$ and follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$.

To create different simulation scenarios, we consider two sample sizes: $N = 27$ (which is the sample size of the original dataset from Pothoff & Roy (1964)), and $N = 54$ (twice the sample size in the original dataset). The repeated measures size was $s = 4$, which is the number of repeated measures of the original dataset.

We also consider two combinations of covariance structures: AR(1) for $\boldsymbol{\Sigma}_\bullet$ with Toeplitz for $\boldsymbol{D}$ (AR(1)-TOEP) and Multiple of the identity for $\boldsymbol{\Sigma}_\bullet$ with Toeplitz for $\boldsymbol{D}$ (MI-TOEP). We propose a TOEP (2×2) for $\boldsymbol{D}$ because we are interested in evaluate our proposed method with a structured matrix with equal variances. We

analyze ML and REML methods for each simulation case and therefore, we have a total of 8 simulation scenarios. When the sample size is $N = 54$, we maintain the same proportion of boys and girls from the original data set, i.e., 22 girls and 32 boys.

For each scenario, we carry out 500 simulations and the results were summarized using Means, Medians and Root-Mean-Square-Error (RMSE) for the proposed method versus Newton-Raphson algorithm of `PROC MIXED` of SAS® version 9.4, respectively. The real parameters used in all simulation scenarios were obtained using the dataset from Pothoff & Roy (1964).

5.1. Calibration of HGAs Tuning Parameters

Before carrying out the simulations it is necessary "to calibrate" the HGAs tuning parameters (see Appendix B). To do this, we selected a simulated dataset for each scenario and then we applied the HGAs varying a set of tuning parameters until reaching "stability" in the optimal value of the ML and REML functions. In this context, the "stability" means that for each one of the 40 simulated datasets, we apply several times (we choose 100 times) the HGA with a set of tuning parameters and all the optimal values obtained are equal. The tuning parameters obtained after this calibration process are:

- *Number of iterations:* $n_{iter} = 250$.
- *Population size:* $n_{pop} = 50$ (by default in "ga" function of `R`, version 4.1.0, GA package).
- *Crossover probability:* $p_{cross} = 0.80$ (by default in "ga" function of `R`, version version 4.1.0, GA package).
- *Mutation probability:* $p_{muta} = 0.20$.
- *Local optimization probability:* $p_{local} = 0.05$ (by default in "ga" function of `R`, version version 4.1.0, GA package).
- *Percentage of elitism:* $p_{elitism} = 5\%$ (by default in "ga" function of `R`, version version 4.1.0, GA package).

The search space for the parameters vector $(\sigma_0, \rho_0, \sigma_1, \rho_1)'$ of AR(1)-TOEP structure is $\Omega = (0, 50) \times (-1, 1) \times (0, 50) \times (-1, 1)$ and the search space for the parameters vector $(\sigma_0, \rho_0, \sigma_1)'$ of MI-TOEP combination is $\Omega = (0, 50) \times (-1, 1) \times (0, 50)$. The Toeplitz structures has correlation parameters instead of covariance parameters because we used a matrix decomposition similar to the one given in the equation (9). Note that the search spaces are large enough to contain the real covariance parameters which is the main objetive in the calibration process.

5.2. Results from Both ML and REML Likelihood Functions

The simulation results for structure AR(1)-TOEP in model (13) are in Table 3, and for MI-TOEP structure are in Table 4. There are small differences for the means and medians between the proposed method and Newton-Raphson values. However the RMSE are similar to each other and relatively small. This differences are explained by the results obtained in Figures 2 for the combination AR(1)-TOEP and 3 for MI-TOEP. They show the comparison of functions $-2\log(\text{ML})$ and $-2\log(\text{REML})$ between the proposed method and Newton-Raphson for all different simulated scenarios. Crosses indicate that the estimated covariance matrix for random effects, $\boldsymbol{D}$, is not PD when Newton-Raphson method is applied but with the proposed method it is PD. In these cases, the proposed method reaches higher values than Newton-Raphson, because the latter continues optimizing the log-likelihood function without taking into account the positive-definiteness of matrix $\boldsymbol{D}$. Table 2 shows the percentages of crosses for both structures. The values are higher for MI-TOEP structure and these percentages show that Newton-Raphson is less efficient to reach PD matrices compared with the proposed method.

Table 2: Percentage of non-PD covariance matrices $\boldsymbol{D}$ (% of crosses) obtained with Newton-Raphson in contrast with the proposed method.

Structure	**ML**		**REML**	
	$N = 27$	$N = 54$	$N = 27$	$N = 54$
AR(1)-TOEP	24.8	19.0	22.6	17.2
MI-TOEP	43.8	33.6	38.6	30.4

Source: Author.

The results obtained in this subsection suggest that our method is better to guarantee PD covariance matrices in an evolutionary way compared to Newton-Raphson. Demidenko (2004) exposes this problem when Newton-Raphson is used to optimize the log-likelihood function and emphasizes the importance that a method yields PD covariance matrices.

6. Conclusions

We presented an estimation method for the covariance parameters in linear mixed models using hybrid genetic algorithms as the optimization method of the likelihood function. This method was created to identify the parametrical regions where the structured covariance matrices (involved in the optimization process) were positive definite. To do this, we considered a penalized function based on either maximum likelihood or restricted maximum likelihood procedures and used the minimum eigenvalue function as the identification technique. We penalized the regions where the eigenvalues were negative in such a way that the hybrid genetic algorithms evolved rejecting them and giving a good fitness for parametrical regions where the eigeinvalues for the covariance matrices were greater than zero.

Table 3: Results of Monte Carlo simulations for the proposed method compared with Newton-Raphson and using the AR(1)-TOEP structure.

Maximum likelihood (ML)													
		4-27						4-54					
		Proposed			Newton-Raphson			Proposed			Newton-Raphson		
Parm	Real	Mean	Median	RMSE	Mean	Median	RMSE	Mean	Median	RMSE	Mean	Median	RMSE
σ_0	2.069	1.989	1.985	0.295	1.994	1.988	0.291	2.029	2.015	0.207	2.032	2.016	0.206
ρ_0	0.755	0.778	0.799	0.362	0.725	0.799	0.294	0.779	0.787	0.246	0.757	0.786	0.213
ρ_1	-0.790	-0.789	-0.796	0.061	-0.793	-0.799	0.059	-0.786	-0.790	0.045	-0.788	-0.791	0.044
σ_1	2.284	2.285	2.276	0.245	2.281	2.270	0.244	2.278	2.273	0.172	2.276	2.272	0.172
β_0	16.270	16.286	16.335	0.876	16.286	16.335	0.876	16.251	16.252	0.624	16.251	16.252	0.624
β_1	1.139	1.082	1.099	1.409	1.080	1.105	1.409	1.114	1.137	0.987	1.113	1.129	0.987
β_2	0.797	0.814	0.816	0.542	0.814	0.815	0.542	0.776	0.775	0.370	0.776	0.774	0.369
β_3	-0.321	-0.353	-0.327	0.858	-0.353	-0.325	0.858	-0.340	-0.349	0.567	-0.340	-0.349	0.567
Restricted Maximum likelihood (REML)													
		4-27						4-54					
		Proposed			Newton-Raphson			Proposed			Newton-Raphson		
Parm	Real	Mean	Median	RMSE	Mean	Median	RMSE	Mean	Median	RMSE	Mean	Median	RMSE
σ_0	2.069	2.075	2.070	0.297	2.081	2.079	0.294	2.071	2.059	0.208	2.074	2.060	0.207
ρ_0	0.755	0.750	0.770	0.352	0.706	0.770	0.296	0.766	0.773	0.242	0.746	0.773	0.213
ρ_1	-0.790	-0.785	-0.792	0.063	-0.788	-0.794	0.061	-0.784	-0.788	0.046	-0.785	-0.789	0.045
σ_1	2.284	2.290	2.283	0.245	2.287	2.277	0.244	2.280	2.276	0.172	2.278	2.275	0.172
β_0	16.270	16.286	16.334	0.875	16.286	16.334	0.875	16.251	16.252	0.624	16.251	16.253	0.624
β_1	1.139	1.082	1.098	1.408	1.081	1.102	1.409	1.113	1.141	0.988	1.113	1.135	0.987
β_2	0.797	0.814	0.815	0.542	0.814	0.815	0.542	0.776	0.775	0.370	0.776	0.774	0.370
β_3	-0.321	-0.353	-0.326	0.858	-0.353	-0.326	0.858	-0.340	-0.349	0.567	-0.340	-0.349	0.567

Source: Author.

We used hybrid genetic algorithms among a broad list of evolutionary algorithms because it combines the search power of genetic algorithms as a global optimization technique with the speed of a local optimizer such as Newton-Raphson.

The simulation results suggest that the proposed method has a good performance, in terms of the correct estimation of the covariance parameters in linear mixed models, compared to Newton-Raphson algorithm and it was even better in obtaining positive definite covariance matrices when both random effects and error are included in the model with structured covariance matrices.

Finally, for the proposed optimization method based on hybrid genetic algorithms, there is still room for improvement in terms of calibration of the tuning parameters for the genetic algorithm and it is an interesting problem for further research.

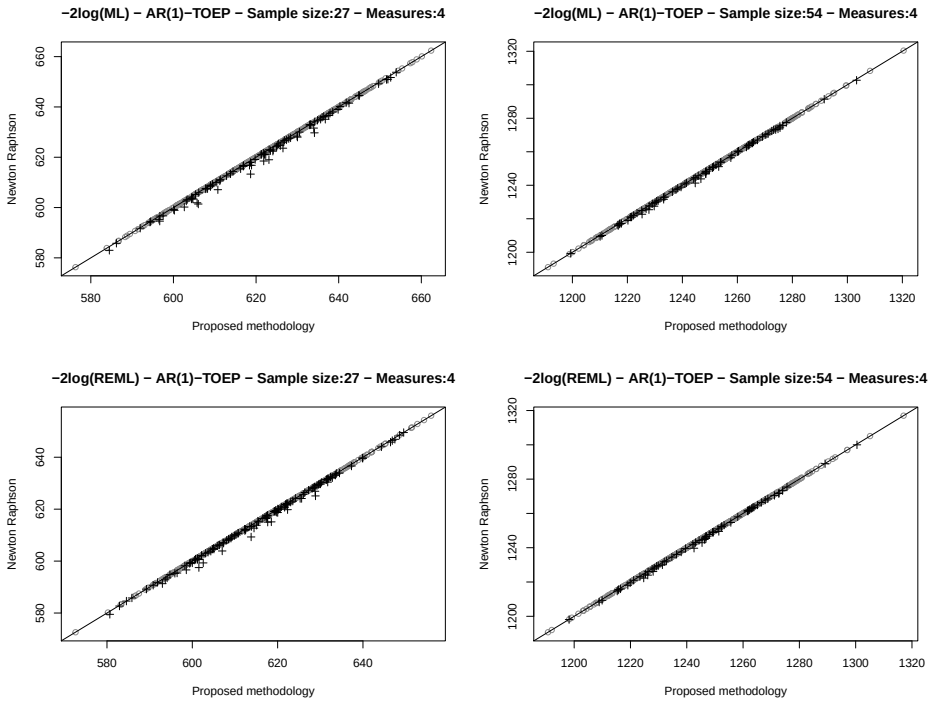

FIGURE 2: Comparison between the proposed methodology and the Newton-Raphson algorithm of the 500 optimal values for both $-2\log(ML)$ and $-2\log(REML)$ using the combination of AR(1)-TOEP structures. The line $x = y$ is added in each scenario to facilitate comparison. Crosses indicate that the estimated covariance matrix $\boldsymbol{D}$ is not PD for Newton-Raphson algorithm. Gray circles indicate that both proposed method and Newton-Raphson are PD.
Source: Author.

TABLE 4: Results of Monte Carlo simulations for the proposed method compared with Newton-Raphson and using the MI-TOEP structure.

Maximum likelihood (ML)													
		4-27						4-54					
		Proposed			Newton-Raphson			Proposed			Newton-Raphson		
Parm	Real	Mean	Median	RMSE	Mean	Median	RMSE	Mean	Median	RMSE	Mean	Median	RMSE
σ_0	2.069	2.040	2.036	0.159	2.025	2.028	0.161	2.060	2.054	0.120	2.053	2.049	0.119
σ_1	2.069	1.971	1.966	0.291	1.994	1.985	0.274	2.015	1.999	0.213	2.027	2.011	0.205
ρ_1	0.755	0.919	0.928	0.673	0.699	0.928	0.399	0.816	0.837	0.427	0.726	0.837	0.300
β_0	16.270	16.182	16.239	1.400	16.182	16.239	1.400	16.293	16.309	1.009	16.293	16.309	1.009
β_1	1.139	1.267	1.229	2.044	1.267	1.229	2.044	1.067	1.098	1.507	1.067	1.098	1.507
β_2	0.797	0.797	0.811	0.531	0.797	0.811	0.531	0.762	0.759	0.377	0.762	0.759	0.377
β_3	-0.321	-0.371	-0.372	0.815	-0.371	-0.372	0.815	-0.300	-0.274	0.604	-0.300	-0.274	0.604
Restricted Maximum likelihood (REML)													
		4-27						4-54					
		Proposed			Newton-Raphson			Proposed			Newton-Raphson		
Parm	Real	Mean	Median	RMSE	Mean	Median	RMSE	Mean	Median	RMSE	Mean	Median	RMSE
σ_0	2.069	2.060	2.056	0.159	2.047	2.048	0.158	2.070	2.064	0.120	2.064	2.060	0.119
σ_1	2.069	2.056	2.051	0.287	2.078	2.070	0.276	2.058	2.041	0.211	2.069	2.051	0.205
ρ_1	0.755	0.867	0.871	0.633	0.680	0.871	0.402	0.792	0.814	0.414	0.713	0.814	0.302
β_0	16.270	16.182	16.239	1.400	16.182	16.239	1.400	16.293	16.309	1.009	16.293	16.309	1.009
β_1	1.139	1.267	1.229	2.044	1.267	1.229	2.044	1.067	1.098	1.507	1.067	1.098	1.507
β_2	0.797	0.797	0.811	0.531	0.797	0.811	0.531	0.762	0.759	0.377	0.762	0.759	0.377
β_3	-0.321	-0.371	-0.372	0.815	-0.371	-0.372	0.815	-0.300	-0.274	0.604	-0.300	-0.274	0.604

Source: Author.

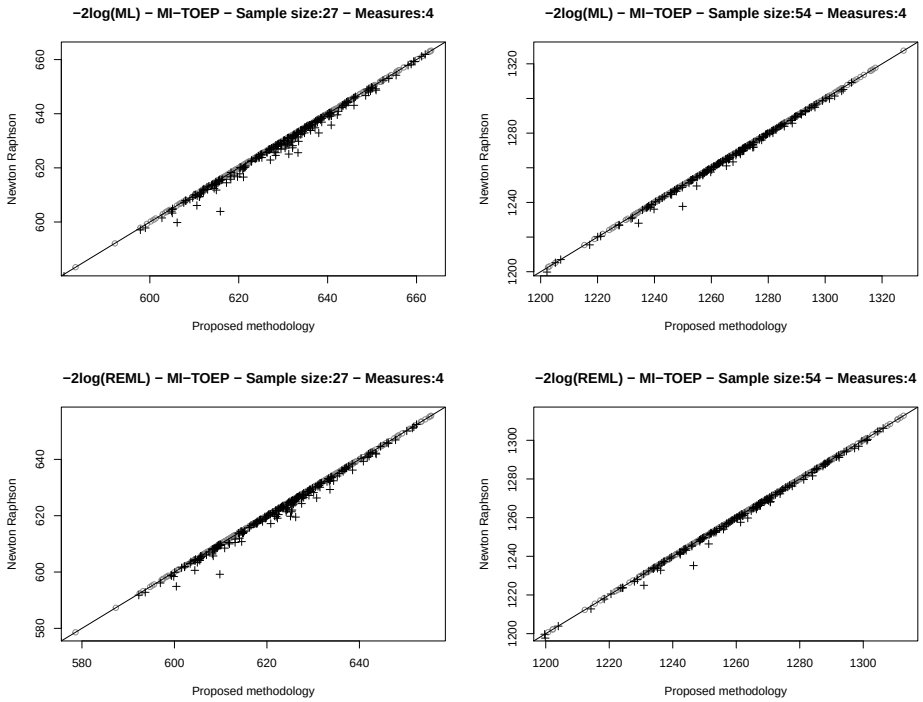

FIGURE 3: Comparison between the proposed methodology and the Newton-Raphson algorithm of the 500 optimal values for both $-2\log(ML)$ and $-2\log(REML)$ using the combination of MI-TOEP structures. The line $x = y$ is added in each scenario to facilitate comparison. Crosses indicate that the estimated covariance matrix $\boldsymbol{D}$ is not PD for Newton-Raphson algorithm. Gray circles indicate that both proposed method and Newton-Raphson are PD.
Source: Author.

Acknowledgments

We thank the school of statistics from the Universidad Nacional de Colombia at Medellín for its constinuous support to research initiatives. This work was partially supported by Colombian Institute for Science and Technology (Colciencias) Scholarship Program No. 647.

[Received: July 2022 — Accepted: March 2023]

References

Box, G., Jenkins, G. & Reinsel, G. (1970), *Time Series Analysis: Forecasting and Control*, fourth edn, John Wiley and Sons, New Jersey.

Chehouri, A., R. Younes, J. P. & Ilinca, A. (2016), 'A constraint-handling technique for genetic algorithms using a violation factor', *Journal of Computer Science* **12(7)**, 350–362.

Coello, C. (2002), 'Theoretical and Numerical Constraint-Handling Techniques used with Evolutionary Algorithms: A Survey of the State of the Art', *Computer Methods in Applied Mechanics and Engineering* **191**, 1245–1287.

Coley, D. A. (1998), *Introduction to genetic algorithms for scientists and engineers*, first edn, World scientific, River Edge, NJ, USA.

Demidenko, E. (2004), *Mixed models: Theory and applications with R*, second edn, Wiley, New Jersey.

Dempster, A., Laird, N. & Rubin, E. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society* **39(1)**, 1–38.

El-Mihoub, T. A., Hopgood, A. A., Nolle, L. & Battersby, A. (2006), 'Hybrid Genetic Algorithms: A review', *Engineering Letters* **13**, 124–137.

Henderson, C. R. (1984), Applications of linear models in animal breeding, Technical Report Press, University of Guelph, Guelph, Canada.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems*, first edn, University of Michigan Press, Ann Arbor.

Kuri-Morales, A. F. & Gutiérrez-García, J. (2002), 'Penalty Function Methods for Constrained Optimization with Genetic Algorithms: A Statistical Analysis', *MICAI 2002: Advances in Artificial Intelligence* **2313**, 108–117.

Laird, N. M. & Ware, J. H. (1982), 'Random-effects models for longitudinal data', *Biometrics* **38**, 963–974.

Lin, C. (2013), 'A rough penalty genetic algorithm for constrained optimization', *Information Sciences* **241**, 119–137.

Lindstrom, M. J. & Bates, D. M. (1988), 'Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data', *Journal of the American Statistical Association* **83**, 1014–1022.

Madar, V. (2015), 'Direct formulation to Cholesky decomposition of a general nonsingular correlation matrix', *Statistics & Probability Letters* **103**, 142–147.

Mebane, W. R. J. & Sekhon, J. S. (2011), 'Genetic optimization using derivatives: The rgenoud package for r', *Journal of Statistical Software* **42**(11), 1–26. `https://www.jstatsoft.org/v42/i11/`

Michalewicz, Z. (1998), *Genetic algorithms + data structures= evolution programs*, second edn, Springer-Verlag, Berlin Heidelberg.

Munkres, J. R. (2000), *Topology*, second edn, Prentice Hall, Upper Saddle River.

Patterson, H. D. & Thompson, R. (1971), 'Recovery of inter-block information when block sizes are unequal', *Biometrika* **58**, 545–554.

Pinheiro, J. C. (1994), Topics in Mixed Effects Models, PhD thesis, University of Wisconsin-Madison, USA.

Pinheiro, J. C. & Bates, D. M. (1996), 'Unconstrained parametrizations for variance-covariance matrices', *Statistics and Computing* **6**, 289–296.

Pothoff, R. & Roy, S. (1964), 'A generalized multivariate analysis of variance model useful especially for growth curve problems', *Biometrika* **51**, 313–326.

Rao, C. (1972), 'Estimation of variance and covariance components in linear models', *Journal of the American Statistical Association* **67(337)**, 112–115.

Reeves, C. & Rowe, J. E. (2002), *Genetic algorithms: Principles and perspectives*, first edn, Springer, New York.

SAS Institute Inc. (2008), *SAS/STAT® 9.2 User's Guide. The MIXED Procedure (Chapter)*. Cary, NC: SAS Institute Inc.

Schott, J. R. (1997), *Matrix analysis for statistics*, third edn, Wiley, New Jersey.

Scrucca, L. (2017), 'On some extensions to ga package: hybrid optimization, parallelisation and islands evolution', *The R Journal.* **9(1)**, 187–206.

Verbeke, G. & Molenberghs, G. (1997), *Linear Mixed Models in Practice- A SAS-Oriented Approach*, first edn, Springer, New York.

West, B., Welch, K. & Galecki, A. (2006), *Linear Mixed Models-A practical guide using statistical software*, second edn, Chapman and Hall/CRC, London.

Wolfinger, R. (1993), 'Covariance structure selection in general mixed models', *Comunications in Statistics - Simulation and Computation* **22**, 1079–1106.

Wolfinger, R., Tobias, R. & Sall, J. (1994), 'Computing gaussian likelihoods and their derivates for general linear mixed models', *SIAM Journal on Scientific Computing* **15(6)**, 1294–1310.

Appendix A.

Appendix A.1.

The matrix $\sigma^2 \boldsymbol{I}_r + \sigma_1 \boldsymbol{J}_r$ has eigenvalues $(r\sigma_1 + \sigma^2)$ with multiplicity 1 and σ^2 with multiplicity $(r-1)$.

Proof. The characteristic polynomial of $\sigma^2\boldsymbol{I}_r + \sigma_1\boldsymbol{J}_r$ is given by

$$\begin{aligned} |\sigma^2\boldsymbol{I}_r + \sigma_1'\boldsymbol{J}_r - \lambda\boldsymbol{I}_r| &= |\sigma_1\boldsymbol{J}_r - (\lambda-\sigma^2)\boldsymbol{I}_r| \\ &= \sigma_1^r\left|\boldsymbol{J}_r - \frac{(\lambda-\sigma^2)}{\sigma_1}\boldsymbol{I}_r\right| \end{aligned}$$

But $\boldsymbol{J}_r = \mathbf{1}\mathbf{1}'$, where $\mathbf{1}$ denotes the all-ones vector of length r, so $\mathbf{1}\mathbf{1}'$ and $\mathbf{1}'\mathbf{1} = r$ have the same positive eigenvalue, r, and the others are equal to 0 with multiplicity $(r-1)$ (Schott, 1997, page 131). So

$$\frac{(\lambda-\sigma^2)}{\sigma_1} = r \quad \text{or} \quad \left(\frac{\lambda-\sigma^2}{\sigma_1}\right)^{r-1} = 0$$

and therefore the eigenvalues of $\sigma^2\boldsymbol{I}_r + \sigma_1\boldsymbol{J}_r$ are $\lambda = (r\sigma_1 + \sigma^2)$ with multiplicity 1 and $\lambda = \sigma^2$ with multiplicity $(r-1)$. □

Appendix A.2. Grandient and Hessian for Newton-Raphson algorithm

The l-th element for gradient vector and (l,s)-th element for Hessian matrix to optimize the function (6) (ML case), Wolfinger et al. (1994) showed that they are given by:

$$\begin{aligned} g_l^{ML} &= \frac{\partial\ell(\boldsymbol{G},\boldsymbol{\Sigma})}{\partial\theta_l} \\ &= \operatorname{tr}\left(\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_l}\right) - \boldsymbol{r}'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_l}\boldsymbol{V}^{-1}\boldsymbol{r} \end{aligned}$$

and

$$\begin{aligned} H_{l,s}^{ML} &= \frac{\partial^2\ell(\boldsymbol{G},\boldsymbol{\Sigma})}{\partial\theta_l\partial\theta_s} \\ &= -\operatorname{tr}\left(\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_l}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_s}\right) + \operatorname{tr}\left(\boldsymbol{V}^{-1}\frac{\partial^2\boldsymbol{V}}{\partial\theta_l\partial\theta_s}\right) + 2\boldsymbol{r}'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_l}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_s}\boldsymbol{V}^{-1}\boldsymbol{r} \\ &\quad -2\boldsymbol{r}'\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_l}\boldsymbol{V}^{-1}\boldsymbol{X}^*\boldsymbol{X}^{*'}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_s}\boldsymbol{V}^{-1}\boldsymbol{r} - \boldsymbol{r}'\boldsymbol{V}^{-1}\frac{\partial^2\boldsymbol{V}}{\partial\theta_l\theta_s}\boldsymbol{V}^{-1}\boldsymbol{r}, \end{aligned}$$

where $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{C}$ for a matrix $\boldsymbol{C}$ satisfying $\boldsymbol{C}\boldsymbol{C}' = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}$.

Wolfinger et al. (1994) also showed that the l-th element for gradient vector and (l,s)-th element for Hessian matrix to optimize the function (8) (REML case) are given by:

$$g_l^{REML} = \frac{\partial\ell_R(\boldsymbol{G},\boldsymbol{\Sigma})}{\partial\theta_l} = g_l^{ML} - \operatorname{tr}\left(\boldsymbol{X}^{*'}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_l}\boldsymbol{V}^{-1}\boldsymbol{X}^*\right)$$

and

$$\begin{aligned} H_{l,s}^{REML} &= H_{l,s}^{ML} + 2\times\operatorname{tr}\left(\boldsymbol{X}^{*'}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_l}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_s}\boldsymbol{V}^{-1}\boldsymbol{X}^*\right) \\ &\quad -\operatorname{tr}\left(\boldsymbol{X}^{*'}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_l}\boldsymbol{V}^{-1}\boldsymbol{X}^*\boldsymbol{X}^{*'}\boldsymbol{V}^{-1}\frac{\partial\boldsymbol{V}}{\partial\theta_s}\boldsymbol{V}^{-1}\boldsymbol{X}^*\right) - \operatorname{tr}\left(\boldsymbol{X}^{*'}\boldsymbol{V}^{-1}\frac{\partial^2\boldsymbol{V}}{\partial\theta_l\theta_s}\boldsymbol{V}^{-1}\boldsymbol{X}^*\right). \end{aligned}$$

Appendix B. Hybrid Genetic Algorithms

Genetic Algorithms are heuristic methods based on processes of natural selection and genetic dynamics, as the key to solve optimization problems. Holland (1975) was the first to implement the idea of a population evolutionary development to find an optimal solution to a specific problem.

The advantage of evolutionary algorithms includes its ability to address problems for which the objetive function has all types of discontinuities and restrictions, which are difficult to control. The optimization of the functions (6) and (8) is not an easy task, given the condition of positive definiteness for the covariance matrices and also the restriction to keep a specific structure for these matrices.

The power of GAs as a global optimization technique could be combined with the speed of a local optimizer (Scrucca, 2017). This extension of GAs is known as *Hybrid Genetic Algorithms* which incorporates a global search and also a local optimization based on the classical methodologies, such as Nelder-Mead or Newton-Raphson. This idea accelerates the search towards the global optimum El-Mihoub et al. (2006) and keeps the GAs ability to address restrictions difficult to manage. Scrucca (2017) describes the HGAs technique and gives different applications by using the package "GA" of the software R.

Appendix B.1. HGAs Components

The HGAs are described by the following components:

- *Fitness function:* used to describe a specific problem. It depends on a vector of parameters and the goal of GAs is to optimize it (Coley, 1998).
- *Search space:* a wide range of possible solutions to locate the optimum of the fitness function. In the continuous case, the search space is, in general, a hypercube.
- *Population:* a set of possible solutions in the search space. Individuals of this population are coded in a *fixed-length bit string* representation.
- *Crossover method:* It mixes genetic fragments of the better solutions to form new, on average even better solutions.
- *Mutation operator:* It allows permanent diversity within the solutions.
- *Local optimizer:* a classical methodology based on derivatives to carry out a local optimization.

Appendix B.2. HGAs Parameters

The parameters into the HGAs, related with the above components are described as follows:

- *Search space:* It depends on the number of parameters in the fitness function and it is, in general, an euclidean subset which contains a wide range of possible solutions. We denote it by Ω.

- *Number of iterations:* It is a fixed integer indicating how many times must be repeated the optimization process. We denote it by n_{iter}.

- *Population size:* It is the number of elements in the search space at each iteration. We denote it by n_{pop}.

- *Crossover probability:* It is a fixed value in the interval $(0; 1)$ which gives the probability of applying the crossover procedure at each iteration. We denote it by p_{cross}.

- *Mutation probability:* It is a fixed value in the interval $(0; 1)$ which gives the probability of applying the mutation procedure at each iteration. We denote it by p_{muta}.

- *Local optimization probability:* It is a fixed value in the interval $(0; 1)$ which gives the probability of applying the local search at each iteration. We denote it by p_{local}.

- *Percentage of elitism:* It is the percentage of the best ranked individuals in the population to be selected for the crossover process. We denote it by $p_{elitism} \times 100$.

For a more detailed development of GAs and HGAs see, for instance Michalewicz (1998) and Reeves & Rowe (2002).