Linear regression analysis for interval-valued data based on the Lasso technique

Paolo Giordani

Department of Statistical Sciences
Sapienza University of Rome
P.le Aldo Moro, 5 - I-00185 Rome
paolo.giordani@uniroma1.it

Abstract

A new method for linear regression analysis of interval-valued data is proposed. In particular, the linear relationship between an interval-valued response variable and a set of interval-valued explanatory variables is investigated by considering two regression models, one for the midpoints (the locations of the intervals) of the response and explanatory variables and the other one for the radii (the imprecision). The regression coefficients of the two models are estimated in such a way that those for the midpoints are close to the corresponding ones for the radii as much as possible. Taking inspiration from the Lasso technique this is done by fixing a threshold expressing the maximum allowed level of diversity between the two sets of regression coefficients. The results of a simulation experiment and some applications to real data are reported in order to show the usefulness of the proposed method, called Lasso-IR (Lasso-based Interval-valued Regression).

Keywords: Interval-valued data, Linear regression analysis, Lasso

1. Introduction

In regression analysis the relationship between a response variable and a number of explanatory variables is investigated. The (response and explanatory) variables usually are single-valued. However, in several real-life situations, the available information is formalized in terms of intervals. The need for interval-valued data may arise in connection with the imprecision of measurement devices (for instance, in the case of mineral concentrations) or

with the data fluctuations in the case of recorded measures during a specific interval of time (for instance, daily pollution, daily stock price). In fact, considering the minimum and maximum recorded values offers a more complete insight about the phenomenon at hand than considering the average values. In all these cases, the attributes involved can be expressed by a lower and an upper bound, providing the boundaries of the interval-valued data. Therefore, an interval-valued datum z can be characterized by the pair of values $\underline{z}$ and $\overline{z}$ with $\overline{z} \geq \underline{z}$ where $\underline{z}$ and $\overline{z}$ denote the lower and upper bound, respectively. Another representation of an interval can be done in terms of the so-called midpoint and radius, say z_M and z_R, with $z_M = \frac{(\overline{z}+\underline{z})}{2}$ and $z_R = \frac{(\overline{z}-\underline{z})}{2}$. The midpoint is the center of an interval (the location), whereas the radius is the half-width of an interval (a measure of the imprecision) with, of course, $z_R \geq 0$. In the following we shall adopt the latter representation. In this work, limiting our attention to the linear case, regression analysis for interval-valued data is studied. In the literature, the topic has been extensively analyzed especially in the last decade. We can roughly distinguish two lines of research according to the use of interval arithmetic. For instance, this is the case for González-Rodríguez et al. (2006, 2007), Gil et al. (2007), Blanco et al. (2008, 2010), whereas the other approach is adopted by Billard and Diday (2000), Domingues et al. (2010), Lima-Neto and De Carvalho (2008, 2010). In this paper we are going to introduce a new linear regression method for interval-valued data that could be assigned to the latter line of research. However, as it will be clarified in the following, a connection with the some papers belonging to the former line of research can also be highlighted. The paper is organized as follows. In the next section the above-mentioned lines of research will be recalled and their features will be discussed. Section 3 contains the here-proposed model. In Section 4 an algorithm to estimate the parameters of the regression model is provided. In order to show how the method works in practice the results of a simulation study and of some applications are discussed in Sections 5 and 6, respectively. Finally, some concluding remarks are given in Section 7.

2. Linear regression models for interval-valued data

We now recall the regression models by Lima-Neto and De Carvalho (2010) and by González-Rodríguez et al. (2007) and we then discuss their peculiarities.

2.1. The Lima-Neto and De Carvalho (2010) model

Let Y be the interval-valued response variable and $X_1, \ldots, X_p$ be the set of interval-valued explanatory variables observed on n units. The so-called Constrained Center and Range Method (CCRM) proposed by Lima-Neto and De Carvalho (2010) assumes the following linear relationship between the response and explanatory variables:

$$\begin{aligned} Y_{Mi} &= b_{M0} + \sum_{j=1}^{p} b_{Mj} X_{Mij} + e_{Mi},\ i = 1, \ldots, n, \\ Y_{Ri} &= b_{R0} + \sum_{j=1}^{p} b_{Rj} X_{Rij} + e_{Ri},\ i = 1, \ldots, n, \end{aligned} \tag{1}$$

where Y_{Mi} and Y_{Ri} denote, respectively, the midpoint and the radius of Y_i $(i = 1, \ldots, n)$, X_{Mij} and X_{Rij} are, respectively, the midpoint and the radius of X_{ij} $(i = 1, \ldots, n,\ j = 1, \ldots, p)$ and e_{Mi} and e_{Ri} $(i = 1, \ldots, n)$ are the (single-valued) residuals for the model of the midpoints and for that of the radii, respectively. Finally, b_{Mj}'s and b_{Rj}'s $(j = 0, \ldots, p)$ are the regression coefficients for the models of the midpoints and of the radii, respectively.
The optimal parameters are obtained according to the least-squares approach by minimizing the sum of squared residuals of the two models constraining the b_{Rj}'s $(j = 0, \ldots, p)$ to be non-negative in order to guarantee the non-negativity of the estimated radii of the response variable. We thus have

$$\begin{aligned} &\min_{b_{M0}, \ldots b_{Mp}, b_{R0}, \ldots b_{Rp}} \sum_{i=1}^{n} \left[(e_{Mi})^2 + (e_{Ri})^2 \right], \\ &\text{s.t. } b_{Rj} \geq 0,\ j = 0, \ldots, p. \end{aligned} \tag{2}$$

In matrix notation (1) can be formalized as

$$\begin{aligned} \mathbf{y}_M &= \mathbf{X}_M \mathbf{b}_M + \mathbf{e}_M, \\ \mathbf{y}_R &= \mathbf{X}_R \mathbf{b}_R + \mathbf{e}_R, \end{aligned} \tag{3}$$

with $\mathbf{y}_M = (Y_{M1}, \ldots, Y_{Mn})'$, $\mathbf{y}_R = (Y_{R1}, \ldots, Y_{Rn})'$, $\mathbf{X}_M$ and $\mathbf{X}_R$ are the $(n \times p + 1)$ matrices of the midpoints and of the radii of the explanatory variables with generic element X_{Mij} and X_{Rij} $(i = 1, \ldots, n,\ j = 0, \ldots, p)$, respectively, assuming that $X_{Mij} = X_{Rij} = 1$ when $j = 0$ $(i = 1, \ldots, n)$. Finally, $\mathbf{e}_M$ and $\mathbf{e}_R$ are the vectors of length n of the residuals and $\mathbf{b}_M$ and $\mathbf{b}_R$ are the vectors of the regression coefficients with generic elements b_{Mj}

and b_{Rj} $(j = 0, \ldots, p)$, respectively. In matrix notation (2) can be rewritten as

$$\begin{array}{l} \min\limits_{\mathbf{b}_M, \mathbf{b}_R} \|\mathbf{e}_M\|^2 + \|\mathbf{e}_R\|^2 = \|\mathbf{Y}_M - \mathbf{X}_M\mathbf{b}_M\|^2 + \|\mathbf{Y}_R - \mathbf{X}_R\mathbf{b}_R\|^2, \\ \text{s.t. } \mathbf{b}_R \geq \mathbf{0}. \end{array} \tag{4}$$

The optimal values of $\mathbf{b}_M$ can be found by solving an ordinary regression problem taking into account that the second norm of (4) does not depend on $\mathbf{b}_M$. Therefore, the estimate of $\mathbf{b}_M$ is $\widehat{\mathbf{b}_M} = (\mathbf{X}'_M\mathbf{X}_M)^{-1}\mathbf{X}'_M\mathbf{y}_M$. Similarly, the first norm of (4) can be considered as a constant for $\mathbf{b}_R$. The estimate of $\mathbf{b}_R$ is obtained solving a constrained regression problem. This problem can be recognized as a particular non-negative least-squares (NNLS) problem, the (iterative) solution of which has been provided by Lawson and Hanson (1995). Therefore, CCRM consists of solving two *separate* regression problems. It follows that CCRM does not take into account jointly the features of the interval-valued data. Although every interval is characterized by two measures, namely midpoint (location) and radius (imprecision), CCRM does not use them jointly when estimating the regression coefficients. This implies that in some cases the results of the analysis could be incomplete. See, for further details, Lima-Neto and De Carvalho (2010).

2.2. The González-Rodríguez et al. (2007) model

Let Y and X be the response and explanatory variables, respectively. The Interval Arithmetic-based Linear Model (IALM) proposed by González-Rodríguez et al. (2007) can be formalized as

$$Y = b_1 X + e, \tag{5}$$

where b_1 is the single-valued regression parameter and e is the residual which takes the form of a random interval-valued set such that its expected value is the interval b_0. It follows that $E(Y|x) = b_1 x + b_0$ for any interval-valued realization x of X being E the expected value in the Aumann sense. According to (5) we then have that $Y_M = b_1 X_M + e_M$ and $Y_R = |b_1| X_R + e_R$ with obvious notation.

Given the interval-valued sample data (x_i, y_i), $i = 1, \ldots, n$ the estimates of b_0 and b_1 can be obtained following the least-squares approach. This requires to introduce a suitable metric for intervals, say d_I^2. The estimates of the parameters of (5) can be found by solving the following minimization

problem:

$$\min_{b_0,b_1} \sum_{i=1}^{n} d_I^2 \left(y_i, b_0 + b_1 x_i\right). \tag{6}$$

A crucial point concerning the minimization of (6) is that the optimal value of b_1 must be limited to the subset of the single-valued (and real) numbers such that the Hukuhara difference $y_i -_H \widehat{b_1} x_i$ is well-defined ($i = 1, \ldots, n$). The main distinctive features of IALM in comparison with CCRM are that interval arithmetic is used allowing us to *jointly* take into account the midpoint and radius information and probabilistic assumptions are made allowing us to make inference on the results of the analysis. Furthermore, it is worth mentioning that the parameter b_1 does not only play the "standard" role of slope. In fact, to some extent it informs us as to whether a decreasing or increasing relationship exists between the midpoints of the response and explanatory variables. Moreover, b_1 can also be interpreted as a sort of "imprecision propagation" factor describing how the imprecisions of the response and explanatory variables are connected. Therefore, b_1 can be seen as a compromise between a measure of slope and a measure of propagation of imprecision (see González-Rodríguez et al. , 2009, in which the fuzzy counterpart of IALM is proposed). As the authors note (see, e.g., Blanco et al. , 2010), the applicability of the model can be limited whenever two different linear relationships for the midpoints and the radii exist. This is so because if the relationships take the form $Y_M = b_{M0} + b_{M1} X_M$ and $Y_R = b_{R0} + b_{R1} X_R$ with $b_{R1} \neq |b_{M1}|$, then there is no $b_1 \in \mathbb{R}$ such that the linear relationship can be expressed by (5). See for further details and for the extension to the multiple linear regression case González-Rodríguez et al. (2007). Finally, it must be noted that Blanco et al. (2010) propose a possible remedy to such a drawback introducing two *different* parameters for the slope and the propagation of the imprecision which, however, are estimated considering *jointly* the midpoint and radius information. In order to make the model more flexible a different approach is followed in this paper as it will be discussed in the next section.

3. A new linear regression model for interval-valued data

In this section we are going to propose a novel regression technique for interval-valued data. Such a model takes inspiration from the idea underlying IALM in the sense that the attempt to seek a common set of regression

coefficients for the midpoint and the radius model is pursued even if *to some extent*. This will be done by adding specific regression coefficients for the radii in such a way to cope properly with all those situations in which the slope is different from the propagation of the imprecision. However, these additive coefficients are constrained to be as small as possible according to a tuning parameter to be chosen by the researcher. Differently from González-Rodríguez et al. (2007), the problem will not managed by means of interval arithmetic. In this respect, the here-proposed technique may resemble the CCRM approach because the problem is addressed as an optimization problem involving the constrained minimization of an objective function.

3.1. The model

Let Y and $X_1, \ldots, X_p$ be the interval-valued response and explanatory variables observed on a set of n units. In order to study the linear relationship between Y and $X_1, \ldots, X_p$ we have:

$$\begin{array}{l} \mathbf{y}_M = \mathbf{y}_M^* + \mathbf{e}_M = \mathbf{X}_M \mathbf{b}_M + \mathbf{e}_M \text{ (midpoint model)}, \\ \mathbf{y}_R = \mathbf{y}_R^* + \mathbf{e}_R = \mathbf{X}_R \mathbf{b}_R + \mathbf{e}_R = \mathbf{X}_R (\mathbf{b}_M + \mathbf{b}_A) + \mathbf{e}_R \text{ (radius model)}, \end{array} \tag{7}$$

where $\mathbf{y}_M$ and $\mathbf{y}_R$ denote the vectors of length n of the observed midpoints and of the observed radii of the response variable and $\mathbf{y}_M^*$ and $\mathbf{y}_R^*$ are the vectors of the theoretical midpoints and radii of the response variable. $\mathbf{X}_M$ and $\mathbf{X}_R$ are the matrices of order $(n \times p+1)$ of the midpoints and of the radii of the explanatory variables containing the unit vector of length n in their first column. $\mathbf{e}_M$ and $\mathbf{e}_R$ denote the residual vectors. Finally, $\mathbf{b}_M$ and $\mathbf{b}_R$ are the vectors of length $(p+1)$ of the regression coefficients for the midpoint and radius models, respectively, where $\mathbf{b}_R = \mathbf{b}_M + \mathbf{b}_A$ being $\mathbf{b}_A$ the vector of the additive coefficients. Therefore, the coefficients of the radius model $\mathbf{b}_R$ are equal to those of the midpoint model $\mathbf{b}_M$ up to the additive coefficients $\mathbf{b}_A$.

3.2. The minimization problem

The parameter vectors $\mathbf{b}_M$ and $\mathbf{b}_A$ are estimated in such a way to minimize a suitable dissimilarity measure between observed and theoretical data. For this purpose, the squared distance d_θ^2 proposed by Trutschnig et al. (2009) is considered. Given two intervals G and H it is

$$d_\theta^2 = (G_M - H_M)^2 + \theta\,(G_R - H_R)^2 \tag{8}$$

with $\theta \in (0, 1]$. When $\theta = 1$, d_θ^2 compares G and H by the sum of the squared distances of their midpoints and of their radii. The choice of θ depends on the relative importance of the radius distance with respect to the midpoint distance. A reasonable choice seems to be $\theta = \frac{1}{3}$. See, for more details, Trutschnig et al. (2009). Using (8), the loss function to be minimized is

$$\min_{\mathbf{b}_M, \mathbf{b}_A} \|\mathbf{e}_M\|^2 + \theta \|\mathbf{e}_R\|^2 = \|\mathbf{y}_M - \mathbf{X}_M \mathbf{b}_M\|^2 + \theta \|\mathbf{y}_R - \mathbf{X}_R (\mathbf{b}_M + \mathbf{b}_A)\|^2 . \quad (9)$$

The loss function in (9) requires some constraints in order to guarantee that the estimated radii are non-negative and that the additive coefficients $\mathbf{b}_A$ are as small as possible. The former requirement can be achieved setting

$$\mathbf{X}_R (\mathbf{b}_M + \mathbf{b}_A) \geq \mathbf{0}. \quad (10)$$

The latter requirement can be managed using the Lasso technique (see Tibshirani , 1996). The Lasso, which is the acronym of Least Absolute Shrinkage and Selection Operator, is a method for estimation in (single-valued) regression aiming at shrinking some regression coefficients and setting some others to 0. This is done by minimizing the residual sum of squares with the constraint that the sum of the absolute values of the regression coefficients is smaller than a threshold. Let $\mathbf{y}$ and $\mathbf{X}$ be the vector of length n of the response variable and the matrix of order $(n \times p + 1)$ of the explanatory variables with the unit vector in its first column, respectively. The Lasso problem is

$$\begin{aligned} &\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 , \\ &\text{s.t.} \sum_{j=1}^{p} |b_j| \leq t, \end{aligned} \quad (11)$$

where $\mathbf{b} = (b_0, \ldots, b_p)'$ is the vector of the regression coefficients and $t \geq 0$ is a tuning parameter which controls the amount of shrinkage applied to the estimates. It can be shown that (11) is a quadratic programming problem with linear inequality constraints, the solution of which can be found in Lawson and Hanson (1995). As already noted, the nature of the constraint is such that it tends to produce some coefficients that are exactly zero. For a better insight into this property we refer to Tibshirani (1996). The use of the Lasso constraint in the here-proposed model for interval-valued data can be carried out as

$$\sum_{j=0}^{p} |b_{Aj}| \leq t. \quad (12)$$

This allows us to limit the magnitude of the additive coefficients as much as possible according to the choice of t. Note that in (12) the Lasso constraint is considered for all the additive coefficients including the intercept.
Taking into account (9), (10) and (12) we then get the following constrained minimization problem:

$$\begin{aligned} &\min_{\mathbf{b}_M, \mathbf{b}_A} \|\mathbf{y}_M - \mathbf{X}_M \mathbf{b}_M\|^2 + \theta \|\mathbf{y}_R - \mathbf{X}_R (\mathbf{b}_M + \mathbf{b}_A)\|^2, \\ &\text{s.t. } \mathbf{X}_R (\mathbf{b}_M + \mathbf{b}_A) \geq \mathbf{0}, \ \sum_{j=0}^{p} |b_{Aj}| \leq t. \end{aligned} \tag{13}$$

We refer to the problem in (13) as Lasso-based Interval-valued Regression (Lasso-IR)

3.3. The choice of t

The minimization of (13) requires to choose the shrinkage parameter t. The possible values of t range from 0 to t_{MAX}. When $t = 0$ it is $\mathbf{b}_A = \mathbf{0}$ and, therefore, the same regression coefficients for the midpoints and the radii are found. We could state that in this case the hypothesis of IALM is adopted. However, when $t = 0$, the solutions from IALM and Lasso-IR in general differ since, as already noted, the former involves interval arithmetic (and the latter does not). It should be underlined that, when $t = 0$, a feasible solution for (13) always exists using non-negative values for $\mathbf{b}_M$. The value of t_{MAX} can be found as follows. One can first set $\mathbf{b}_M$ equal to the unconstrained regression coefficients of $\mathbf{y}_M$ with respect to $\mathbf{X}_M$, say $\widehat{\mathbf{b}_M}$. Then, $\widehat{\mathbf{b}_A}$ can be found by determining the optimal constrained regression coefficients of $\mathbf{y}_R - \mathbf{X}_M \widehat{\mathbf{b}_M}$ with respect to $\mathbf{X}_R$ provided that the estimates of $\mathbf{y}_R$ are non-negative. Then $t_{\text{MAX}} = \sum_{j=0}^{p} \left|\widehat{b_{Aj}}\right|$, being $\widehat{b_{Aj}}$ the j-th element of $\widehat{\mathbf{b}_A}$ $(j = 0, \ldots, p)$. In practice, t_{MAX} is the smallest value such that two *separate* regression problems for the midpoint and the radius models are solved. Of course, if $t > t_{\text{MAX}}$ is chosen the same solution for the case with $t = t_{\text{MAX}}$ is obtained. To some limited extent the solution with $t \geq t_{\text{MAX}}$ is comparable with the CCRM one even if the two solutions differ as explained in Remark 1.

Remark 1
When $t \geq t_{\text{MAX}}$, it does not necessarily hold that the solution obtained from Lasso-IR coincides with the CCRM one. In fact, in CCRM, the non-negativity of the estimated radii is achieved by seeking *non-negative* regression coefficients for the radius model. In our model such a constraint is

relaxed admitting negative regression coefficients for the radius model. In order to better clarify this point let us consider the following example:

$$\mathbf{y}_R = \begin{pmatrix} 0.9 \\ 0.5 \\ 0.4 \\ 0 \end{pmatrix} \quad \mathbf{X}_R = \begin{pmatrix} 2.4 & 5 & 1 \\ 1 & 3.5 & 3 \\ 0 & 1 & 1.2 \\ 2 & 3 & 0 \end{pmatrix}.$$

By applying CCRM we get $\widehat{\mathbf{b}}_R^{CCRM} = \begin{pmatrix} 0 & 0.11 & 0.10 \end{pmatrix}'$, whereas if we do not impose the non-negativity of the regression coefficients a perfect fit solution with $\widehat{\mathbf{b}}_R = \begin{pmatrix} -1.5 & 1.0 & -0.5 \end{pmatrix}'$ is found.

Prior knowledge on the data under investigation could facilitate the choice of t. If the value of t chosen by the researcher is close to 0, then approximately the same linear relationships for midpoints and radii are assumed. The opposite comment holds when t is high. However, in order to choose t, we suggest to consider cross-validation techniques, such as the k-fold cross-validation procedure (see, e.g., Efron and Tibshirani , 1993) in which the data are split into k subsets of (approximately) equal size. Then, the model is fitted to the data consisting of $k-1$ subsets (training set) and its predictive accuracy is assessed using the remaining subset (test set). This process is repeated k times in such a way that the k subsets are used once as test set. In Lasso-IR for different values of t ranging from 0 to t_{MAX} we can compute the predictive accuracy as

$$CV\left(t\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_{Mi} - \widehat{y_{Mi}}^{(-k(i))}\left(t\right)\right)^2 + \theta\left(y_{Ri} - \widehat{y_{Ri}}^{(-k(i))}\left(t\right)\right)^2\right], \quad (14)$$

where $\widehat{y_{M}}_i^{(-k(i))}$ and $\widehat{y_{R}}_i^{(-k(i))}$ denote the i-th fitted midpoint and radius, respectively, computed setting t and removing the k-th part of the data. Then the optimal value of t is

$$t_{\text{OPT}} = \underset{0\leq t\leq t_{\text{MAX}}}{\arg\min}\, CV\left(t\right). \quad (15)$$

4. The estimation procedure

To solve the minimization problem in (13) an alternating least squares algorithm is proposed. It consists of updating separately the vectors $\mathbf{b}_M$ and

$\mathbf{b}_A$ keeping fixed the remaining one. Whenever a vector is updated, the loss function to be minimized decreases. After updating both the vectors, if the loss function value decreases less than a specified percentage (e.g. 0.0001%) from the previous function value, we consider the algorithm converged, otherwise we repeat the updates of $\mathbf{b}_M$ and $\mathbf{b}_A$. The function in (13) has a lower bound and, therefore, the function value converges to a stable value. The updates of $\mathbf{b}_M$ and $\mathbf{b}_A$ are described below.

Update of $\mathbf{b}_M$

First of all it should be noted that the constraints in (13) do not play an active role in the update of $\mathbf{b}_M$. Therefore, from (13) the optimal value of $\mathbf{b}_M$ can be found as

$$\min_{\mathbf{b}_M} \|\mathbf{y}_M - \mathbf{X}_M \mathbf{b}_M\|^2 + \theta \|\mathbf{y}_R - \mathbf{X}_R (\mathbf{b}_M + \mathbf{b}_A)\|^2, \tag{16}$$

where $\mathbf{b}_A$ is fixed. The function in (16) can be rewritten as

$$\left\| \begin{bmatrix} \mathbf{y}_M \\ \theta^{1/2} (\mathbf{y}_R - \mathbf{X}_R \mathbf{b}_A) \end{bmatrix} - \begin{bmatrix} \mathbf{X}_M \\ \theta^{1/2} \mathbf{X}_R \end{bmatrix} \mathbf{b}_M \right\|^2 = \|\mathbf{c} - \mathbf{D}\mathbf{b}_M\|^2, \tag{17}$$

where $\mathbf{c}$ and $\mathbf{D}$ are implicitly defined in (17), from which we get

$$\widehat{\mathbf{b}_M} = (\mathbf{D}'\mathbf{D})^{-1} \mathbf{D}'\mathbf{c}. \tag{18}$$

Update of $\mathbf{b}_A$

In order to update $\mathbf{b}_A$ and starting from (13) it is easy to see that the problem to be solved is

$$\begin{aligned} &\min_{\mathbf{b}_A} \|\mathbf{y}_R - \mathbf{X}_R (\mathbf{b}_M + \mathbf{b}_A)\|^2, \\ &\text{s.t. } \mathbf{X}_R (\mathbf{b}_M + \mathbf{b}_A) \geq \mathbf{0}, \sum_{j=0}^{p} |b_{Aj}| \leq t, \end{aligned} \tag{19}$$

keeping $\mathbf{b}_M$ fixed. The problem in (19) can be recognized as a constrained regression problem where the response variable is $\mathbf{y}_R - \mathbf{X}_R \mathbf{b}_M$ and the explanatory ones are $\mathbf{X}_R$. The first set of constraints requires that the n estimates of $\mathbf{y}_R$ are non-negative. The second set of constraints can be exploited as 2^{p+1} inequality constraints corresponding to the 2^{p+1} different possible signs for the $\mathbf{b}_{Aj}$'s. This allows us to rewrite (19) as

$$\begin{aligned} &\min_{\mathbf{b}_A} \|(\mathbf{y}_R - \mathbf{X}_R \mathbf{b}_M) - \mathbf{X}_R \mathbf{b}_A\|^2, \\ &\text{s.t. } \begin{bmatrix} \mathbf{X}_R \\ \mathbf{H} \end{bmatrix} \mathbf{b}_A \geq \begin{bmatrix} -\mathbf{X}_R \mathbf{b}_M \\ t\mathbf{1}_{2^{p+1}} \end{bmatrix}, \end{aligned} \tag{20}$$

in which $\mathbf{1}_{2^{p+1}}$ is the unit column vector of length 2^{p+1} and $\mathbf{H}$ is a $(2^{p+1} \times p+1)$ matrix containing in its rows all the 2^{p+1} combinations of length $(p+1)$ of ± 1. The problem in (20) is a particular case of a linear least squares problem with inequality constrains of the form

$$\begin{array}{l} \min\limits_{\mathbf{w}} \|\mathbf{z} - \mathbf{V}\mathbf{w}\|^2, \\ \text{s.t. } \mathbf{F}\mathbf{w} \geq \mathbf{g}, \end{array} \tag{21}$$

where $\mathbf{z}$, $\mathbf{w}$ and $\mathbf{g}$ are vectors of length q, r and s, respectively, and $\mathbf{V}$ and $\mathbf{F}$ are matrices of order $(q \times r)$ and $(s \times r)$, respectively. It is straightforward to see that (20) coincides with (21) when $\mathbf{z} = \mathbf{y}_R - \mathbf{X}_R\mathbf{b}_M$, $\mathbf{V} = \mathbf{X}_R$, $\mathbf{w} = \mathbf{b}_A$, $\mathbf{F} = \begin{bmatrix} \mathbf{X}_R \\ \mathbf{H} \end{bmatrix}$ and $\mathbf{g} = \begin{bmatrix} -\mathbf{X}_R\mathbf{b}_M \\ t\mathbf{1}_{2^{p+1}} \end{bmatrix}$.
In the literature there exist several methods to solve (21) and, therefore, (20). See, for instance, Lawson and Hanson (1995) and Gill et al. (1981). A more relevant point is connected with the computational burden of (20) since the number of constraints is exponentially related to the number of explanatory variables. In particular, the total number of constraints is n (from (10)) + 2^{p+1} (from (12)). It is clear that, from a computational point of view, the number of the Lasso constraints can represent a problem. This has been already recognized by Tibshirani (1996), who suggests two procedures to overcome it. Both the procedures can easily be adapted to solve (20). However, in our analyses based on the method of Gill et al. (1981), we saw that the algorithm seemed to converge quickly. In this respect, some results are reported in the simulation experiment of Section 5.

The algorithm can be summarized as follows.
Step 1 (Inizialization): Randomly generate initial values for $\mathbf{b}_A$ fulfilling the constraints in (13). For instance the elements of $\mathbf{b}_A$ can be generated randomly from U[0,1] rescaling them if necessary.
Step 2 (Update of $\mathbf{b}_M$): Minimize (17) using (18).
Step 3 (Update of $\mathbf{b}_A$): Minimize (20) using the procedure provided by Gill et al. (1981).
Step 4 (Convergence): Check convergence.

The above described algorithm has been implemented using Matlab (see the appendix). Note that the update of $\mathbf{b}_A$ requires the Matlab routine `lsqlin` that can be found in the Matlab toolbox `optim`.

5. Simulation experiment

A simulation study has been carried out in order to evaluate the performance of Lasso-IR. Specifically, the simulation study aims at offering a better insight into the efficiency of the algorithm, its tendency to hit the global optimum, the reproducibility of the solution and the recovery performance. For this purpose interval-valued data sets were randomly generated and noise was added. We considered different numbers of units ($n = 60, 90, 120, 150$) and of explanatory variables ($p = 6, 9, 12, 15$ including the intercept). The midpoints and the radii of the explanatory variables were generated from U[-1,1] and U[0,1], respectively. They were stored in the matrices $\mathbf{X}_M$ and $\mathbf{X}_R$, which contained all 1's in their first column. The regression parameters in $\mathbf{b}_M$ and $\mathbf{b}_A$ were randomly generated according to six different cases (C1–C6). In cases C1–C3 the elements of $\mathbf{b}_M$ and $\mathbf{b}_A$ were generated from U[-1,1]. Furthermore, in Cases 1 and 2, respectively $\frac{2}{3}$ and $\frac{1}{3}$ of the elements of $\mathbf{b}_A$ were set to 0. A similar set-up was considered for cases C4–C6 but the coefficients were generated from U[0,1]. Finally, noise vectors for the midpoints $\mathbf{n}_M$ (generated from U[-1,1]) and for the radii $\mathbf{n}_R$ (generated from U[0,1]) were added. To tune the level of noise added we used $e = 0.1$ (low level), 0.5 (medium level), 1.0 (high level). Summing up, the randomly generated data sets took the form

$$\begin{aligned} \mathbf{y}_M &= \mathbf{X}_M \mathbf{b}_M + \varepsilon \mathbf{n}_M, \\ \mathbf{y}_R &= \mathbf{X}_R \left(\mathbf{b}_M + \mathbf{b}_A\right) + \varepsilon \mathbf{n}_R. \end{aligned} \tag{22}$$

Note that we scaled $\mathbf{n}_M$ and $\mathbf{n}_R$ in such a way that their sum of squares was equal to the sum of squares of $\mathbf{X}_M \mathbf{b}_M$ and $\mathbf{X}_R \left(\mathbf{b}_M + \mathbf{b}_A\right)$, respectively. Also note that, in the data generation process, we discarded and replaced every generated data set for which at least one radius of the response variable took a negative value.

For every level of every design variable five replications were considered. The design was fully crossed. Therefore, the total number of randomly generated data sets was 4 (numbers of units) $\times$ 4 (numbers of variables) $\times$ 3 (levels of noise added) $\times$ 6 (structures of the coefficients) $\times$ 5 (replications) = 1440. To limit the risk of hitting local optima we ran the algorithm considering five random starts and we chose the solution corresponding to the lowest value of the loss function. Finally, note that we set $t = \sum_{j=0}^{p} |b_{Aj}|$.

Concerning the frequency of hitting the global optimum we checked, for each data set, how many times the function vale was less than 0.1% bigger than

the lowest one (the one corresponding to the global optimum). We observed very satisfactory results. In fact, during the entire simulation study the global optimum was always attained. For each data set, we found negligible differences (lower than 10^{-4}) among the regression coefficients obtained using different random starts. It follows that the solution was fully reproducible.
The results about the computation time are reported in Table 1 in which the average computation time for every level of every design variable is given The average computation time was 7.42s. As expected it mainly depended

Table 1: Average computation time for every level of every design variable (seconds)

$n = 60$	$n = 90$	$n = 120$	$n = 150$	$p = 6$	$p = 9$	$p = 12$	$p = 15$	
8.51	7.53	7.21	6.41	0.07	0.29	2.14	27.16	
$\varepsilon = 0.1$	$\varepsilon = 0.5$	$\varepsilon = 1.0$	C1	C2	C3	C4	C5	C6
7.18	9.25	5.82	11.34	4.97	4.25	9.21	7.18	7.55

on the number of variables passing from 0.07s when $p = 6$ to 27.16s when $p = 15$. Also note that only 195 times (out of 7200) the algorithm took more than 60s and just once more than 120s.
To investigate the recovery of the regression parameters $\mathbf{b}_M$ and $\mathbf{b}_A$ we used the mean absolute difference (MAD)

$$\begin{aligned} MAD_M &= \sum_{j=1}^{p} \left| b_{Mj} - \widehat{b_{Mj}} \right| \Big/ p, \\ MAD_A &= \sum_{j=1}^{p} \left| b_{Aj} - \widehat{b_{Aj}} \right| \Big/ p. \end{aligned} \tag{23}$$

Table 2 contains the average MAD values for the coefficients of the midpoints and for the additive ones distinguished with respect to every level of every design variable. From Table 2 we can see that the recovery performance of the coefficients of the midpoint model was better than that of the additive coefficients. The average MAD values increased when the level of noise added increased and seemed to be affected by the structure of the coefficients. In particular, in cases C4–C6 the average MAD_M values were higher than those in cases C1–C3 and the same comment holds for MAD_A in cases C5 and C6 in comparison with cases C1–C4.
Finally, we investigated the ability of the method to detect the coefficients in $\mathbf{b}_A$ set to 0. For this purpose, we checked the average percentage of times

Table 2: Average MAD_M and MAD_A values for every level of every design variable

	$n=60$	$n=90$	$n=120$	$n=150$	$p=6$	$p=9$	$p=12$	$p=15$	
MAD_M	0.15	0.13	0.12	0.11	0.11	0.12	0.14	0.15	
MAD_A	0.32	0.32	0.30	0.30	0.29	0.31	0.32	0.32	
	$\varepsilon=0.1$	$\varepsilon=0.5$	$\varepsilon=1.0$	C1	C2	C3	C4	C5	C6
MAD_M	0.02	0.11	0.26	0.09	0.09	0.09	0.17	0.17	0.17
MAD_A	0.09	0.36	0.47	0.14	0.20	0.24	0.22	0.42	0.62

in which coefficients set to 0 were estimated by 0 and the average MAD values limited to the coefficients set to 0. The values are reported in Table 3 distinguishing with respect to the number of variables and the structure of the coefficients. By observing Table 3 we can state that the method well

Table 3: Recovery of the additive coefficients set to 0: average percentage of times of obtained estimates equal to 0 and average MAD_A values distinguished with respect to the numbers of variables and the structures of the coefficients

	$p=6$		$p=9$		$p=12$		$p=15$	
	%	MAD_A	%	MAD_A	%	MAD_A	%	MAD_A
C1	70.00	0.08	73.61	0.08	62.71	0.08	69.80	0.08
C2	55.00	0.12	61.66	0.08	50.00	0.11	36.33	0.12
C4	81.67	0.12	88.33	0.13	89.58	0.13	91.50	0.12
C5	88.34	0.15	87.22	0.20	84.58	0.31	90.00	0.25

recovered coefficients set to zero, although it tended to underestimate their number. Nonetheless, by inspecting the average MAD values, it is worth mentioning that the estimated values of the coefficients set to zero were in general quite low in absolute sense with the exception of case C5. This suggested that the estimates of coefficients set to zero usually were zero or values close to zero. Furthermore, it must be noted that an abnormal number of coefficients estimated by 0 has been observed in some specific conditions regardless whether their known-in-advance values were 0 or not. More specifically, we sometimes observed a tendency of the method to give only one estimated additive coefficient different from 0 (and equal to t). Such a poor performance was mainly observed (95% of times) when the level of noise added was high (note that, when $\varepsilon = 1.0$, 100%, of noise was added to the data) and the structure of the coefficients were defined by cases C4–C6. We did not extensively analyze this problem, but we found that it depended

on a low value of t and, therefore, can be solved by increasing t.

6. Applications

In this section the results of two applications to real data are discussed. Both data sets refer to the values of three cardiological variables, namely the pulse rate, the systolic pressure and the diastolic pressure observed on a set of patients. The two data sets can be found in Lima-Neto and De Carvalho (2010) and González-Rodríguez et al. (2007), respectively, and, hence, are not reported here.

6.1. Cardiological data set (Lima-Neto and De Carvalho , 2010)

The recorded values take the form of an interval and concern $n = 11$ patients. In order to study the linear dependence of the pulse rate (Y) with respect to the systolic pressure (X_1) and the diastolic pressure (X_2) we applied Lasso-IR. Since the number of units is low, the leave-one-out procedure (i.e. k-fold with $k = 1$) has been considered for determining the tuning parameter t, where $t_{\text{MAX}} = 1.29$ has been obtained according to Section 3.3. The $CV(t)$ values for increasing values of t from 0 to 1.29 with increasing step equal to 0.01 are reported in Figure 1, from which we can see that the optimal value of t was $t_{\text{OPT}} = 0.79$. By setting $t = 0.79$ we got $\widehat{b_M} = \begin{pmatrix} 11.12 & -0.07 & 0.90 \end{pmatrix}'$ and $\widehat{b_A} = \begin{pmatrix} 0 & 0 & -0.79 \end{pmatrix}'$ from which

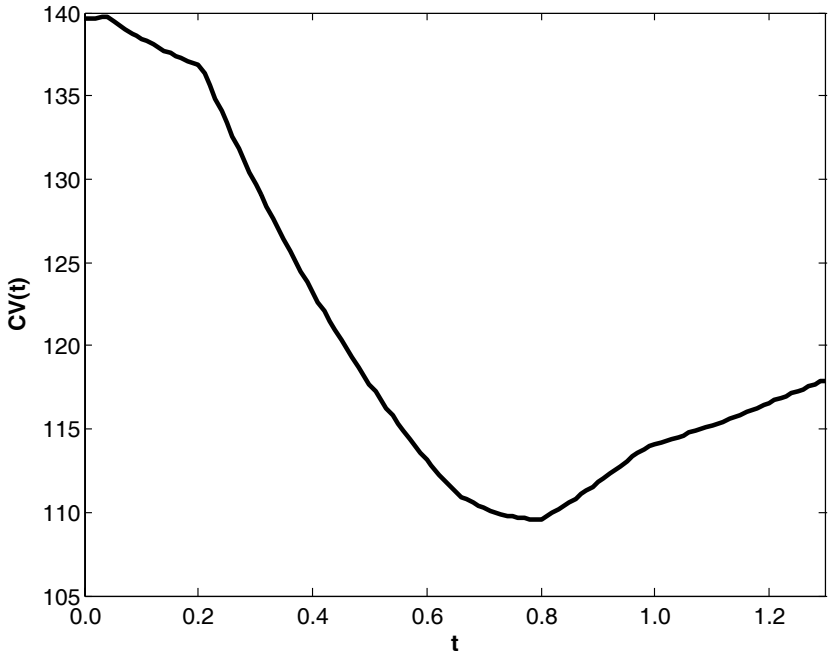
$$\begin{aligned} Y_M &= 11.12 - 0.07X_{1M} + 0.90X_{2M}, \\ Y_R &= 11.12 - 0.07X_{1R} + 0.11X_{2R}. \end{aligned}$$

The value of $\widehat{b_{A2}}$ suggested that there exist different linear relationships between Y and X_2 for the midpoints and for the radii. Conversely, since $\widehat{b_{A1}} = 0$ the same relationship for the midpoints and the radii was found with regard to Y and X_1. The same comment holds for the intercept. By comparing the obtained results with the CCRM ones, we can note that the negative relationship between the pulse rate and the diastolic pressure is not captured by CCRM due to its constraints.

6.2. Cardiological data set (González-Rodríguez et al. , 2007)

The values of the pulse rate and of the systolic and diastolic pressures were observed on a sample of $n = 59$ patients in an hospital in Asturias (Spain) from a population of 3000 patients hospitalized per year (González-Rodríguez

Figure 1: $CV(t)$ values for $t \in [0, 1.29]$ with increasing step = 0.01

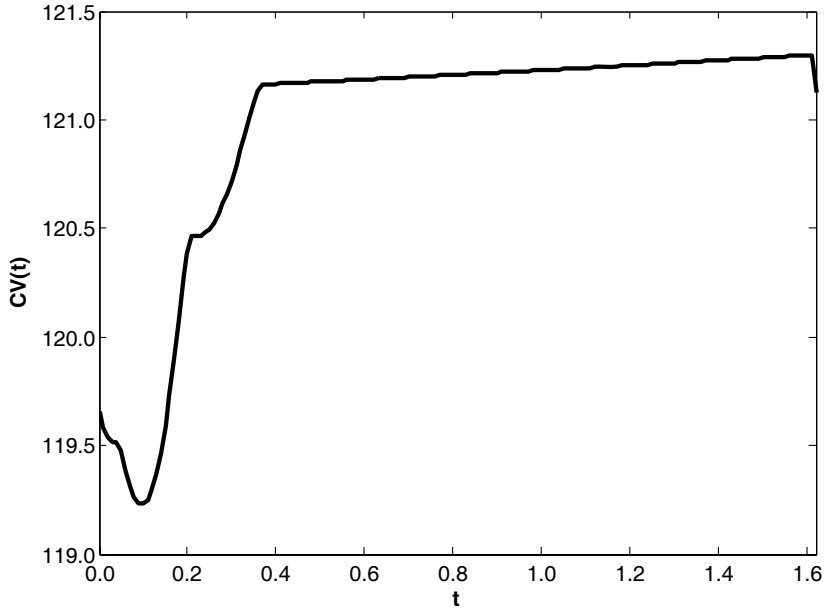


et al. , 2007). In this case, the linear relationship of the diastolic pressure (Y) with respect to the pulse rate (X_1) and the systolic pressure (X_2) was analyzed. We determined the optimal value of t by k-fold cross-validation with $k = 5$ and we estimated the regression coefficients. The optimal value of t was $t_{\text{OPT}} = 0.10$ varying t from 0 to $t_{\text{MAX}} = 1.62$ with increasing step 0.01 (see Figure 2). Setting $t = 0.10$ we got $\widehat{b_M} = \begin{pmatrix} 10.94 & 0.08 & 0.45 \end{pmatrix}'$ and $\widehat{b_A} = \begin{pmatrix} 0 & 0 & -0.10 \end{pmatrix}'$ from which

$$Y_M = 10.94 + 0.08X_{1M} + 0.45X_{2M},$$
$$Y_R = 10.94 + 0.08X_{1R} + 0.35X_{2R}.$$

The Lasso-IR method does not involve probabilistic assumptions. Nonetheless, since the data were a random sample, we were interested in assessing the statistical validity of the obtained regression coefficients. The standard errors of the estimates of the regression parameters were found by a non-parametric bootstrap procedure. $B = 1000$ bootstrap samples of size n were generated.

Figure 2: $CV(t)$ values for $t \in [0, 1.62]$ with increasing step = 0.01



Using t_{OPT} previously obtained, for each bootstrap sample b ($b = 1, \ldots, B$), the regression parameters were estimated. Denoting by $\widehat{b_M^b}$ and $\widehat{b_A^b}$ the estimates of b_M^b and b_A^b , respectively, for the b-th bootstrap sample, the standard errors were computed as

$$\begin{aligned} \widehat{se}\left(\widehat{b_{Mj}}\right) &= \sqrt{\sum_{b=1}^{B}\left(\widehat{b_{Mj}^b} - \overline{\widehat{b_{Mj}}}\right)\Big/ B}, j = 0, \ldots, p, \\ \widehat{se}\left(\widehat{b_{Aj}}\right) &= \sqrt{\sum_{b=1}^{B}\left(\widehat{b_{Aj}^b} - \overline{\widehat{b_{Aj}}}\right)\Big/ B}, j = 0, \ldots, p, \end{aligned} \tag{24}$$

where $\overline{\widehat{b_{Mj}}} = \sum_{b=1}^{B} b_{Mj}^b \Big/ B$ and $\overline{\widehat{b_{Aj}}} = \sum_{b=1}^{B} b_{Aj}^b \Big/ B$, $j = 0, \ldots, p$. We got $\widehat{se}\left(\widehat{b_M}\right) = \begin{pmatrix} 2.78 & 0.10 & 0.05 \end{pmatrix}'$ and $\widehat{se}\left(\widehat{b_A}\right) = \begin{pmatrix} 0.00 & 0.03 & 0.03 \end{pmatrix}'$. By comparing the bootstrap estimates of the standard errors with the corresponding estimates we can gather some evidence that the intercept and the coefficients of X_2 were accurate estimates of the corresponding population

coefficients. On the contrary, less accurate estimates were found for X_1. This suggested that the population linear relationship between Y and X_1 was not clearly detected by the sample data. Note that such results are consistent with the ones by González-Rodríguez et al. (2007), who decided to include only X_2 in the estimated model.

7. Conclusions

In this paper we proposed a tool called Lasso-IR for performing linear regression analysis of interval-valued data. It consists of two regression models, one for the midpoints of the intervals and the other one for the radii. The two regression models are characterized by the same regression coefficients *as much as possible* according to a given criterion based on the Lasso technique. Namely, taking inspiration from González-Rodríguez et al. (2007), we look for a unique set of regression coefficients. A unique set of coefficients is desirable for the sake of parsimony. Unfortunately, this can limit the applicability of the model in some cases. Thus, to make the model more flexible than IALM, the regression coefficients for the radii are allowed to differ to some extent from the ones for the midpoint model. This is achieved by introducing additive coefficients for the radius model such that their sum in absolute value is not bigger than a shrinking parameter t that can be either fixed in advance or chosen by cross-validation techniques. A relevant difference between Lasso-IR and IALM is that the latter is based on interval arithmetic. In the simple linear case, the estimates of the regression coefficients can be found exactly. Instead, in the multiple linear case, a stepwise algorithm can be adopted for finding reasonable estimates. See, for more details, González-Rodríguez et al. (2007). In our method, no distinction is needed between the simple and multiple linear cases. In this respect, Lasso-IR seems to be more related to CCRM proposed by Lima-Neto and De Carvalho (2010). Both the methods consist of two linear regression models for the midpoints and the radii and the optimal regression coefficients are obtained in such a way that a given loss function is minimized and the estimated radii of the response variable are non-negative. Nonetheless, two relevant distinctive features between CCRM and Lasso-IR can be found. The first one is that in CCRM two distinct sets of coefficients for the midpoint and the radius models are assumed, whereas, in Lasso-IR a common set of coefficients is sought as much as possible for the sake of parsimony. The second one concerns the way to constrain the estimated radii of the response to be non-negative. To achieve

it CCRM requires that all the regression coefficients for the radius model are non-negative, whereas in Lasso-IR the non-negativity of the estimated radii is guaranteed without imposing non-negative regression coefficients. It follows that Lasso-IR is more flexible than CCRM because it allows us to handle those situations in which a negative relationship between the radii of the response variable and of the explanatory ones occurs.
In the near future it will be interesting to further investigate the probabilistic properties of Lasso-IR in order to make inference on its results.

Appendix A. Matlab routine of Lasso-IR

```
function[bMh,bAh,bRh,yMh,yRh,lfv,cpt]=lassoir(yM,yR,XM,XR,theta,rs,t);
% Lasso-based interval-valued regression
% Note: lassoir requires lsqlin (MATLAB Optimization Toolbox)
% Input:
% yM: vector of the midpoints of the dependent variable
% yR: vector of the radii of the dependent variable
% XM: matrix of the midpoints of the independent variables
% XR: matrix of the radii of the independent variables
% rs: number of random starts
% t: shrinkage parameter for the lasso penalization term
% Output:
% bMh: vector of the estimated coefficients for the midpoints
% bAh: vector of the estimated additive coefficients
% bRh: vector of the estimated coefficients for the radii
% yMh: vector of the estimated midpoints of the dependent variable
% yRh: vector of the estimated radii of the dependent variable
% lfv: vector of the loss function values
% cpt: vector of the computation times
eps=10^-10; fopt=10^6; [n,p]=size(XM); CONST=ones(2^p,p);
options=optimset('Display','off','LargeScale','off');
for col=1:p;
 CONST(1:2^(p-col),col)=-ones(2^(p-col),1); inc=1;
 while 2^(p-col)+2*inc*2^(p-col)<=2^p;
  CONST(1+2*inc*2^(p-col):2^(p-col)+2*inc*2^(p-col),col)=-ones(2^(p-col),1);
  inc=inc+1;
 end;
end;
```

```
for st=1:rs;
 bA=rand(p,1); if sum(bA)>t; bA=bA/sum(bA)*t; end;
 func=10^6; fold=func+2*eps*func; iter=1; tic;
 while abs(fold-func)>eps*func;
  fold=func;
  y=[yM;(yR-XR*bA)*theta^.5]; X=[XM;XR*theta^.5]; bM=inv(X'*X)*X'*y;
  G=[-XR; CONST]; h=[XR*bM; t*ones(2^p,1)]; bRold=bR;
  [bA,nr,r,ef]=lsqlin(XR,yR-XR*bM,G,h,[],[],[],[],[],options);
  if ef==-2;
   bA=bAold;func=sum((yM-XM*bM).^2)+theta*sum((yR-XR*(bM+bA)).^2);
   fold=func;
  else
   iter=iter+1;
   func=sum((yM-XM*bM).^2)+theta*sum((yR-XR*(bM+bA)).^2);
  end;
 end;
 cpt(st)=toc; lfv(st)=func;
 if func<fopt;
  fopt=func; bMh=aM; bAh=bA; bRh=bMh+bAh;
 end;
end;
yMh=XM*bMh; yRh=XR*bRh;
```

References

Billard, L., Diday, E., 2000. Regression analysis for interval-valued data. In: Kiers, H.A.L., Rasson, J.-P., Groenen, P.J.F., Schader, M., (Eds.), Data Analysis, Classification and Related Methods. Springer-Verlag, Heidelberg, pp. 369–374.

Blanco, A., Corral, N., Colubi, A., González-Rodríguez, G., 2008. On a linear independence test for interval-valued random sets. In: Dubois, D., Lubiano, M.A., Prade, H., Gil, M.A., Grzegorzewski, P., Hryniewics, O., (Eds.), Soft Methods for Handling Variability and Imprecision. Springer-Verlag, Heidelberg, pp. 331–337.

Blanco, A., Corral, N., González-Rodríguez, G., Palacio, A., 2010. On some confidence regions to estimate a linear regression model for interval data. In: Borgelt, C., González-Rodríguez, G., Trutschnig, W., Lubiano M.A.,

Gil, M.A., Grzegorzewski, P., Hryniewicz, O., (Eds.), Combining Soft Computing and Statistical Methods in Data Analysis. Springer-Verlag, Heidelberg, 263–271.

Domingues, M.A.O., de Souza, R.M.C.R., Cysneiros, F.J.A., 2010. A robust method for linear regression of symbolic interval data. Pattern Recognition Letters 31, 1991–1996.

Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap. Chapman & Hall, New York.

Gil, M.A., González-Rodríguez, G., Colubi, A., Montenegro, M., 2007. Testing linear independence in linear models with interval-valued data. Computational Statistics and Data Analysis 51, 3002–3015.

Gill, P.E., Murray, W., Wright, M.H., 1981. Practical Optimization. Academic Press, London.

González-Rodríguez, G., Colubi, A., Coppi, R., Giordani, P., 2006. On the estimation of linear models with interval-valued data. In: Rizzi, A., Vichi, M., (Eds.), Proceedings in Computational Statistics. Physica-Verlag, New York, pp. 697–704.

González-Rodríguez, G., Blanco, A., Corral, N., Colubi, A., 2007. Least squares estimation of linear regression models for convex compact random sets. Advances in Data Analysis and Classification 1, 67–81.

González-Rodríguez, G., Blanco, A., Colubi, A., Lubiano, M.A., 2009. Estimation of a simple linear regression model for fuzzy random variables. Fuzzy sets and Systems 160, 357–370.

Lawson, C.L, Hanson, R.J., 1995. Solving Least Squares Problems, (Classics in Applied Mathematics, Vol. 15). SIAM, Philadelphia.

Lima-Neto, E.A., De Carvalho, F.A.T., 2008. Centre and range method to fitting a linear regression model on symbolic interval data. Computational Statistics and Data Analysis 52, 1500–1515.

Lima-Neto, E.A., De Carvalho, F.A.T., 2010. Constrained linear regression models for symbolic interval-valued variables. Computational Statistics and Data Analysis 54, 333–347.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society - Series B 58, 267–288.

Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A., 2009. A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Information Sciences, 179 3964–3972.