Information Sciences 454–455 (2018) 419–442

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An exponential-type kernel robust regression model for interval-valued variables

Eufrásio de A. Lima Neto[a], Francisco de A.T. de Carvalho[b,*]

[a] Universidade Federal da Paraíba, Departamento de Estatística, Cidade Universitária, 58051-900 João Pessoa, PB, Brazil
[b] Centro de Informatica, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes s/n – Cidade Universitaria, CEP 50740-560 Recife, PE, Brazil

ARTICLE INFO

Article history:
Received 31 October 2017
Revised 29 March 2018
Accepted 2 May 2018
Available online 3 May 2018

Keywords:
Interval-valued variables
Exponential-type kernel functions
Robust regression models
Width hyper-parameter estimators
Outliers

ABSTRACT

The presence of outliers is very common in regression problems and the use of robust regression methods is strongly recommended such that the bad fitted observations not affect the parameter estimates of the model. Interval-valued variables are becoming common in data analysis problems since this type of data represents either the uncertainty existing in an error measurement or the natural variability present in the data. Regarding the presence of outliers in interval-valued data sets, few robust regression methods have been proposed in literature. This paper introduces a new robust regression method for interval-valued variables that penalizes the presence of outliers in the midpoints and/or in the ranges of interval-valued observations through the use of exponential-type kernel functions. Thus, the weight given to the midpoint and range of each interval-valued observation is updated at each iteration in order to optimize a suitable objective function. The convergence of the parameter estimation algorithm is guaranteed with a low computational cost. A comparative study between the proposed method against some previous robust regression approaches for interval-valued variables is also considered. The performance of these methods are evaluated based on the bias and mean squared error (MSE) of the parameter estimates for the midpoints and ranges of the intervals, considering synthetic data sets with X-space outliers, Y-space outliers and leverage points, different sample sizes and percentage of outliers in a Monte Carlo framework. The results suggest that the proposed approach presents a competitive performance (or best), in comparison with the previous approaches, on interval-valued outliers scenarios that are comparable to those found in practices. Applications to real interval-valued data sets corroborates the usefulness of the proposed method.

1. Introduction

Understand the relationship between a set of variables is a common task to solve practical problems in many areas. The regression analysis represents an interesting tool that can be used for this aim, representing an important research field with new methods in constant development.

* Corresponding author.
E-mail addresses: fatc@cin.ufpe.br, francisco.carvalho@pq.cnpq.br (F.d.A.T. de Carvalho).

https://doi.org/10.1016/j.ins.2018.05.008
0020-0255/

Nowadays information systems allow to collect and storage data with low cost and faster. Moreover, massive data sets have been generated in a easy way in many areas like Economy, Meteorology, Business, Telecommunications, Biology, among others. These data sets tend to be released in an aggregated format due to confidentiality reasons or because the interest of study is not the individual unit but a group of units. Thus, the researcher does not analyze a classical data set with single values in the real line, but a complex data set aggregated with new types of data, like interval-valued data, that offer information on the lower and upper bound of the variable of interest.

Therefore, interval datasets are becoming common in data analysis problems. This type of data can represents the imprecision and/or uncertainty existing in an error measurement but also can represents the natural variability present in the data. Some examples of interval variables are technical specifications, temperatures in meteorological stations and daily stock prices. In this context, statistical tools to analyze interval variables are very much required.

Interval data representing imprecision or uncertainty has been mainly dealt with by means of fuzzy-valued data, with various research developing regression models for fuzzy-valued variables. In this framework, two main approaches are present in literature, fuzzy models using linear and non-linear programming [8,22,33,34,36,41,50] and fuzzy models using least squares method [9,10,12,15–18,46].

This paper is concerned with interval-valued data representing natural variability in the data, which have been mainly treated in Symbolic Data Analysis (SDA) field [3,4] with various research addressed to regression models for interval-valued variables taking into account parametric and nonparametric regression algorithms as well as linear and nonlinear relationships.

Regarding the SDA field, many approaches have been proposed in order to consider a regression model for interval-valued variables. Some of these approaches represent extensions or modifications of regression models for real-valued data. A seminal paper was proposed by Billard and Diday [2] considering a linear regression model for interval-valued variable. Other works were proposed in the same direction, most of them, taking into account a parametric linear relationship between a response interval-valued variable Z and a set of explanatory interval-valued variables $W_1, \ldots, W_p$, represented in terms of the midpoints (centers) and half-ranges (radius) ([38,49,52] and the references therein). The use of constraints was also considered by some authors in order to guarantee that the radius is greater or equal than zero and, consequently, the lower bound is less or equal than the upper bound ([23–25,39] and the references therein). However, the use of constraints limits the domain of the objective function, penalizing the parameter estimates obtained in the optimization process. Other regression methods for interval-valued data have taken into account a probabilistic support for the response interval-valued variable Z, allowing the use of inference techniques over the parameter estimates [1,5,37,40]. More recently, semi-parametric and nonparametric regression models for interval-valued variables were proposed by Refs. [20], [29], [30] and [51]. Besides, a model where the intervals are represented by quantile functions and that considers the distribution Uniform or Symmetric Triangular within the intervals has been proposed by Ref. [13].

Robust regression attempts to cope with outliers and leverage points [26,27,31,44,45,53]. The regression outliers are data that move away from the linear model pattern of the majority of the observations and the use of a non-robust techniques typically leads to biased inferences. Concerning interval-valued data, some contributions were proposed related to robust regression models for interval-valued variables. Ref. [14] presented the symbolic symmetrical linear regression model for interval variables that take into account a Student-t distribution for the midpoints of interval and a Gaussian distribution for the ranges. Ref. [19] considers two independent classic robust regressions over the midpoints and ranges of the intervals. The regression outliers (in the midpoints and ranges) are penalized according to the Tukey's bi-weight function. Ref. [21] adapted the technique of quantile regression for interval-valued variables.

The use of positive definite kernels has become popular in the computational intelligence community. The idea of using exponential-type kernel functions to measure the similarity between two objects have been successfully applied in computer vision, signal processing, clustering, pattern classification, among others, with a large literature on the family of kernel-based algorithms ([11,47,48] and the references contained therein). Recently, Ref. [7] proposed a robust regression method for real-valued data based on exponential-type kernel function (called ETKRR method) which presented a competitive performance (or best) in comparison with the well established classical robust linear regression models like L1-regression, MM-Estimator regression, weighted least squares, among others.

This paper introduces a robust regression method for interval-valued variables, hereafter named iETKRR (Exponential-type kernel robust regression for interval-valued variables). It extends to interval-valued variables the robust regression model for real-valued variables proposed by Ref. [7]. Its main contributions are as follows:

- the iETKKR provides a new objective function that has two terms aiming to take into account the informations provided either by the center and the radius of the intervals or by the lower and upper boundaries of the intervals. Therefore, the new objective function is suitable to manage interval-valued data.
- Besides, the proposed method allows to combine different hyper-parameter estimators, respectively, one for the center and one for the radius (or one for the lower bound and one for the upper bound), and thus provides more flexibility and robustness to treat the different outlier's types present in interval-valued data sets.

The iETKRR method re-weights the interval observations based on exponential-type kernel functions, in such a way that the weight assigned to an interval outlier observation is as small as possible, considering an iterative process to minimize a suitable objective function. The weighting is provided by calculating the similarities between the observed and predicted

values for the midpoint an range, respectively, of the response variables and updating it at each iteration in order to optimize the objective function. The convergence of the estimation algorithm is guaranteed with a low computational cost.

A comparative study between the iETKRR method and the robust regression approaches for interval-valued variables present in literature [14,19,21] is considered. These methods will be evaluated in terms of the bias and mean squared error (MSE) of the parameter estimates taking into account X-space outliers, Y-space outliers, leverage points, different sample sizes and percentage of outliers in the sample, representing a total of 138 different configurations in a Monte Carlo simulation framework with 10,000 replications. Applications on real interval-valued data sets also illustrate the usefulness of the iETKRR method.

The paper is organized as follows: Section 2 reviews some concepts about exponential-type kernel functions and presents the iETKRR method for interval-valued variable as well as the parameter estimate algorithm. Section 3 exhibits the Monte Carlo experiments that evaluates the convergence of the parameter estimation algorithm, compares the iETKRR method with the existing robust regression methods for interval-valued variables and discuss the results obtained in the numerical analysis. Section 4 brings the applications to real interval-valued data sets and Section 5 gives the concluding remarks.

2. Exponential-type kernel robust regression for interval-valued variables (iETKRR)

This section reviews some concepts about exponential-type kernel functions, presents the iETKRR method for interval-valued variable and provides the corresponding parameter estimation algorithm.

2.1. Some important concepts about kernel functions

Hereafter we briefly recall the basic theory about kernel functions. Let $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a non-empty set where $\mathbf{v}_i \in \mathbb{R}^p$. A function $\mathcal{K} : V \times V \to \mathbb{R}$ is called a *positive definite* or *Mercer kernel* [32] if and only if $\mathcal{K}$ is symmetric (i.e. $\mathcal{K}(\mathbf{v}_i, \mathbf{v}_k) = \mathcal{K}(\mathbf{v}_k, \mathbf{v}_i)$) and the following inequality holds:

$$\sum_{i=1}^{n}\sum_{k=1}^{n} c_i c_k \mathcal{K}(\mathbf{v}_i, \mathbf{v}_k) \geq 0, \quad \forall n \geq 2,$$

where $c_r \in \mathbb{R}$, $\forall r = 1, \ldots, n$.

Let $\Phi : V \to \mathcal{F}$ be a non-linear mapping from the input space V to a high dimensional feature space $\mathcal{F}$. By applying the mapping Φ, the dot product $\mathbf{v}_i^\top \mathbf{v}_k$ in the input space is mapped to $\Phi(\mathbf{v}_i)^\top \Phi(\mathbf{v}_k)$ in the feature space. The key idea supporting the kernel algorithms is that the non-linear mapping Φ does not need to be explicitly specified because each Mercer kernel can be expressed as $\mathcal{K}(\mathbf{v}_i, \mathbf{v}_k) = \Phi(\mathbf{v}_i)^\top \Phi(\mathbf{v}_k)$ [35].

One of the most relevant aspects in applications is that it is possible to compute Euclidean distances in $\mathcal{F}$ without knowing explicitly Φ. This can be done using the so called *distance kernel trick* [35]:

$$\begin{aligned} ||\Phi(\mathbf{v}_i) - \Phi(\mathbf{v}_k)||^2 &= (\Phi(\mathbf{v}_i) - \Phi(\mathbf{v}_k))^\top (\Phi(\mathbf{v}_i) - \Phi(\mathbf{v}_k)) \\ &= \Phi(\mathbf{v}_i)^\top \Phi(\mathbf{v}_i) - 2\Phi(\mathbf{v}_i)^\top \Phi(\mathbf{v}_k) + \Phi(\mathbf{v}_k)^\top \Phi(\mathbf{v}_k) \\ &= \mathcal{K}(\mathbf{v}_i, \mathbf{v}_i) - 2\mathcal{K}(\mathbf{v}_i, \mathbf{v}_k) + \mathcal{K}(\mathbf{v}_k, \mathbf{v}_k). \end{aligned} \tag{1}$$

In this paper, we consider the exponential-type Gaussian kernel function, the most commonly used in the literature, for the iETKRR model:

$$\mathcal{K}_{\mathcal{G}}(\mathbf{v}_i, \mathbf{v}_k) = \exp\left\{-\frac{||\mathbf{v}_i - \mathbf{v}_k||^2}{2\gamma^2}\right\}, \tag{2}$$

where γ^2 is the width hyper-parameter of the Gaussian kernel.

The width hyper-parameter γ^2 plays an important role in the computation of the Gaussian kernel function and must be carefully adjusted to the problem in question. Overestimated values of γ^2 lead to a linear projection and higher dimension in the behavior of the exponential function. On the other hand, for underestimated values of γ^2, the function will lack regularization and the decision boundary will be highly sensitive to noise in data.

2.2. The iETKRR method

Let $E = \{e_1, \ldots, e_n\}$ be a set of examples that are described by $p+1$ interval-valued variables $Z, W_1, \ldots, W_p$. The interval-valued variable Z is a dependent variable and it is related to a set of interval-valued variables W_j ($j = 1, 2, \ldots, p$), known as independent variables. Each example $e_i \in E$ ($1 \leq i \leq n$) is represented by an interval-valued feature vector $(\mathbf{w}_i, z_i)$, with $\mathbf{w}_i = (w_{i1}, \ldots, w_{ip})$, where $w_{ij} = [a_{ij}, b_{ij}] \in \Im = \{[a, b] : a, b \in \Re, a \leq b\}$ ($1 \leq j \leq p$) and $z_i = [z_{Li}, z_{Ui}] \in \Im$ are the observed values of W_j and Z, respectively. Now, let Y_1, X_{1j} and Y_2, X_{2j} ($j = 1, 2, \ldots, p$) be quantitative variables that represent the lower and upper bounds or the midpoints (centers) and the half-ranges (radius) of the intervals defined by the interval-valued variables Z and W_j, respectively.

In case where the quantitative variables, Y_1, X_{1j} and Y_2, X_{2j} ($j = 1, 2, \ldots, p$) represent, respectively, the lower and upper boundaries of the interval variables Z and W_j, we denote $Y_1 = Y^L$, $X_{1j} = X_j^L$, $Y_2 = Y^U$, $X_{2j} = X_j^U$, and each example $e_i \in E$

$(i = 1, \ldots, n)$ will be represented by two vectors: $(\mathbf{x}_i^L, y_i^L)$ and $(\mathbf{x}_i^U, y_i^U)$, with $\mathbf{x}_i^L = (x_{i1}^L, \ldots, x_{ip}^L)$ and $\mathbf{x}_i^U = (x_{i1}^U, \ldots, x_{ip}^U)$, where $x_{ij}^L = a_{ij}$, $x_{ij}^U = b_{ij}$, $y_i^L = z_{Li}$ and $y_i^U = z_{Ui}$ are the observed values of the quantitative variables X_j^L, X_j^U, Y^L and Y^U, respectively.

In the same way, for the case where the quantitative variables, Y_1, X_{1j} and Y_2, X_{2j} $(1 \leq j \leq p)$ represent, respectively, the midpoint and half-range of the interval variables Z and W_j, we denote $Y_1 = Y^c$, $X_{1j} = X_j^c$, $Y_2 = Y^r$, $X_{2j} = X_j^r$, and each example $e_i \in E$ $(1 \leq i \leq n)$ will be represented by the vectors $(\mathbf{x}_i^c, y_i^c)$ and $(\mathbf{x}_i^r, y_i^r)$, with $\mathbf{x}_i^c = (x_{i1}^c, \ldots, x_{ip}^c)$ and $\mathbf{x}_i^r = (x_{i1}^r, \ldots, x_{ip}^r)$, where $x_{ij}^c = (a_{ij} + b_{ij})/2$, $x_{ij}^r = (b_{ij} - a_{ij})/2$, $y_i^c = (z_{Li} + z_{Ui})/2$ and $y_i^r = (z_{Ui} - z_{Li})/2$ are the observed values of the variables X_j^c, X_j^r, Y^c and Y^r, respectively.

Let be the response feature vector $\mathbf{y}_i = (y_{1i}, y_{2i})$ and its corresponding mean vector $\boldsymbol{\mu}_i = (\mu_{1i}, \mu_{2i})$, where the subscripts 1 and 2 represent the lower bound (or the midpoint) and the upper bound (or the half-range) of the intervals, respectively, with $\mu_{1i} = \mathbf{x}_{1i}^\top \boldsymbol{\beta}_1 = \beta_{10} + \sum_{j=1}^p \beta_{1j} x_{1ij}$ and $\mu_{2i} = \mathbf{x}_{2i}^\top \boldsymbol{\beta}_2 = \beta_{20} + \sum_{j=1}^p \beta_{1j} x_{2ij}$ representing the mean of the response variables Y_1 and Y_2, respectively, with $\mathbf{x}_{1i} = (1, x_{1i1}, \ldots, x_{1ip})$ and $\mathbf{x}_{2i} = (1, x_{2i1}, \ldots, x_{2ip})$ related to the values of the set of the explanatory variables X_{1j} and X_{2j} $(1 \leq j \leq p)$, respectively, and with $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \ldots, \beta_{1p})$ and $\boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}, \ldots, \beta_{2p})$ representing the respective vector of parameters to be estimated.

The iETKRR method manages the response variables Y_1 and Y_2 based on the kernel framework and the parameter estimation is guided by the minimization of an objective function taking into account the distance kernel trick [35]. In this way, the sum of squared errors is defined not in the original space but in a high dimensional space $\mathcal{F}$ through non-linear mapping Φ_1 and Φ_2 applied, respectively, on the observed response variables y_{1i} and y_{2i} and its corresponding mean μ_{1i} and μ_{2i} Based on the distance kernel trick, the sum of squared errors defined in a high dimensional space is computed in the original space as follows:

$$\begin{aligned} S &= \sum_{i=1}^n (||\Phi_1(y_{1i}) - \Phi_1(\mu_{1i})||^2 + ||\Phi_2(y_{2i}) - \Phi_2(\mu_{2i})||^2) \\ &= \sum_{i=1}^n [\mathcal{K}_1(y_{1i}, y_{1i}) - 2\mathcal{K}_1(y_{1i}, \mu_{1i}) + \mathcal{K}_1(\mu_{1i}, \mu_{1i})] \\ &\quad + [\mathcal{K}_2(y_{2i}, y_{2i}) - 2\mathcal{K}_2(y_{2i}, \mu_{2i}) + \mathcal{K}_2(\mu_{2i}, \mu_{2i})]. \end{aligned} \tag{3}$$

Let $\mathcal{K}_\mathcal{G}$ be the Gaussian kernel function. Then, the following properties are verified: $\mathcal{K}_{1\mathcal{G}}(y_{1i}, y_{1i}) = \mathcal{K}_{2\mathcal{G}}(y_{2i}, y_{2i}) = 1$, $\mathcal{K}_{1\mathcal{G}}(\mu_{1i}, \mu_{1i}) = \mathcal{K}_{2\mathcal{G}}(\mu_{1i}, \mu_{1i}) = 1 \quad (1 \leq i \leq n)$ and the functional (3) can be rewritten as

$$\begin{aligned} S &= \sum_{i=1}^n 2[2 - \{\mathcal{K}_{1\mathcal{G}}(y_{1i}, \mu_{1i}) + \mathcal{K}_{2\mathcal{G}}(y_{2i}, \mu_{2i})\}], \\ &= \sum_{i=1}^n 2\left[2 - \left\{\exp\left[-\frac{1}{2}\left(\frac{y_{1i} - \mu_{1i}}{\gamma_1}\right)^2\right] + \exp\left[-\frac{1}{2}\left(\frac{y_{2i} - \mu_{2i}}{\gamma_2}\right)^2\right]\right\}\right], \end{aligned} \tag{4}$$

with $\gamma_1 > 0$ and $\gamma_2 > 0$. Thus, as close we have the terms y_{1i} and μ_{1i} (y_{2i} and μ_{2i}), $\forall i = 1, \ldots, n$, as close to zero becomes the objective function for the iETKRR model.

We can observe that the objective function of Eq. (4) is suitable to manage interval-valued data because it is able to take into account the informations provided either by the center and the radius of the intervals or the lower and upper boundaries of the intervals. Besides, this objective function allows to combine different hyper-parameter estimators either on the center and on the radius of the intervals or on the lower and upper boundaries of the intervals, and thus it provides more flexibility and robustness to the model to treat the different outliers types present in interval-valued data sets.

Differentiating Eq. (4) with respect to each element of the vectors of parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ we obtain, respectively, the set of normal equations for the iETKRR model, given by

$$\begin{aligned} \frac{\partial S}{\partial \beta_{10}} &= -2\sum_{i=1}^n \mathcal{K}_{1\mathcal{G}}(y_{1i}, \hat{\mu}_{1i}) \frac{(y_{1i} - \hat{\mu}_{1i})}{\gamma_1} = 0, \\ \frac{\partial S}{\partial \beta_{1j}} &= -2\sum_{i=1}^n \mathcal{K}_{1\mathcal{G}}(y_{1i}, \hat{\mu}_{1i}) \frac{x_{1ij}(y_{1i} - \hat{\mu}_{1i})}{\gamma_1} = 0, \ (1 \leq j \leq p) \\ \frac{\partial S}{\partial \beta_{20}} &= -2\sum_{i=1}^n \mathcal{K}_{2\mathcal{G}}(y_{2i}, \hat{\mu}_{2i}) \frac{(y_{2i} - \hat{\mu}_{2i})}{\gamma_2} = 0, \\ \frac{\partial S}{\partial \beta_{2j}} &= -2\sum_{i=1}^n \mathcal{K}_{2\mathcal{G}}(y_{2i}, \hat{\mu}_{2i}) \frac{x_{2ij}(y_{2i} - \hat{\mu}_{2i})}{\gamma_2} = 0, \ (1 \leq j \leq p) \end{aligned} \tag{5}$$

The solution of the normal equation system (5) is obtained through an iterative re-weighting unconstrained least squares process, based on the regression model

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \tag{6}$$

where

$$\mathbf{y}^* = \begin{bmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \end{bmatrix},\ \mathbf{X}^* = \begin{bmatrix} \mathbf{X}_1^* & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2^* \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}, \text{and } \boldsymbol{\epsilon}^* = \begin{bmatrix} \boldsymbol{\epsilon}_1^* \\ \boldsymbol{\epsilon}_2^* \end{bmatrix},$$

with $\mathbf{y}_1^* = \mathbf{K}_1^{1/2}\mathbf{y}_1$ and $\mathbf{y}_2^* = \mathbf{K}_2^{1/2}\mathbf{y}_2$ being two $n \times 1$ weighted vectors of the response variables Y_1 and Y_2, respectively, $\mathbf{X}_1^* = \mathbf{K}_1^{1/2}\mathbf{X}_1$ and $\mathbf{X}_2^* = \mathbf{K}_2^{1/2}\mathbf{X}_2$ being two $n \times p$ weighted matrices of the explanatory variables X_1 and X_2, respectively, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ being the respective vectors of the regression parameters, $\boldsymbol{\epsilon}_1^* = \mathbf{K}_1^{1/2}\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2^* = \mathbf{K}_2^{1/2}\boldsymbol{\epsilon}_2$ being two $n \times 1$ weighted vectors of independent errors, and $\mathbf{K_1} = \text{diag}(k_{1_1}, k_{1_2}, \ldots, k_{1_n})$ and $\mathbf{K_2} = \text{diag}(k_{2_1}, k_{2_2}, \ldots, k_{2_n})$ being two $n \times n$ diagonal kernel weight matrices with k_{1_i} and k_{2_i} defined as $k_{1_i} = \mathcal{K}_{1\mathcal{G}}(y_{1i}, \mu_{1i})$ and $k_{2_i} = \mathcal{K}_{2\mathcal{G}}(y_{2i}, \mu_{2i})$ $(1 \leq i \leq n)$.

Note that $0 \leq \mathcal{K}_{1\mathcal{G}}(y_{1i}, \mu_{1i}) \leq 1$ and $0 \leq \mathcal{K}_{2\mathcal{G}}(y_{2i}, \mu_{2i}) \leq 1$. Therefore, observations with large residual will have a small weight for the parameter estimates.

Finally, given a new observation e described by an interval-valued feature vector $(\mathbf{w}, z)$, the value $z = [z_L, z_U]$ of Z will be predicted, considering the rule proposed by Xu [52], as follow:

- if e is described in terms of the lower and upper interval bounds, then $\hat{z}_L = \min\{\mathbf{x}_1^\top \hat{\boldsymbol{\beta}}_1, \mathbf{x}_2^\top \hat{\boldsymbol{\beta}}_2\}$ and $\hat{z}_U = \max\{\mathbf{x}_1^\top \hat{\boldsymbol{\beta}}_1, \mathbf{x}_2^\top \hat{\boldsymbol{\beta}}_2\}$, where $\mathbf{x}_1^\top = (1, x_{11}, \ldots, x_{1p})$ and $\mathbf{x}_2^\top = (1, x_{21}, \ldots, x_{2p})$, $x_{1j} = a_j$ and $x_{2j} = b_j$ $(1 \leq j \leq p)$;
- if e is described in terms of the midpoint and half-range of the interval, then $\hat{z}_L = \min\{\hat{y}^c - \hat{y}^r, \hat{y}^c + \hat{y}^r\}$ and $\hat{z}_U = \max\{\hat{y}^c - \hat{y}^r, \hat{y}^c + \hat{y}^r\}$, where: $\hat{y}^c = \mathbf{x}_1^\top \hat{\boldsymbol{\beta}}_\mathbf{1}$ and $\hat{y}^r = \mathbf{x}_2^\top \hat{\boldsymbol{\beta}}_\mathbf{2}$, with $\mathbf{x}_1^\top = (1, x_{11}, \ldots, x_{1p})$ and $\mathbf{x}_2^\top = (1, x_{21}, \ldots, x_{2p})$, $x_{1j} = \frac{a_j + b_j}{2}$ and $x_{2j} = \frac{b_j - a_j}{2}$ $(1 \leq j \leq p)$.

2.3. Estimation algorithm

The estimation of the vectors of parameters $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ can be performed through the iterative algorithm presented hereafter.

Input : $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{y_1}$, $\mathbf{y_2}$, a tolerance limit ϵ, a maximum number of iterations T, the kernel width hyper-parameters γ_1 and γ_2
Output: $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\mu}}_1 = (\mu_{11}, \ldots, \mu_{1n})^\top$, $\hat{\boldsymbol{\mu}}_2 = (\mu_{21}, \ldots, \mu_{2n})^\top$

Initialization:
Set $t = 0$ and $\mathbf{K_1}^{(0)} = \mathbf{K_2}^{(0)} = \mathbf{I}_n$; // $\mathbf{I}_n$ is an $n \times n$ identity matrix

Compute $\hat{\boldsymbol{\beta}}_1^{(0)}$and $\hat{\boldsymbol{\beta}}_2^{(0)}$; // started values for $\hat{\boldsymbol{\beta}}_\mathbf{1}$ and $\hat{\boldsymbol{\beta}}_\mathbf{2}$

Compute $\hat{\boldsymbol{\mu}}_\mathbf{1}^{(0)} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_\mathbf{1}^{(0)}$ and $\hat{\boldsymbol{\mu}}_\mathbf{2}^{(0)} = \mathbf{X}_2\hat{\boldsymbol{\beta}}_\mathbf{2}^{(0)}$;

Compute $S^{(0)}$according to Eq. (4);

Fitting steps:
repeat

Set $t = t + 1$;

$\mathbf{K}_1^{(t)} = \text{diag}\{k_{11}^{(t)}, \ldots, k_{1n}^{(t)}\}$, $\mathbf{K}_2^{(t)} = \text{diag}\{k_{21}^{(t)}, \ldots, k_{2n}^{(t)}\}$, with $k_{\theta i}^{(t)} = \mathcal{K}_{\theta\mathcal{G}}(y_{\theta i}, \hat{\mu}_{\theta i}^{(t-1)})$ and $\hat{\mu}_{\theta i}^{(t-1)} = \mathbf{x}_{\theta i}^\top \hat{\boldsymbol{\beta}}_\theta^{(t-1)}$ $(\theta = \{1, 2\})$;

Compute $\hat{\boldsymbol{\beta}}_\mathbf{1}^{(t)} = (\mathbf{X}_1^\top \mathbf{K}_1^{(t)} \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{K}_1^{(t)} \mathbf{y}_1$ and $\hat{\boldsymbol{\beta}}_\mathbf{2}^{(t)} = (\mathbf{X}_2^\top \mathbf{K}_2^{(t)} \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{K}_2^{(t)} \mathbf{y}_2$;

Compute $\hat{\boldsymbol{\mu}}_\mathbf{1}^{(t)} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_\mathbf{1}^{(t)}$ and $\hat{\boldsymbol{\mu}}_\mathbf{2}^{(t)} = \mathbf{X}_2\hat{\boldsymbol{\beta}}_\mathbf{2}^{(t)}$;

Compute $S^{(t)}$according to Eq. (4);

until $|S^{(t)} - S^{(t-1)}| \leq \varepsilon$ *or* $t \geq T$;

The algorithm starts from an initial solution for $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ and alternates between the estimation of the regression model (coefficients and predicted responses) and the estimation of the weights (of each observation) until the convergence to a local minimum of the objective function.

As the kernel weight matrices need to be computed only once at each iteration, the complexity of estimate the coefficients vectors $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ is $\mathcal{O}(2n(p+1))$, where n is the number of interval observations and p is the number of interval-valued variables. At each iteration t, the kernel weight matrices require $2n$ kernel function evaluations $k_{\theta_i}^{(t)} = \mathcal{K}_{\theta\mathcal{G}}(y_{\theta_i}, \hat{\mu}_{\theta_i}^{(t-1)})$ $(1 \leq i \leq n; \theta = \{1, 2\})$ and, since $\hat{\mu}_{1_i}^{(t-1)} = \beta_{1_0}^{(t-1)} + \beta_{1_1}^{(t-1)} x_{1_{i1}} + \cdots + \beta_{1_p}^{(t-1)} x_{1_{ip}}$ and $\hat{\mu}_{2_i}^{(t-1)} = \beta_{2_0}^{(t-1)} +$

$\beta_{2_1}^{(t-1)} x_{2_{i1}} + \cdots + \beta_{2_p}^{(t-1)} x_{2_{ip}} \quad (1 \le i \le n)$, the computational complexity of the iETKRR algorithm is $\mathcal{O}(2n(p+1))$ for a single iteration.

2.4. Convergence properties

The convergence of the proposed algorithm can be studied from two series: $w_t = \hat{\boldsymbol{\mu}}^{(t)} = [\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2]^\top$ and $z_t = S(w_t) = S(\hat{\boldsymbol{\mu}}^{(t)})$ $(t = 0, 1, \ldots)$. From an initial term $w_0 = \hat{\boldsymbol{\mu}}^{(0)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(0)}$, where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix}, \ \hat{\boldsymbol{\beta}}^{(0)} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1^{(0)} \\ \hat{\boldsymbol{\beta}}_2^{(0)} \end{bmatrix},$$

$\hat{\boldsymbol{\beta}}^{(0)}$ represents the start value for $\hat{\boldsymbol{\beta}}$ and $\mathbf{K_1}^{(0)} = \mathbf{K_2}^{(0)} = \mathbf{I}_n$, the algorithm computes iteratively $\hat{\boldsymbol{\mu}}^{(t)}$ until the convergence when the criterion S achieves a stationary value.

Proposition 2.1. *The series $z_t = S(w_t)$ decreases at each iteration and converges.*

Proof. The proof of Proposition 2.1 follows a similar scheme of that developed in Ref. [7]. □

Proposition 2.2. *The series $w_t = \hat{\boldsymbol{\mu}}^{(t)}$ converges.*

Proof. The proof of Proposition 2.2 follows a similar scheme of that developed in Ref. [7]. □

3. Monte Carlo experiments

A Monte Carlo simulation study is performed aiming to compare the iETKRR method against the robust interval regression methods IRR [19], IQR [21] and SSLR [14]. All experiments were implemented in the R language [43] and performed on the same machine (OS: Windows 7 Professional 64-bits, Memory: 16 GiB, Processor: Intel Core i7-X990 CPU @ 3.47 GHz). The code with the iETKRR parameter estimation algorithm can be requested to the authors.

3.1. Experimental settings

The parameter estimates for the iETKRR method were obtained according to the algorithm presented in Section 2.3 with a tolerance limit $\epsilon = 1 \times 10^{-10}$ and a maximum number of iterations $T = 100$.

The simulation scenarios were built in terms of midpoints and ranges of the intervals. Let $E = \{e_1, \ldots, e_n\}$ be a set of examples that are described by 2 interval-valued variables Z and W_1, respectively, dependent and independent variable. Each example $e_i \in E$ $(1 \le i \le n)$ is represented by an interval-valued feature vector (w_{i1}, z_i), where $w_{i1} = [a_{i1}, b_{i1}] \in \Im = \{[a, b] : a, b \in \Re, a \le b\}$ and $z_i = [z_{Li}, z_{Ui}] \in \Im$ are the observed values of W_1 and Z, respectively.

Let Y_1, X_{11} and Y_2, X_{21} be quantitative variables that represent the midpoints (centers) and the half-ranges (radius) of the intervals defined by the interval-valued variables Z and W_1, respectively. Therefore, hereafter we assume that $Y_1 = Y^c$, $X_{11} = X^c$, $Y_2 = Y^r$, $X_{21} = X^r$, and each example $e_i \in E$ $(1 \le i \le n)$ will be represented by the vectors (x_i^c, y_i^c) and (x_i^r, y_i^r), where $x_i^c = (a_{i1} + b_{i1})/2$, $x_i^r = (b_{i1} - a_{i1})/2$, $y_i^c = (z_{Li} + z_{Ui})/2$ and $y_i^r = (z_{Ui} - z_{Li})/2$ are the observed values of the real-valued variables X^c, X^r, Y^c and Y^r, respectively.

The simulation settings reflect scenarios found in real problems. We fixed a percentage of clean data and added the complementary percentage of outliers. The clean interval-valued data for the interval explanatory variable X came from two uniform distributions. The first one generates the values of the midpoints $X^c \sim U[5, 10]$ while the second one generates the values of the ranges $X^r \sim U[1, 2]$. The ith clean interval-valued data $z_i = [z_{Li}, z_{Ui}] = [y_i^c - y_i^r/2, y_i^c + y_i^r/2]$ for the response variable Z is generated according to the two linear models $y_i^c = 11 + 2x_i^c + \epsilon_i^c$ and $y_i^r = 1 + 1x_i^r + \epsilon_i^r$, where $\epsilon_i^c \sim N(0, 0.5^2)$ and $\epsilon_i^r \sim N(0, 0.1^2)$ are the random errors terms for the midpoint and ranges, respectively, with Gaussian distribution.

The outliers interval-valued observations are generated according to nine different scenarios often found in real life problems. Fig. 1 illustrates these scenarios for a sample size $n = 200$ and percentage of outliers equal to 5%.

The three scatter plots in the top of Fig. 1 illustrates the scenarios 1, 2 and 3, representing outliers observations (red rectangles) with the presence of atypical values in the midpoints of Y^c-space, X^c-space and leverage on the midpoints, respectively. On the other side, the three plots in the middle of the figure illustrates the scenarios 4, 5 and 6, representing outliers observations (red rectangles) with unusual values just in the ranges of Y^r-space, X^r-space and leverage on the ranges, respectively. Finally, the three plots in the bottom of the figure illustrates the scenarios 7, 8 and 9 representing outliers observations (red rectangles) with atypical values in the midpoints and ranges, simultaneously. We also considered a scenario without the presence of outliers interval-valued observations, named as scenario 0.

We used two bivariate Gaussian distributions to generate these types of outliers interval observations (in the midpoint, in the ranges and in both). The first one generates the outliers in the midpoints – $\mathcal{N}_2^c(\boldsymbol{\mu}^c = (\mu_x^c, \mu_y^c), \Sigma_c = diag(0.5))$ – while the second one generates the outliers in the ranges – $\mathcal{N}_2^r(\boldsymbol{\mu}^r = (\mu_x^r, \mu_y^r), \Sigma_r = diag(0.1))$. Table 1 presents the parameters used to obtain the outliers interval observations, according with the scenario. In this Table:

- $\bar{x}^c$, $\bar{x}^r$, $\bar{y}^c$ and $\bar{y}^r$ are, respectively, the average of the observed values of the real-valued variables X^c, X^r, Y^c and Y^r;

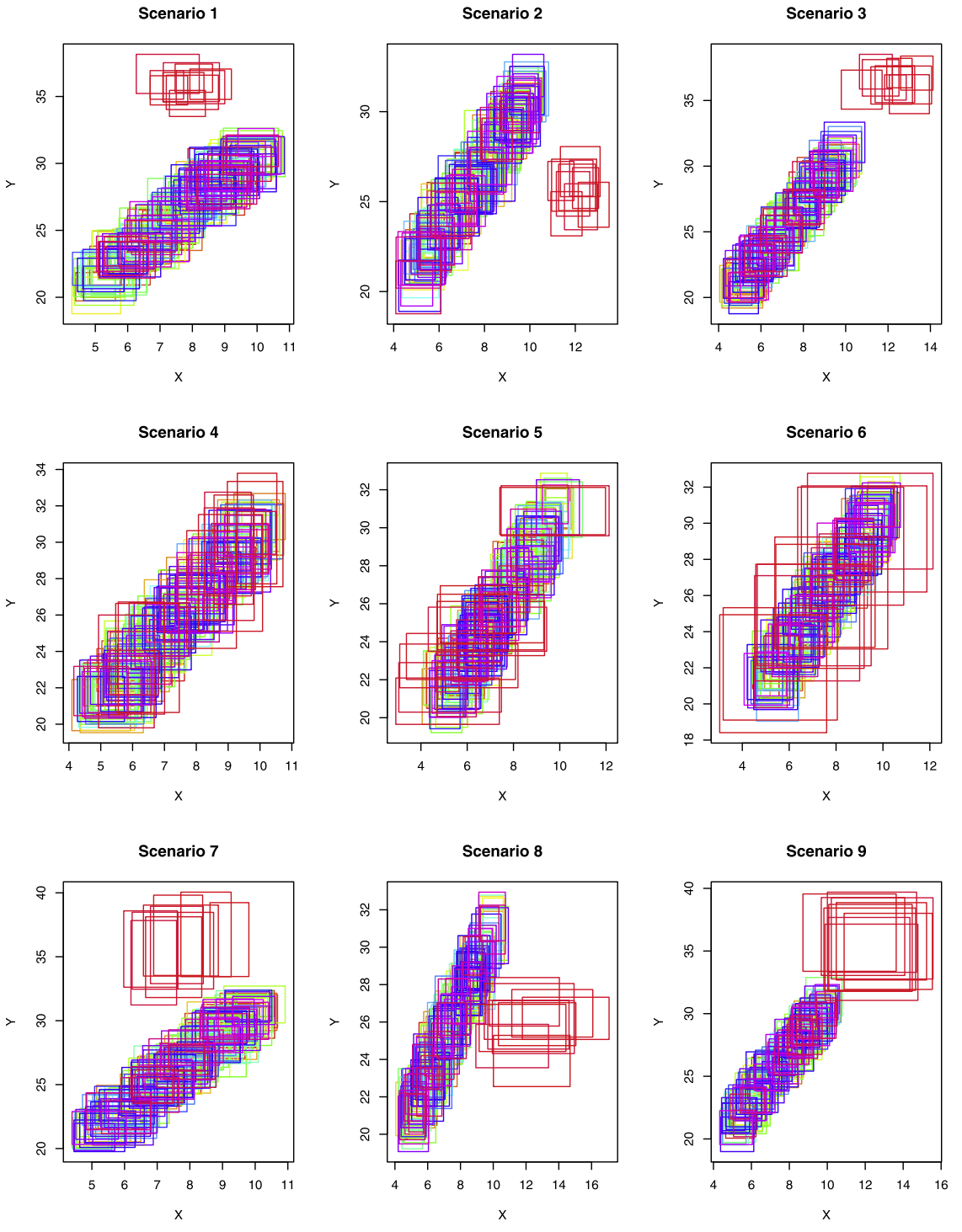

Fig. 1. The interval-valued outliers scenarios. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

- x^c_{max}, x^r_{max}, y^c_{max} and y^r_{max} are, respectively, the maximum of the observed values of the real-valued variables X^c, X^r, Y^c and Y^r;
- S_{x^c}, S_{x^r}, S_{y^c} and S_{y^r} are, respectively, the standard-deviation of the observed values of the real-valued variables X^c, X^r, Y^c and Y^r.

Table 1
Outliers parameters according to bivariate Gaussian distributions.

	Outlier's scenario	μ_x^c	μ_y^c	μ_x^r	μ_y^r
0	No-outliers	–	–	–	–
1	Y^c-space	$\bar{x}^c$	$y_{max}^c + 1.5S_{y^c}$	–	–
2	X^c-space	$x_{max}^c + 1.5S_{x^c}$	$\bar{y}^c$	–	–
3	(Y^c, X^c)-space	$x_{max}^c + 1.5S_{x^c}$	$y_{max}^c + 1.5S_{y^c}$	–	–
4	Y^r-space	–	–	$\bar{x}^r$	$y_{max}^r + 1.5S_{y^r}$
5	X^r-space	–	–	$x_{max}^r + 1.5S_{x^r}$	$\bar{y}^r$
6	(Y^r, X^r)-space	–	–	$x_{max}^r + 1.5S_{x^r}$	$y_{max}^r + 1.5S_{y^r}$
7	$Y^c - Y^r$-space	$\bar{x}^c$	$y_{max}^c + 1.5S_{y^c}$	$\bar{x}^r$	$y_{max}^r + 1.5S_{y^r}$
8	$X^c - X^r$-space	$x_{max}^c + 1.5S_{x^c}$	$\bar{y}^c$	$x_{max}^r + 1.5S_{x^r}$	$\bar{y}^r$
9	(Y^c, X^c)-(Y^r, X^r)-space	$x_{max}^c + 1.5S_{x^c}$	$y_{max}^c + 1.5S_{y^c}$	$x_{max}^r + 1.5S_{x^r}$	$y_{max}^r + 1.5S_{y^r}$

Each outlier scenario taken into account three different sample sizes $n = \{50, 200, 1000\}$ and five different percentage of outliers $p.out = \{5\%, 10\%, 15\%, 20\%, 30\%\}$ in a total of 138 different configurations. A Monte Carlo simulation based on 10,000 replicates was considered for each configuration. Bias and mean squared error (MSE) of the parameter estimates as well as the computational time (t – in seconds) have been computed for each robust regression method for interval-valued variables (iETKRR, IRR, IQR and SSLR).

3.2. Initialization and width hyper-parameter estimators for the parameter estimation algorithm

The performance of the parameter estimation algorithm given in the Section 2.3 depends from initial values for the vector of parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as well as from estimators for the hyper-parameter γ_1 and γ_2. Ref. [7] presented a full experimental study about these points for exponential-type kernel functions in a robust regression approach for real-valued data.

This paper evaluates the performance of the algorithm in terms of bias and MSE of the parameter estimates and recommended the OLS estimator as the best initial value for $\boldsymbol{\beta}$ as well as four estimators for the hyper-parameter γ. The first one (S1) is the Caputo's estimator [6]. The estimator S2 is the median of the values of $r_{ij} = (y_i - \hat{\mu}_j^{\text{OLS}})^2$, $\forall r_{ij} \neq 0$, $i, j = 1, \ldots, n$, while the estimator S3 is defined as

$$\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{\mu}_i^{\text{OLS}})^2,$$

where p is the number of explanatory variables. Finally, S4 is the estimator proposed by Hurvich et al. [28] based on the improved Akaike information criterion (AIC_c).

The final solution provided by this kind of iterative algorithm is very dependent from its starting point. Thus, alternatives for the initialization must be considered. One of the options is provide randomly an initial solution, but in this case the algorithm needs to be executed several times and then to select the best solution (minimal value of the objective function). Another possibility is to provide already a good solution in the initialization stage, that is afterwards improved by the algorithm. The algorithm is executed only once time because the initialization is based on a fixed solution. For the sake of simplicity and also because preliminary experiments showed that the parameter estimates of the CRM method (the OLS version for interval-value data) can provide an initial solution of good quality, in this paper we used the parameter estimates of the CRM method as initial values to the vectors of parameters $\boldsymbol{\beta}_1^{(0)} = \boldsymbol{\beta}_c^{CRM}$ and $\boldsymbol{\beta}_2^{(0)} = \boldsymbol{\beta}_r^{CRM}$, respectively.

Monte Carlo simulation 10,000 times repeated was considered to evaluated how the estimators S1, S2, S3 and S4 affect the performance of the parameter estimation algorithm in terms of bias and MSE. Table 2 shows the hyper-parameter estimators with the lower bias and MSE according with the scenario, percentage of outliers (p.out) and sample size (n). Remark that the pair (S_I, S_{II}) represents the best hyper-parameter estimators for the iETKRR method in the midpoints and in the ranges, respectively.

The estimator S1 was the best one in the scenario 0 due to the small bias, lowest values of MSE and computational time. The estimator S3 is recommended when the percentage of outliers is up to 10% in scenarios 1, 2, 4, 5, 7 and 8. The estimator S3 is also the best one when leverage interval-valued outliers are present in the sample (scenarios 3, 6 and 9), for all sample sizes and percentage of outliers. The estimator S2 is recommended in scenarios 4, 5 and 7, particularly when the percentage of outliers is greater than 10% and when the sample size greater than $n = 50$. The estimator S4 improved the performance of the iETKRR method when the percentage of outliers oscillated between 15% and 20% in the scenarios 1, 2, 7 and 8. A complete description with the values of bias and MSE for all scenarios, sample sizes and percentage of outliers is available in the Supplementary material.

3.3. Evaluating the convergence of parameter estimation algorithm

This section evaluates the parameter estimation algorithm of the iETKRR method about the risk of local optima and/or the maximum number of iterations to be attained. Based on the width hyper-parameter setup presented in Table 2 and

Table 2
Best hyper-parameter estimator for iETKRR method according with the scenario, percentage of outliers and sample size.

Scenario	p.out	n			Scenario	p.out	n		
		50	200	1000			50	200	1000
0	–	(S1,S1)	(S1,S1)	(S1,S1)	1	0.05	(S3,S1)	(S3,S1)	(S3,S1)
						0.10	(S3,S1)	(S3,S1)	(S4,S1)
						0.15	(S4,S1)	(S4,S1)	(S4,S1)
						0.20	(S4,S1)	(S4,S1)	(S4,S1)
						0.30	(S4,S1)	(S4,S1)	(S4,S1)
2	0.05	(S3,S1)	(S3,S1)	(S3,S1)	3	0.05	(S3,S1)	(S3,S1)	(S3,S1)
	0.10	(S3,S1)	(S3,S1)	(S3,S1)		0.10	(S3,S1)	(S3,S1)	(S3,S1)
	0.15	(S4,S1)	(S4,S1)	(S4,S1)		0.15	(S3,S1)	(S3,S1)	(S3,S1)
	0.20	(S4,S1)	(S4,S1)	(S4,S1)		0.20	(S3,S1)	(S3,S1)	(S3,S1)
	0.30	(S3,S1)	(S3,S1)	(S3,S1)		0.30	(S3,S1)	(S3,S1)	(S3,S1)
4	0.05	(S1,S3)	(S1,S3)	(S1,S3)	5	0.05	(S1,S3)	(S1,S3)	(S1,S3)
	0.10	(S1,S3)	(S1,S2)	(S1,S2)		0.10	(S1,S3)	(S1,S3)	(S1,S3)
	0.15	(S1,S4)	(S1,S2)	(S1,S2)		0.15	(S1,S3)	(S1,S2)	(S1,S2)
	0.20	(S1,S4)	(S1,S2)	(S1,S2)		0.20	(S1,S2)	(S1,S2)	(S1,S2)
	0.30	(S1,S4)	(S1,S2)	(S1,S2)		0.30	(S1,S2)	(S1,S2)	(S1,S2)
6	0.05	(S1,S3)	(S1,S3)	(S1,S3)	7	0.05	(S3, S3)	(S3, S3)	(S3, S3)
	0.10	(S1,S3)	(S1,S3)	(S1,S3)		0.10	(S3, S3)	(S3, S2)	(S2, S2)
	0.15	(S1,S3)	(S1,S3)	(S1,S3)		0.15	(S4, S4)	(S4, S2)	(S4, S2)
	0.20	(S1,S3)	(S1,S3)	(S1,S3)		0.20	(S4, S4)	(S4, S2)	(S4, S2)
	0.30	(S1,S3)	(S1,S3)	(S1,S3)		0.30	(S4, S4)	(S4, S2)	(S4, S2)
8	0.05	(S2, S3)	(S2, S3)	(S3, S3)	9	0.05	(S3, S3)	(S3, S3)	(S3, S3)
	0.10	(S2, S3)	(S3, S3)	(S3, S3)		0.10	(S3, S3)	(S3, S3)	(S3, S3)
	0.15	(S4, S3)	(S4, S2)	(S4, S2)		0.15	(S3, S3)	(S3, S3)	(S3, S3)
	0.20	(S4, S3)	(S4, S2)	(S4, S2)		0.20	(S3, S3)	(S3, S3)	(S3, S3)
	0.30	(S2, S3)	(S2, S2)	(S3, S2)		0.30	(S3, S3)	(S3, S3)	(S3, S3)

detailed in the Supplementary material, we verified that the parameter estimation algorithm converges for a region close to the global minimum in the majority of the scenarios and configurations. This is justified due to the bias close to zero and the small MSE of the parameters estimates, mainly, when the sample size increase and when the percentage of outliers is small or in the scenarios without the presence of outliers. In all these situations, the parameter estimates are very close to the true parameters indicating that these parameter estimates are close to the minimal argument of the objective function and, consequently, close to the global minimum.

Another important aspect related to the algorithm is the number of steps required until the convergence of the parameter estimates. We evaluated this point in terms of the average number of steps until convergence ($\overline{iter}$) and its respective standard deviation (S_{iter}). Table 3 brings the results based on Monte Carlo simulation with 10,000 replicates for each configuration. We remember that it was considered a tolerance limit $\epsilon = 1 \times 10^{-10}$ and a maximum number of iterations $T = 100$. Remark that Table 3 exhibits the results for a sample size $n = 50$. The results for the sample sizes $n = \{200, 1000\}$ are available in the Supplementary material.

For scenario 0, where no outliers are considered, the algorithm requires just 6 steps until the convergence of the parameter estimates. Less than 13 steps are required until the convergence of the algorithm in the scenarios 1, 4 and 7, that consider Y-space outliers. In the scenarios with X-Space outliers (2, 5 and 8), the number of steps until convergence increases when the percentage of outliers also increases. Moreover, for interval data set with the presence of leverage points (scenarios 3, 6 and 9) and up to 15% of outliers, the algorithm requires about 30 steps to reach the convergence.

For the sample sizes $n = \{200, 1000\}$, the values of $\overline{iter}$ are close to those presented in Table 3, however the values of S_{iter} decreases when n increases (see Supplementary material for details). Finally, it is important to highlight that the fixed maximum number of iterations was never hit.

3.4. Comparative study between the robust methods for interval-valued data

This section considers a comparative study between the iETKRR method and the robust regression methods IRR, IQR and SSLR. For the SSLR method, we follow the same set up presented by Domingues et al. [14], i.e., we used a Student-t distribution for the midpoint and a Gaussian distribution for the ranges of the intervals. We considered a grid between 2 and 10 for the degree of freedom of Student-t distribution, choosing the SSLR model with lower bias and MSE. The SSLR with 2 degree of freedom demonstrated the best results and will be presented in the next Tables. For the IQR approach, we considered the median ($\tau = 0.5$) for the midpoint and for the range of the intervals.

All methods were evaluated in the 10 outliers scenarios presented in Table 1, taking into account three different sample sizes and five different percentage of outliers, in a total of 138 different configurations. A Monte Carlo simulation based on 10,000 replicates was considered for each configuration and the comparison between the robust regression methods and the

Table 3
Number of iterations until convergence of the parameter estimation algorithm, $n = 50$.

Scenario	p.out	$\overline{iter}$	S_{iter}	Scenario	p.out	$\overline{iter}$	S_{iter}
0	–	5.80	(0.900)	1	0.05	8.26	0.90
					0.10	6.84	0.52
					0.15	6.81	0.52
					0.20	7.16	0.56
					0.30	8.93	0.86
2	0.05	9.91	1.30	3	0.05	31.09	15.22
	0.10	9.21	0.84		0.10	28.14	14.26
	0.15	10.03	1.57		0.15	28.01	14.69
	0.20	26.82	15.78		0.20	27.99	14.80
	0.30	34.59	18.07		0.30	29.23	14.93
4	0.05	12.86	4.39	5	0.05	16.88	6.17
	0.10	9.38	1.47		0.10	15.40	3.65
	0.15	10.82	2.29		0.15	17.16	5.13
	0.20	10.97	2.22		0.20	22.41	9.68
	0.30	11.95	2.14		0.30	40.14	19.09
6	0.05	27.58	14.09	7	0.05	12.88	4.45
	0.10	23.52	12.81		0.10	9.38	1.42
	0.15	23.44	12.91		0.15	10.84	2.34
	0.20	24.28	13.48		0.20	10.95	2.21
	0.30	26.46	14.84		0.30	11.95	2.15
8	0.05	16.85	5.99	9	0.05	31.06	15.27
	0.10	15.42	3.69		0.10	28.13	14.59
	0.15	19.50	9.65		0.15	27.88	14.39
	0.20	28.40	17.83		0.20	28.08	15.20
	0.30	40.50	19.24		0.30	29.42	15.31

Table 4
Scenario 0: bias and MSE for the parameter estimates, according to the interval robust regression method.

Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			t
	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
iETKRR	11.0027	0.0027	0.1446	1.9997	−0.0003	0.0025	0.9991	−0.0009	0.0060	1.0004	0.0004	0.0026	0.0020
IRR	11.0027	0.0027	0.1541	1.9997	−0.0003	0.0026	0.9990	−0.0010	0.0063	1.0004	0.0004	0.0027	0.0133
IQR	11.0017	0.0017	0.2291	1.9998	−0.0002	0.0039	0.9986	−0.0014	0.0094	1.0007	0.0007	0.0040	0.0024
SSLR	11.0029	0.0029	0.1622	1.9997	−0.0003	0.0028	0.9990	−0.0010	0.0059	1.0004	0.0004	0.0025	0.0095

proposed approach occurred in terms of bias, MSE of the parameter estimates and the computational time (t, in seconds). Remark that the tables below consider the sample size $n = 50$. The results for the sample sizes $n = \{200, 1000\}$ are available in the Supplementary material.

3.4.1. Outliers scenario 0

Table 4 presents the performance of each method in scenario 0, where only clean observations are considered. As mentioned in the previous section, the hyper-parameter estimator S1 is recommended in this scenario.

All robust methods presented a good performance in terms of bias. However, the iETKRR approach presented the best precision for the parameter estimates due to the smallest MSE as well as the lowest computational time. The IQR method presented the worst performance in terms of MSE.

3.4.2. Outliers scenario 1

Scenario 1 considers interval-valued data with the presence of atypical values (outliers) in the midpoint of the response variable (Y^c-space). No outliers are considered for the ranges of the intervals. In this way, the comparison is done in term of the bias and MSE for the parameters β_0^c and β_1^c. For the iETKRR method, we considered the hyper-parameter estimator S3 up to $p.out = 10\%$ and the estimator S4 for the other configurations, as suggested in Table 2.

Table 5 presents the results of the comparative study between the robust regression methods for interval-valued variables. The iETKRR method outperforms the methods IQR and SSLR in all configurations. Up to 10% of outliers, the iETKRR method presents a similar performance to the IRR method, but with about of a half of computational time required by the previous method. When the percentage of outliers is greater than 10% the IRR method presented the best results.

3.4.3. Outliers scenario 2

Scenario 2 considers interval-valued data with the presence of outliers in the midpoint of the explanatory variable (X^c-space). No outliers are considered for the ranges of the intervals. In this way, the comparison is done in term of the bias

Table 5
Scenario 1: bias and MSE for the parameter estimates, according to the interval robust regression method.

p.out	Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			t
		Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
5%	*iETKRR*	11.0033	0.0023	0.1537	1.9995	−0.0005	0.0026	1.0003	0.0003	0.0058	0.9997	−0.0003	0.0025	0.0069
	IRR	11.0024	0.0024	0.1578	1.9996	−0.0004	0.0027	1.0003	0.0003	0.0061	0.9997	−0.0003	0.0026	0.0133
	IQR	11.0282	0.0282	0.2385	1.9996	−0.0004	0.0041	1.0005	0.0005	0.0091	0.9997	−0.0003	0.0039	0.0024
	SSLR	11.0038	0.0038	0.1891	1.9998	−0.0002	0.0032	1.0003	0.0003	0.0057	0.9997	−0.0003	0.0025	0.0139
10%	*iETKRR*	10.9973	−0.0027	0.1599	2.0005	0.0005	0.0027	1.0009	0.0009	0.0059	0.9995	−0.0005	0.0025	0.0067
	IRR	11.0023	0.0023	0.1635	1.9997	−0.0003	0.0028	1.0008	0.0008	0.0062	0.9995	−0.0005	0.0027	0.0134
	IQR	11.0889	0.0889	0.2710	1.9990	−0.0010	0.0045	1.0008	0.0008	0.0093	0.9995	−0.0005	0.0040	0.0024
	SSLR	11.0119	0.0119	0.1863	2.0002	0.0002	0.0032	1.0008	0.0008	0.0058	0.9995	−0.0005	0.0025	0.0160
15%	*iETKRR*	10.9943	−0.0057	0.2007	2.0009	0.0009	0.0034	1.0014	0.0014	0.0060	0.9992	−0.0008	0.0026	0.0436
	IRR	10.9987	−0.0013	0.1676	2.0002	0.0002	0.0029	1.0014	0.0014	0.0063	0.9992	−0.0008	0.0027	0.0138
	IQR	11.1177	0.1177	0.2892	1.9995	−0.0005	0.0048	1.0013	0.0013	0.0092	0.9992	−0.0008	0.0040	0.0026
	SSLR	11.0143	0.0143	0.1868	2.0010	0.0010	0.0032	1.0013	0.0013	0.0059	0.9993	−0.0007	0.0025	0.0184
20%	*iETKRR*	11.0062	0.0062	0.2097	1.9992	−0.0008	0.0036	0.9991	−0.0009	0.0058	1.0008	0.0008	0.0025	0.0446
	IRR	10.9940	−0.0060	0.1842	2.0009	0.0009	0.0032	0.9992	−0.0008	0.0061	1.0007	0.0007	0.0026	0.0148
	IQR	11.1500	0.1500	0.3304	2.0012	0.0012	0.0053	0.9999	−0.0001	0.0089	1.0004	0.0004	0.0039	0.0028
	SSLR	11.0154	0.0154	0.1990	2.0032	0.0032	0.0034	0.9991	−0.0009	0.0058	1.0008	0.0008	0.0025	0.0227
30%	*iETKRR*	11.0058	0.0058	0.6090	1.9996	−0.0004	0.0078	0.9999	−0.0001	0.0062	1.0001	0.0001	0.0026	0.0442
	IRR	11.0041	0.0041	0.2088	1.9996	−0.0004	0.0036	0.9998	−0.0002	0.0063	1.0002	0.0002	0.0027	0.0132
	IQR	11.2965	0.2965	0.5158	1.9996	−0.0004	0.0074	1.0004	0.0004	0.0093	0.9999	−0.0001	0.0040	0.0024
	SSLR	11.1431	0.1431	0.2647	2.0112	0.0112	0.0043	0.9998	−0.0002	0.0058	1.0002	0.0002	0.0025	0.0380

Table 6
Scenario 2: bias and MSE for the parameter estimates, according to the interval robust regression method.

p.out	Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			t
		Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
5%	*iETKRR*	11.0039	0.0039	0.1575	1.9994	−0.0006	0.0027	0.9993	−0.0007	0.0059	1.0003	0.0003	0.0026	0.0074
	IRR	11.0034	0.0034	0.1590	1.9995	−0.0005	0.0027	0.9993	−0.0007	0.0063	1.0003	0.0003	0.0027	0.0132
	IQR	11.4182	0.4182	0.4268	1.9406	−0.0594	0.0079	0.9986	−0.0014	0.0092	1.0008	0.0008	0.0040	0.0024
	SSLR	11.0620	0.0620	0.1970	1.9911	−0.0089	0.0034	0.9994	−0.0006	0.0058	1.0003	0.0003	0.0025	0.0138
10%	*iETKRR*	11.0007	0.0007	0.1593	1.9999	−0.0001	0.0027	1.0007	0.0007	0.0059	0.9996	−0.0004	0.0025	0.0045
	IRR	10.9950	−0.0050	0.1647	2.0007	0.0007	0.0028	1.0006	0.0006	0.0062	0.9997	−0.0003	0.0027	0.0132
	OQR	12.4458	1.4458	2.4692	1.7941	−0.2059	0.0493	1.0006	0.0006	0.0093	0.9997	−0.0003	0.0040	0.0024
	SSLR	11.2261	0.2261	0.2476	1.9679	−0.0321	0.0044	1.0006	0.0006	0.0058	0.9996	−0.0004	0.0025	0.0170
15%	*iETKRR*	11.0287	0.0287	0.4218	1.9958	−0.0042	0.0084	0.9999	−0.0001	0.0059	1.0000	0.0000	0.0025	0.0427
	IRR	11.0010	0.0010	0.1725	1.9999	−0.0001	0.0029	0.9998	−0.0002	0.0062	1.0001	0.0001	0.0026	0.0132
	IQR	13.4762	2.4762	7.1886	1.6447	−0.3553	0.1486	1.0003	0.0003	0.0092	0.9997	−0.0003	0.0039	0.0024
	SSLR	11.5291	0.5291	1.8323	1.9234	−0.0766	0.0408	0.9999	−0.0001	0.0058	1.0000	0.0000	0.0025	0.0213
20%	*iETKRR*	14.6497	3.6497	44.2366	1.4632	−0.5368	0.9300	1.0000	0.0000	0.0061	1.0000	0.0000	0.0026	0.0445
	IRR	11.0025	0.0025	0.1844	1.9997	−0.0003	0.0032	0.9999	−0.0001	0.0061	1.0001	0.0001	0.0026	0.0132
	IQR	18.9867	7.9867	74.1869	0.8385	−1.1615	1.5730	0.9999	−0.0001	0.0091	1.0001	0.0001	0.0039	0.0024
	SSLR	19.0406	8.0406	78.1596	0.8316	−1.1684	1.6516	0.9998	−0.0002	0.0057	1.0002	0.0002	0.0024	0.0205
30%	*iETKRR*	22.3511	11.3511	133.8339	0.3253	−1.6747	2.8348	0.9994	−0.0006	0.0058	1.0003	0.0003	0.0025	0.0115
	IRR	11.0028	0.0028	0.2362	1.9998	−0.0002	0.0042	0.9994	−0.0006	0.0061	1.0003	0.0003	0.0027	0.0134
	IQR	22.5374	11.5374	134.2646	0.3342	−1.6658	2.7833	1.0004	0.0004	0.0091	0.9997	−0.0003	0.0039	0.0024
	SSLR	22.2378	11.2378	126.9670	0.3739	−1.6261	2.6495	0.9995	−0.0005	0.0058	1.0003	0.0003	0.0025	0.0137

and MSE for the parameters β_0^c and β_1^c. For the iETKRR method, we considered the hyper-parameter estimator S4 for $p.out = \{15\%, 20\%\}$ and the estimator S3 for the other configurations, as suggested in Table 2.

Table 6 illustrates the results of the comparative study. The iETKRR method presented the best performance for $n = 50$ and *p.out* up to 10%. Up to 15% of outliers, the method iETKRR outperforms the methods SSLR and IQR. However, after 15% of outliers, just the IRR method presented a good performance.

3.4.4. Outliers scenario 3

Scenario 3 considers interval-valued data with the presence of midpoint leverage observations, i.e., unusual observations in the midpoints of the response Z and explanatory W interval-valued variables are present in the data set. No outliers are considered in the ranges of these variables. Again, the comparison is done in term of the bias and MSE for the parameters β_0^c and β_1^c. For the iETKRR method, we considered the hyper-parameter estimator S3 in all configurations, as suggested in Table 2.

Table 7
Scenario 3: bias and MSE for the parameter estimates, according to the interval robust regression method.

p.out	Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			t
		Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
5%	*iETKRR*	10.9994	−0.0006	0.3525	2.0003	0.0003	0.0062	1.0012	0.0012	0.0059	0.9993	−0.0007	0.0025	0.0109
	IRR	10.9859	−0.0141	0.2292	2.0023	0.0023	0.0041	1.0013	0.0013	0.0062	0.9993	−0.0007	0.0027	0.0132
	IQR	10.9704	−0.0296	0.2947	2.0046	0.0046	0.0052	1.0013	0.0013	0.0093	0.9993	−0.0007	0.0040	0.0024
	SSLR	10.9845	−0.0155	0.2423	2.0025	0.0025	0.0043	1.0012	0.0012	0.0058	0.9994	−0.0006	0.0025	0.0129
10%	*iETKRR*	10.9745	−0.0255	0.3644	2.0043	0.0043	0.0066	1.0002	0.0002	0.0059	0.9998	−0.0002	0.0025	0.0103
	IRR	10.9483	−0.0517	0.3382	2.0083	0.0083	0.0063	1.0003	0.0003	0.0062	0.9997	−0.0003	0.0027	0.0134
	IQR	10.9311	−0.0689	0.3739	2.0113	0.0113	0.0068	1.0000	0.0000	0.0093	0.9999	−0.0001	0.0040	0.0024
	SSLR	10.9474	−0.0526	0.3147	2.0086	0.0086	0.0058	1.0001	0.0001	0.0058	0.9999	−0.0001	0.0025	0.0136
15%	*iETKRR*	10.9666	−0.0334	0.4004	2.0053	0.0053	0.0073	1.0009	0.0009	0.0058	0.9995	−0.0005	0.0025	0.0104
	IRR	10.9327	−0.0673	0.3991	2.0109	0.0109	0.0075	1.0007	0.0007	0.0061	0.9997	−0.0003	0.0026	0.0134
	IQR	10.9188	−0.0812	0.4286	2.0134	0.0134	0.0079	1.0005	0.0005	0.0090	0.9998	−0.0002	0.0039	0.0025
	SSLR	10.9348	−0.0652	0.3681	2.0106	0.0106	0.0068	1.0010	0.0001	0.0057	0.9995	−0.0005	0.0024	0.0136
20%	*iETKRR*	10.9603	−0.0397	0.4594	2.0064	0.0064	0.0084	0.9999	−0.0001	0.0059	1.0002	0.0002	0.0025	0.0103
	IRR	10.9294	−0.0706	0.4601	2.0119	0.0119	0.0087	0.9998	−0.0002	0.0062	1.0002	0.0002	0.0027	0.0133
	IQR	10.9202	−0.0798	0.4653	2.0137	0.0137	0.0085	1.0001	0.0001	0.0091	1.0000	0.0000	0.0039	0.0024
	SSLR	10.9332	−0.0668	0.4152	2.0113	0.0113	0.0077	0.9998	−0.0002	0.0059	1.0003	0.0003	0.0025	0.0139
30%	*iETKRR*	10.9389	−0.0611	0.5575	2.0104	0.0104	0.0105	0.9997	−0.0003	0.0060	1.0002	0.0002	0.0026	0.0107
	IRR	10.9365	−0.0635	0.5453	2.0130	0.0130	0.0104	0.9998	−0.0002	0.0064	1.0002	0.0002	0.0028	0.0133
	IQR	10.9465	−0.0535	0.5354	2.0121	0.0121	0.0099	0.9992	−0.0008	0.0094	1.0005	0.0005	0.0041	0.0024
	SSLR	10.9391	−0.0609	0.4990	2.0122	0.0122	0.0094	0.9996	−0.0004	0.0060	1.0003	0.0003	0.0026	0.0141

Table 8
Scenario 4: bias and MSE for the parameter estimates, according to the interval robust regression method.

p.out	Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			t
		Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
5%	*iETKRR*	11.0030	0.0030	0.1434	1.9998	−0.0002	0.0025	0.9992	−0.0008	0.0073	1.0008	0.0008	0.0031	0.0077
	IRR	11.0032	0.0032	0.1530	1.9998	−0.0002	0.0026	0.9982	−0.0018	0.0066	1.0015	0.0015	0.0028	0.0132
	IQR	11.0057	0.0057	0.2289	1.9996	−0.0004	0.0039	1.0071	0.0071	0.0101	0.9987	−0.0013	0.0043	0.0024
	SSLR	11.0041	0.0041	0.1886	1.9997	−0.0003	0.0032	1.1112	0.1112	0.0515	0.9526	−0.0474	0.0185	0.0123
10%	*iETKRR*	10.9987	−0.0013	0.1468	2.0001	0.0001	0.0025	0.9914	−0.0086	0.0072	1.0073	0.0073	0.0031	0.0030
	IRR	10.9986	−0.0014	0.1561	2.0001	0.0001	0.0027	0.9936	−0.0064	0.0069	1.0054	0.0054	0.0030	0.0132
	IQR	11.0006	0.0006	0.2313	1.9998	−0.0002	0.0040	1.0226	0.0226	0.0123	0.9955	−0.0045	0.0051	0.0024
	SSLR	10.9982	−0.0018	0.1901	2.0001	0.0001	0.0033	1.3140	0.3140	0.1813	0.8682	−0.1318	0.0518	0.0123
15%	*iETKRR*	11.0004	0.0004	0.1426	1.9998	−0.0002	0.0024	0.9861	−0.0139	0.0080	1.0121	0.0121	0.0035	0.0468
	IRR	11.0009	0.0009	0.1516	1.9997	−0.0003	0.0026	0.9855	−0.0145	0.0076	1.0115	0.0115	0.0034	0.0133
	IQR	11.0003	0.0003	0.2245	1.9999	−0.0001	0.0038	1.0287	0.0287	0.0136	0.9956	−0.0044	0.0056	0.0023
	SSLR	11.0013	0.0013	0.1854	1.9997	−0.0003	0.0032	1.4170	0.4170	0.2762	0.8252	−0.1748	0.0732	0.0123
20%	*iETKRR*	10.9996	−0.0004	0.1473	2.0000	0.0000	0.0025	0.9770	−0.0230	0.0095	1.0199	0.0199	0.0044	0.0455
	IRR	10.9995	−0.0005	0.1562	2.0000	0.0000	0.0027	0.9738	−0.0262	0.0093	1.0210	0.0210	0.0042	0.0131
	IQR	10.9987	−0.0013	0.2331	2.0001	0.0001	0.0040	1.0389	0.0389	0.0177	0.9948	−0.0052	0.0070	0.0024
	SSLR	10.9987	−0.0013	0.1907	2.0001	0.0001	0.0033	1.5376	0.5376	0.4117	0.7749	−0.2251	0.1019	0.0122
30%	*iETKRR*	11.0011	0.0011	0.1458	2.0000	0.0000	0.0025	0.9481	−0.0519	0.0152	1.0454	0.0454	0.0078	0.0455
	IRR	11.0010	0.0010	0.1545	2.0000	0.0000	0.0027	0.9264	−0.0736	0.0195	1.0634	0.0634	0.0106	0.0132
	IQR	11.0033	0.0033	0.2302	1.9998	−0.0002	0.0040	1.0821	0.0821	0.0317	0.9833	−0.0167	0.0111	0.0024
	SSLR	11.0019	0.0019	0.1879	2.0000	0.0000	0.0032	1.8282	0.8282	0.8158	0.6518	−0.3482	0.1756	0.0121

Table 7 presents the comparative study between the robust regression methods for interval-valued variables in scenario 3. The iETKRR method outperforms the previous methods in all configurations. Moreover, the iETKRR method also demonstrated a good performance when the percentage of outliers is 20% or 30%.

3.4.5. Outliers scenario 4

The scenario 4 considers interval-valued data with the presence of outliers in the ranges of the response interval-valued variable Z. No outliers are considered in the midpoints of the intervals variables and the comparison is done in term of the bias and MSE for the parameters β_0^r and β_1^r. For the iETKRR method, we considered the hyper-parameter estimator S3 for $p.out = 5\%$ and the estimator S4 for $n = 50$ and $p.out$ greater than 10%. Finally, the estimator S2 is considered in the remaining configurations, as suggested in Table 2.

Table 8 brings the results of the comparative study between the robust regression methods for interval-valued variables. The iETKRR outperforms the other approaches when $p.out = 5\%$ and presents a competitive performance with the IRR approach for $n = 50$. The IQR method is the best one for $p.out = 30\%$. However, the IQR method presented a high MSE

Table 9
Scenario 5: bias and MSE for the parameter estimates, according to the interval robust regression method.

p.out	Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			t
		Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
5%	*iETKRR*	11.0000	0.0000	0.1467	1.9999	−0.0001	0.0025	1.0011	0.0011	0.0081	0.9990	−0.0010	0.0035	0.0085
	IRR	11.0005	0.0005	0.1566	1.9998	−0.0002	0.0027	1.0051	0.0051	0.0069	0.9961	−0.0039	0.0030	0.0132
	IQR	10.9991	−0.0009	0.2344	2.0001	0.0001	0.0040	1.0803	0.0803	0.0169	0.9429	−0.0571	0.0078	0.0024
	SSLR	10.9988	−0.0012	0.1925	2.0001	0.0001	0.0033	1.4258	0.4258	0.2149	0.6990	−0.3010	0.1065	0.0124
10%	*iETKRR*	11.0031	0.0031	0.1490	1.9997	−0.0003	0.0026	1.0169	0.0169	0.0080	0.9876	−0.0124	0.0035	0.0082
	IRR	11.0041	0.0041	0.1583	1.9996	−0.0004	0.0027	1.0320	0.0320	0.0104	0.9768	−0.0232	0.0048	0.0133
	IQR	10.9999	−0.0001	0.2324	2.0001	0.0001	0.0040	1.2806	0.2806	0.0963	0.8003	−0.1997	0.0480	0.0025
	SSLR	11.0024	0.0024	0.1914	1.9998	−0.0002	0.0033	1.7997	0.7997	0.6635	0.4361	−0.5639	0.3293	0.0122
15%	*iETKRR*	10.9993	−0.0007	0.1443	2.0001	0.0001	0.0025	1.0314	0.0314	0.0149	0.9769	−0.0231	0.0074	0.0084
	IRR	11.0001	0.0001	0.1530	2.0000	0.0000	0.0026	1.0553	0.0553	0.0167	0.9600	−0.0400	0.0079	0.0134
	IQR	11.0007	0.0007	0.2231	1.9999	−0.0001	0.0038	1.4385	0.4385	0.2253	0.6872	−0.3128	0.1141	0.0024
	SSLR	11.0002	0.0002	0.1854	1.9999	−0.0001	0.0032	1.8991	0.8991	0.8286	0.3664	−0.6336	0.4109	0.0122
20%	*iETKRR*	10.9984	−0.0016	0.1498	2.0002	0.0002	0.0026	1.0964	0.0964	0.0767	0.9289	−0.0711	0.0407	0.0051
	IRR	10.9970	−0.0030	0.1592	2.0004	0.0004	0.0027	1.1246	0.1246	0.0504	0.9105	−0.0895	0.0248	0.0130
	IQR	10.9932	−0.0068	0.2368	2.0010	0.0010	0.0040	1.6719	0.6719	0.5022	0.5190	−0.4810	0.2566	0.0024
	SSLR	10.9957	−0.0043	0.1944	2.0006	0.0006	0.0033	1.9889	0.9889	0.9958	0.3039	−0.6961	0.4928	0.0122
30%	*iETKRR*	10.9982	−0.0018	0.1465	2.0002	0.0002	0.0025	1.4586	0.4586	0.4459	0.6646	−0.3354	0.2355	0.0080
	IRR	10.9992	−0.0008	0.1559	2.0000	0.0000	0.0027	1.6777	0.6777	0.6043	0.5178	−0.4822	0.3045	0.0134
	IQR	10.9988	−0.0012	0.2284	2.0001	0.0001	0.0039	1.9696	0.9696	0.9719	0.3062	−0.6938	0.4954	0.0023
	SSLR	10.9990	−0.0010	0.1894	2.0000	0.0000	0.0032	2.1104	1.1104	1.2459	0.2194	−0.7806	0.6150	0.0122

for the intercept parameter β_0^r in comparison with the previous approaches. Finally, the SSLR method presented the worst performance.

3.4.6. Outliers scenario 5

Scenario 5 considers interval-valued data with the presence of outliers in the ranges of the explanatory interval-valued variable (X^r-space outlier). No outliers are considered in the midpoints of the intervals. In this way, the comparison is done in term of the bias and MSE for the parameters β_0^r and β_1^r. For the iETKRR method, we considered the hyper-parameter estimator S3 for $p.out = \{5\%, 10\%\}$ and the estimator S2 in the other configurations, as suggested in Table 2.

From Table 9, it is possible to verify that the iETKRR method outperforms the other approach in all configurations. The IRR method presented the second best performance followed by the IQR approach. These methods presented a small bias and MSE up to $p.out = 15\%$.

3.4.7. Outliers scenario 6

Scenario 6 considers interval-valued data with the presence of range leverage points, i.e., we have the presence of unusual observations in the ranges of the response Z and explanatory W interval-valued variables. No outliers are considered in the midpoints of these variables. Again, the comparison is done in term of the bias and MSE for the parameters β_0^r and β_1^r. For the iETKRR method, we considered the hyper-parameter estimator S3 in all configurations, as suggested in Table 2.

The iETKRR method outperforms the other robust regression methods in the majority of the configuration settings. Until $p.out = 10\%$, the IRR method presents the second best performance and outperforms the IQR and SSLR approaches. Table 10 provides the experimental results for this scenario.

3.4.8. Outliers scenario 7

The scenario 7 considers outliers in the response interval-valued variable Z, with the presence of atypical values in the midpoints Y^c and ranges Y^r, simultaneously. No outliers are considered in the explanatory interval-valued variable W. The comparison between the methods will be done in term of the bias and MSE for the vector of parameters $\boldsymbol{\beta}^c$ and $\boldsymbol{\beta}^r$. We considered the hyper-parameter estimators for iETKRR method according to Table 2. The estimator S2 was recommended when the presence of atypical observations occurs in the ranges of the response variable (Y^r). In relation to the midpoints, we used the estimator S3 when $p.out = \{5\%, 10\%\}$ and the estimator S4 in the remaining configurations.

Table 11 presents the results of the comparative study between the robust regression methods for interval-valued variables. The iETKRR method outperformed the IQR and SSLR approaches and presented a competitive performance in comparison with IRR method. Moreover, the iETKRR method required half of computational time in comparison with the IRR method, for *p.out* up to 10%.

3.4.9. Outliers scenario 8

The scenario 8 considers outliers in the explanatory interval-valued variable W, with the presence of atypical values in the midpoint X^c and range X^r real-valued variables, simultaneously. No outliers are considered in the response interval-

Table 10
Scenario 6: bias and MSE for the parameter estimates, according to the interval robust regression method.

p.out	Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			*t*
		Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
5%	*iETKRR*	10.9996	−0.0004	0.1458	2.0001	0.0001	0.0025	0.9989	−0.0011	0.0128	1.0006	0.0006	0.0056	0.0102
	IRR	10.9991	−0.0009	0.1550	2.0002	0.0002	0.0027	0.9978	−0.0022	0.0090	1.0016	0.0016	0.0040	0.0134
	IQR	10.9962	−0.0038	0.2332	2.0006	0.0006	0.0040	0.9974	−0.0026	0.0122	1.0023	0.0023	0.0054	0.0023
	SSLR	10.9976	−0.0024	0.1905	2.0004	0.0004	0.0033	0.9881	−0.0119	0.0248	1.0109	0.0109	0.0120	0.0124
10%	*iETKRR*	10.9968	−0.0032	0.1446	2.0003	0.0003	0.0025	1.0002	0.0002	0.0132	1.0003	0.0003	0.0059	0.0095
	IRR	10.9966	−0.0034	0.1545	2.0003	0.0003	0.0026	0.9987	−0.0013	0.0140	1.0018	0.0018	0.0065	0.0134
	IQR	10.9932	−0.0068	0.2263	2.0008	0.0008	0.0039	0.9976	−0.0024	0.0158	1.0038	0.0038	0.0071	0.0024
	SSLR	10.9967	−0.0033	0.1884	2.0003	0.0003	0.0032	1.0009	0.0009	0.0304	1.0084	0.0084	0.0146	0.0123
15%	*iETKRR*	11.0034	0.0034	0.1484	1.9995	−0.0005	0.0025	0.9991	−0.0009	0.0150	1.0011	0.0011	0.0068	0.0096
	IRR	11.0035	0.0035	0.1571	1.9996	−0.0004	0.0027	1.0000	0.0000	0.0172	1.0012	0.0012	0.0081	0.0134
	IQR	11.0038	0.0038	0.2330	1.9994	−0.0006	0.0040	1.0021	0.0021	0.0174	1.0014	0.0014	0.0079	0.0024
	SSLR	11.0040	0.0040	0.1921	1.9995	−0.0005	0.0033	1.0188	0.0188	0.0305	0.9996	−0.0004	0.0145	0.0122
20%	*iETKRR*	11.0112	0.0112	0.1483	1.9986	−0.0014	0.0026	1.0014	0.0014	0.0184	1.0002	0.0002	0.0086	0.0095
	IRR	11.0101	0.0101	0.1566	1.9988	−0.0012	0.0027	1.0007	0.0007	0.0214	1.0018	0.0018	0.0102	0.0132
	IQR	11.0060	0.0060	0.2293	1.9993	−0.0007	0.0040	1.0069	0.0069	0.0196	0.9995	−0.0005	0.0090	0.0024
	SSLR	11.0087	0.0087	0.1891	1.9989	−0.0011	0.0033	1.0381	0.0381	0.0313	0.9908	−0.0092	0.0143	0.0124
30%	*iETKRR*	11.0044	0.0044	0.1453	1.9995	−0.0005	0.0025	1.0022	0.0022	0.0245	1.0011	0.0011	0.0116	0.0099
	IRR	11.0039	0.0039	0.1550	1.9996	−0.0004	0.0027	1.0064	0.0064	0.0273	1.0008	0.0008	0.0130	0.0135
	IQR	11.0030	0.0030	0.2308	1.9997	−0.0003	0.0040	1.0165	0.0165	0.0226	0.9959	−0.0041	0.0103	0.0023
	SSLR	11.0029	0.0029	0.1891	1.9997	−0.0003	0.0033	1.0877	0.0877	0.0361	0.9695	−0.0305	0.0141	0.0124

Table 11
Scenario 7: bias and MSE for the parameter estimates, according to the interval robust regression method.

p.out	Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			t
		Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
5%	*iETKRR*	10.9941	−0.0059	0.1548	2.0010	0.0010	0.0026	0.9974	−0.0026	0.0072	1.0019	0.0019	0.0031	0.0036
	IRR	10.9940	−0.0060	0.1596	2.0010	0.0010	0.0027	0.9972	−0.0028	0.0066	1.0021	0.0021	0.0028	0.0132
	IQR	11.0192	0.0192	0.2389	2.0012	0.0012	0.0041	1.0054	0.0054	0.0104	0.9998	−0.0002	0.0044	0.0024
	SSLR	10.9973	−0.0027	0.1900	2.0011	0.0011	0.0033	1.1107	0.1107	0.0517	0.9531	−0.0469	0.0186	0.0138
10%	*iETKRR*	10.9957	−0.0043	0.1602	2.0008	0.0008	0.0027	0.9889	−0.0111	0.0069	1.0093	0.0093	0.0030	0.0027
	IRR	11.0008	0.0008	0.1648	1.9999	−0.0001	0.0028	0.9920	−0.0080	0.0067	1.0065	0.0065	0.0029	0.0133
	IQR	11.0823	0.0823	0.2698	1.9999	−0.0001	0.0045	1.0215	0.0215	0.0119	0.9962	−0.0038	0.0049	0.0024
	SSLR	11.0095	0.0095	0.1905	2.0006	0.0006	0.0033	1.3180	0.3180	0.1860	0.8656	−0.1344	0.0533	0.0162
15%	*iETKRR*	10.9944	−0.0056	0.1929	2.0009	0.0009	0.0033	0.9855	−0.0145	0.0080	1.0124	0.0124	0.0036	0.0417
	IRR	11.0022	0.0022	0.1694	1.9995	−0.0005	0.0029	0.9843	−0.0157	0.0075	1.0124	0.0124	0.0033	0.0132
	IQR	11.1190	0.1190	0.2959	1.9991	−0.0009	0.0048	1.0268	0.0268	0.0134	0.9970	−0.0030	0.0055	0.0024
	SSLR	11.0173	0.0173	0.1896	2.0005	0.0005	0.0032	1.4135	0.4135	0.2699	0.8273	−0.1727	0.0710	0.0179
20%	*iETKRR*	11.0035	0.0035	0.2173	1.9995	−0.0005	0.0037	0.9766	−0.0234	0.0097	1.0201	0.0201	0.0045	0.0407
	IRR	11.0044	0.0044	0.1860	1.9994	−0.0006	0.0032	0.9763	−0.0237	0.0092	1.0193	0.0193	0.0042	0.0131
	IQR	11.1586	0.1586	0.3494	1.9998	−0.0002	0.0056	1.0415	0.0415	0.0173	0.9927	−0.0073	0.0068	0.0023
	SSLR	11.0263	0.0263	0.2031	2.0016	0.0016	0.0035	1.5449	0.5449	0.4154	0.7694	−0.2306	0.1025	0.0206
30%	*iETKRR*	11.0060	0.0060	0.3115	1.9992	−0.0008	0.0050	0.9488	−0.0512	0.0156	1.0451	0.0451	0.0081	0.0398
	IRR	11.0005	0.0005	0.2019	2.0002	0.0002	0.0035	0.9259	−0.0741	0.0195	1.0638	0.0638	0.0107	0.0132
	IQR	11.2875	0.2875	0.4990	2.0008	0.0008	0.0072	1.0801	0.0801	0.0319	0.9846	−0.0154	0.0112	0.0024
	SSLR	11.1377	0.1377	0.2531	2.0120	0.0120	0.0041	1.8138	0.8138	0.7972	0.6620	−0.3380	0.1707	0.0380

valued variable Z. The comparison between the methods is done in term of the bias and MSE for the vector of parameters $\boldsymbol{\beta}^c$ and $\boldsymbol{\beta}^r$. We considered the hyper-parameter estimators for iETKRR method according to Table 2.

Table 12 presents the results of the comparative study between the robust regression methods for interval-valued variables. Up to 10% of outliers in the sample, the iETKRR method outperforms the other robust approaches and presents a half of computational time in comparison with the second best method. The iETKRR and IRR methods presented a very close performance for $p.out = 15\%$. When the percentage of outliers is greater than 15%, the IRR method showed the best performance. The IQR and SSLR methods were not competitive in this scenario.

3.4.10. Outliers scenario 9

Finally, the scenario 9 considers outliers in both interval-valued variable Z and W, with the presence of atypical values in the midpoints (X^c, Y^c) and ranges (X^r, Y^r), simultaneously. The comparison between the methods is done in term of the bias and MSE for the vector of parameters $\boldsymbol{\beta}^c$ and $\boldsymbol{\beta}^r$. We considered the hyper-parameter estimator S3 for the iETKRR method, according to Table 2.

Table 12
Scenario 8: bias and MSE for the parameter estimates, according to the interval robust regression method.

p.out	Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			t
		Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
5%	*iETKRR*	10.9998	−0.0002	0.1498	1.9998	−0.0002	0.0026	1.0025	0.0025	0.0078	0.9983	−0.0017	0.0033	0.0036
	IRR	10.9984	−0.0016	0.1562	2.0000	0.0000	0.0027	1.0066	0.0066	0.0067	0.9953	−0.0047	0.0029	0.0132
	IQR	11.4194	0.4194	0.4217	1.9404	−0.0596	0.0078	1.0823	0.0823	0.0170	0.9419	−0.0581	0.0078	0.0023
	SSLR	11.0578	0.0578	0.1918	1.9916	−0.0084	0.0033	1.4299	0.4299	0.2184	0.6967	−0.3033	0.1078	0.0139
10%	*iETKRR*	10.9962	−0.0038	0.1641	2.0002	0.0002	0.0028	1.0157	0.0157	0.0079	0.9885	−0.0115	0.0035	0.0048
	IRR	10.9931	−0.0069	0.1669	2.0006	0.0006	0.0028	1.0307	0.0307	0.0100	0.9777	−0.0223	0.0046	0.0132
	IQR	12.4421	1.4421	2.4537	1.7941	−0.2059	0.0492	1.2805	0.2805	0.0953	0.8008	−0.1992	0.0474	0.0025
	SSLR	11.2255	0.2255	0.2501	1.9676	−0.0324	0.0045	1.7999	0.7999	0.6636	0.4363	−0.5637	0.3290	0.0172
15%	*iETKRR*	11.0145	0.0145	0.3925	1.9978	−0.0022	0.0079	1.0297	0.0297	0.0135	0.9783	−0.0217	0.0066	0.0490
	IRR	10.9969	−0.0031	0.1717	2.0002	0.0002	0.0029	1.0564	0.0564	0.0169	0.9592	−0.0408	0.0081	0.0133
	IQR	13.4622	2.4622	7.1321	1.6463	−0.3537	0.1476	1.4374	0.4374	0.2240	0.6877	−0.3123	0.1137	0.0024
	SSLR	11.5377	0.5377	1.9778	1.9220	−0.0780	0.0438	1.8979	0.8979	0.8262	0.3676	−0.6324	0.4092	0.0210
20%	*iETKRR*	14.7681	3.7681	45.8579	1.4465	−0.5535	0.9617	1.0903	0.0903	0.0610	0.9335	−0.0665	0.0326	0.0500
	IRR	11.0028	0.0028	0.1911	1.9995	−0.0005	0.0033	1.1247	0.1247	0.0490	0.9105	−0.0895	0.0241	0. 0130
	IQR	19.0103	8.0103	74.7378	0.8353	−1.1647	1.5819	1.6664	0.6664	0.4938	0.5226	−0.4774	0.2528	0.0023
	SSLR	19.0055	8.0055	77.9001	0.8362	−1.1638	1.6471	1.9858	0.9858	0.9899	0.3057	−0.6943	0.4904	0.0207
30%	*iETKRR*	22.4913	11.4913	139.0338	0.3048	−1.6952	2.9173	1.4733	0.4733	0.4227	0.6547	−0.3453	0.2237	0. 0136
	IRR	11.0071	0.0071	0.2131	1.9991	−0.0009	0.0036	1.6823	0.6823	0.6091	0.5143	−0.4857	0.3072	0. 0135
	IQR	22.5450	11.5450	134.4470	0.3326	−1.6674	2.7887	1.9711	0.9711	0.9751	0.3048	−0.6952	0.4973	0.0022
	SSLR	22.2450	11.2450	127.1242	0.3723	−1.6277	2.6543	2.1131	1.1131	1.2520	0.2175	−0.7825	0.6180	0.0137

Table 13
Scenario 9: bias and MSE for the parameter estimates, according to the interval robust regression method.

p.out	Method	$\hat{\beta}_0^c$			$\hat{\beta}_1^c$			$\hat{\beta}_0^r$			$\hat{\beta}_1^r$			t
		Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE	
5%	*iETKRR*	10.9929	−0.0071	0.3479	2.0008	0.0008	0.0060	0.9990	−0.0010	0.0125	1.0007	0.0007	0.0055	0.0098
	IRR	10.9778	−0.0222	0.2226	2.0032	0.0032	0.0040	0.9985	−0.0015	0.0089	1.0013	0.0013	0.0040	0.0134
	IQR	10.9711	−0.0289	0.2922	2.0043	0.0043	0.0052	0.9971	−0.0029	0.0122	1.0027	0.0027	0.0054	0.0023
	SSLR	10.9786	−0.0214	0.2353	2.0031	0.0031	0.0042	0.9875	−0.0125	0.0248	1.0116	0.0116	0.0120	0.0128
10%	*iETKRR*	10.9736	−0.0264	0.3623	2.0040	0.0040	0.0065	1.0005	0.0005	0.0134	1.0002	0.0002	0.0060	0.0088
	IRR	10.9492	−0.0508	0.3329	2.0080	0.0080	0.0062	1.0006	0.0006	0.0143	1.0006	0.0006	0.0067	0.0135
	IQR	10.9354	−0.0646	0.3733	2.0103	0.0103	0.0068	1.0003	0.0003	0.0159	1.0021	0.0021	0.0072	0.0024
	SSLR	10.9487	−0.0513	0.3117	2.0081	0.0081	0.0057	1.0045	0.0045	0.0304	1.0062	0.0062	0.0147	0.0135
15%	*iETKRR*	10.9509	−0.0491	0.4024	2.0071	0.0071	0.0073	1.0016	0.0016	0.0148	0.9996	−0.0004	0.0068	0.0087
	IRR	10.9199	−0.0801	0.3954	2.0123	0.0123	0.0074	1.0026	0.0026	0.0173	0.9997	−0.0003	0.0082	0.0135
	IQR	10.9093	−0.0907	0.4187	2.0144	0.0144	0.0077	1.0043	0.0043	0.0173	1.0001	0.0001	0.0079	0.0023
	SSLR	10.9237	−0.0763	0.3620	2.0118	0.0118	0.0067	1.0207	0.0207	0.0304	0.9984	−0.0016	0.0144	0.0138
20%	*iETKRR*	10.9604	−0.0396	0.4470	2.0065	0.0065	0.0082	0.9980	−0.0020	0.0180	1.0024	0.0024	0.0083	0.0086
	IRR	10.9339	−0.0661	0.4536	2.0113	0.0113	0.0086	0.9978	−0.0022	0.0208	1.0037	0.0037	0.0099	0.0132
	IQR	10.9264	−0.0736	0.4665	2.0129	0.0129	0.0085	1.0024	0.0024	0.0193	1.0025	0.0025	0.0089	0.0024
	SSLR	10.9368	−0.0632	0.4105	2.0108	0.0108	0.0076	1.0323	0.0323	0.0317	0.9946	−0.0054	0.0147	0.0139
30%	*iETKRR*	10.9524	−0.0476	0.5540	2.0084	0.0084	0.0104	1.0021	0.0021	0.0240	1.0014	0.0014	0.0113	0.0094
	IRR	10.9492	−0.0508	0.5463	2.0113	0.0113	0.0104	1.0058	0.0058	0.0264	1.0014	0.0014	0.0126	0.0135
	IQR	10.9524	−0.0476	0.5315	2.0112	0.0112	0.0098	1.0145	0.0145	0.0219	0.9975	−0.0025	0.0100	0.0024
	SSLR	10.9487	-0.0513	0.4984	2.0108	0.0108	0.0094	1.0855	0.0855	0.0350	0.9714	−0.0286	0.0137	0.0141

Table 13 presents the results of the comparative study This is one of the hardest configuration due the presence of outliers in the midpoints and ranges of the interval-valued variables Z and W. The iETKRR method outperforms the robust methods IRR, IQR and SSRL in all configurations. Moreover, the iETKRR method demonstrates a low computational time in various configurations settings as well as a good performance when the percentage of outliers is 20% or 30%.

3.4.11. iETKRR performance overview

The results of the experimental section suggest that the iETKRR method presents a satisfactory performance in comparison with the other robust regression methods for interval-valued variables (IRR, IQR and SSLR). In the majority of the configuration settings, the iETKRR method outperforms the previous approaches or demonstrates to be a strong competitor model.

In the scenario 0 (clean observations) all the robust methods presented a good performance in terms of bias. However, the iETKRR exhibited the best precision (MSE) for the parameter estimates and the lowest computational time.

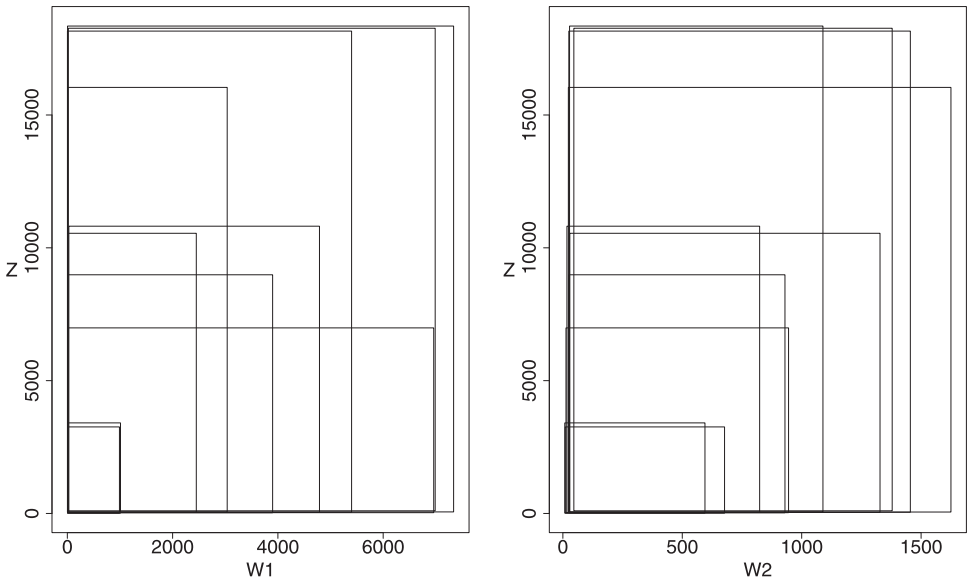

Fig. 2. Interval variables for the Taobao online retail platform.

In the scenarios with the presence of leverage interval-valued observations (3, 6 and 9), the iETKRR method outperformed the other approaches in all configurations. Concerning the presence of X-Space outliers (scenarios 2, 5 and 8), the iETKRR method had the best performance for *p.out* up to 10% and presented a competitive performance for $p.out = 15\%$, which represent outliers percentages that often are found in practice. In comparison with IRR method, the iETKRR method was competitive in scenarios with Y-Space outliers (1, 4 and 7), having the best performance and/or the lower computational time in some specific configurations. Finally, we would like to mention that always $\hat{z}_L < \hat{z}_U$ for iETKRR method. Thus, the rule proposed by Ref. [52] was never required in the experimental study.

4. Applications on real data sets

This section evaluates the proposed robust regression method iETKRR in applications concerning real interval-valued data sets, as well as presents a comparative study in relation to other robust regression methods. The aim is to illustrate the usefulness of the iETKRR method in comparison with the other robust methods SSLR [14], IRR [19] and IQR [21]. We considered seven real data sets with the presence of outliers interval-valued observations. The non-robust method *CRM* [38] is also considered to show the importance of the use of robust approaches to manage interval-valued data with the presence of outliers.

The robustness of each method will be evaluated based on the percentage of change of the parameter estimates when the outliers are suppressed from the data set. Based on this criterion it will be possible to verify which approach will demonstrate more robustness to the presence of outliers.

For the SSLR method, we considered a Student-t distribution for both the midpoint and the range of intervals. The Akaike Information Criterion was used to select the degree of freedom for Student-*t* distribution in a grid between 2 and 10. For the IQR approach [21], we considered the median ($\tau = 0.5$) for the midpoint and for the range of the intervals.

We also consider the interval outlier definition presented in Ref. [19]. An interval-valued observation with Studentized residual larger than 2.0 can be a potential outlier. Moreover, according to the authors, an interval-valued observation can be an outlier in the midpoints, in the radius, or in both. Finally, it is important to emphasize that the outliers presented in the real interval-valued data sets were not inputed, they were labeled as natural outliers observations according to the definition introduced in Ref. [19].

4.1. Taobao seller credit data set

This application refers to a seller credit data set from Taobao, a popular online retail platform in China. The data set is available in Ref. [25]. It has 10 interval-valued observations and 3 interval-valued variables for ten Chinese cities, such as Beijing, Shanghai, and others in 2014. The record of seller credit is considered as response interval-valued variable Z. The popularity of shop (W_1) and the quantity of goods (W_2) are considered as explanatory interval-valued variables.

Fig. 2 presents the scatter plots of the interval-valued data. The interval-valued observation 3 was considered an outlier due to a Studentized residual larger than 2.0 in the midpoint and in the range.

Table 14
Taobao interval-valued data set: percentage change (%) in the parameter estimates after suppression of the outliers observations.

Method	Center			Range		
	$\hat{\beta}_0^c$	$\hat{\beta}_1^c$	$\hat{\beta}_2^c$	$\hat{\beta}_0^r$	$\hat{\beta}_1^r$	$\hat{\beta}_2^r$
CRM	20.39	52.63	21.38	18.57	51.11	20.58
iETKRR$_{(S3,\ S3)}$	5.66	3.39	5.20	3.17	3.33	3.57
IRR	0.06	0.02	0.01	0.04	0.01	0.01
IQR	3.50	33.74	7.85	2.99	25.48	6.67
SSLR$_{(2,\ 2)}$	1.63	5.53	0.69	1.16	4.14	0.56

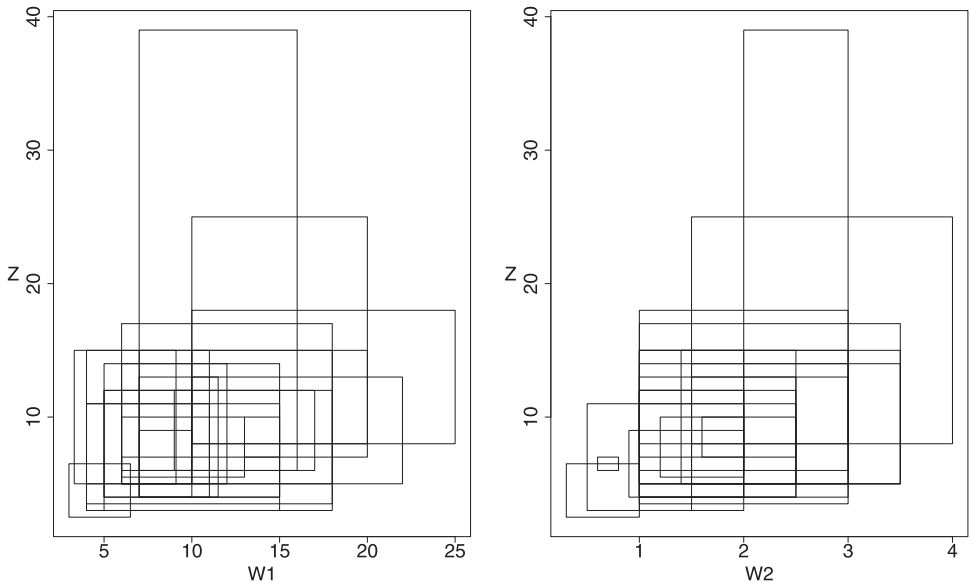

Fig. 3. Interval variables for the Mushroom species.

Table 14 presents the performance of the robust regression methods in this interval-valued data set. Remark that the label iETKRR$_{(S3,\ S3)}$ means that the best performance for this method was provided by the estimator S3 in both width hyper-parameters γ_1 and γ_2, respectively. The convergence of the parameter estimation algorithm occurred with 12 iterations and no negative range was estimated by the iETKRR model, not being required the rule proposed by Ref. [52]. Besides, the label SSLR$_{(2,\ 2)}$ means that the best performance for this method occurred when the selected degree of freedom for the Student-t distribution was equal to 2, in both midpoint and range components of the SSLR model.

One can observe that all robust methods presented a better performance in comparison with the non-robust method CRM. Particularly, the method IRR presented the lowest percentage of change in parameter estimates for the midpoint and range regression models, followed by SSLR$_{(2,\ 2)}$ and iETKRR$_{(S3,\ S3)}$ approaches. The IQR method exhibited the worst performance between the robust regression methods for this interval-valued data set.

4.2. Amanita mushroom data set

This real interval-valued data set presents 23 species of a mushroom family, called Amanita. Each specie is described by two explanatory interval-valued variables, namely stipe length (W_1) and stipe thickness (W_2). The problem is to predict the response interval-valued variables pileus cap (Z) using the explanatory variables. The complete interval-valued data set is available in Ref. [14].

Fig. 3 presents the scatter plots of the interval-valued data. Only the interval-valued observation 12 was considered as an interval outlier due to a Studentized residual larger than 2.0 in the midpoint and in the range, which represents 5% of complete data set. It also was identified as an Y-space outlier.

Table 15 exhibits the performance of the robust regression methods in comparison with the non-robust method CRM. The method iETKRR with the estimator S4 (for the width hyper-parameter γ_1) and the method IRR presented the lowest percentage of change in parameter estimates for the midpoints. Besides, the method iETKRR with the estimator S4 (for the width hyper-parameter γ_2) provided the best performance for the range of intervals. The convergence of the parameter estimation algorithm occurred with 21 iterations and no negative range was estimated by the iETKRR model. The IQR method exhibited the worst performance between the robust regression methods for this interval-valued data set.

Table 15
Mushroom interval-valued data set: percentage change (%) in the parameter estimates after suppression of the outliers observations.

Method	Center			Range		
	$\hat{\beta}_0^c$	$\hat{\beta}_1^c$	$\hat{\beta}_2^c$	$\hat{\beta}_0^r$	$\hat{\beta}_1^r$	$\hat{\beta}_2^r$
CRM	323.31	19.37	33.66	82.19	40.97	72.56
iETKRR$_{(S4,\,S4)}$	8.61	2.96	0.01	1.05	0.26	0.59
IRR	7.14	5.10	0.16	2.56	2.86	2.53
IQR	1362.0	9.52	21.86	53.84	17.95	2.56
SSLR$_{(3,\,3)}$	10.91	0.49	3.11	20.27	4.68	2.04

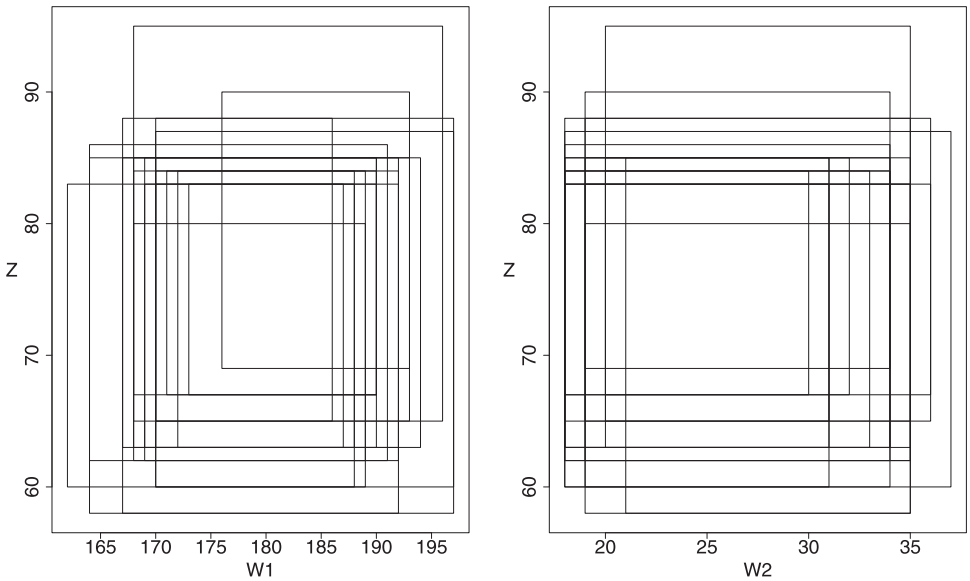

Fig. 4. Interval variables for the Soccer teams.

Table 16
Soccer interval-valued data set: percentage change (%) in the parameter estimates after suppression of the outlier observation.

Method	Center			Range		
	$\hat{\beta}_0^c$	$\hat{\beta}_1^c$	$\hat{\beta}_2^c$	$\hat{\beta}_0^r$	$\hat{\beta}_1^r$	$\hat{\beta}_2^r$
CRM	119.21	13.25	107.08	108.52	10.43	2.46
iETKRR$_{(S3,\,S3)}$	2.68	1.81	2.01	20.91	6.55	1.49
IRR	674.11	7.20	16.32	109.99	10.32	1.61
IQR	308.74	16.93	2.01	95.55	10.90	24.32
SSLR$_{(2,\,2)}$	36.95	7.40	5.29	108.52	10.43	2.46

4.3. Soccer interval-valued data set

This interval-valued data set brings the record of weight (Z), height (W_1) and age (W_2) for 531 soccer players of the French Football Professional Championship grouped in 20 teams. The complete data set can be accessed in Ref. [40].

Fig. 4 presents the scatter plots of the interval-valued data. The interval-valued observation 4 was considered as an interval outlier on the midpoint due to a Studentized residual larger than 2.0. For the ranges of the intervals was not identified outliers. Based on this fact, we considered a Gaussian distribution in the SSLR model for the ranges.

Table 16 exhibits the performance of the robust methods in comparison with the non-robust method CRM. The method iETKRR$_{(S3,\,S3)}$ presented the lowest percentage of change in the parameter estimates for both models (midpoint and range). The convergence of the parameter estimates occurred after 50 iterations and no negative range was estimated by method iETKRR. Note that the other robust methods presented a high percentage of change for the intercept for the midpoint and range models. The IQR method exhibited the worst performance between the robust regression methods in this interval-valued data set. Concerning the ranges of the intervals where no outliers were detected, the iETKRR method outperformed the CRM approach as well as the other robust methods.

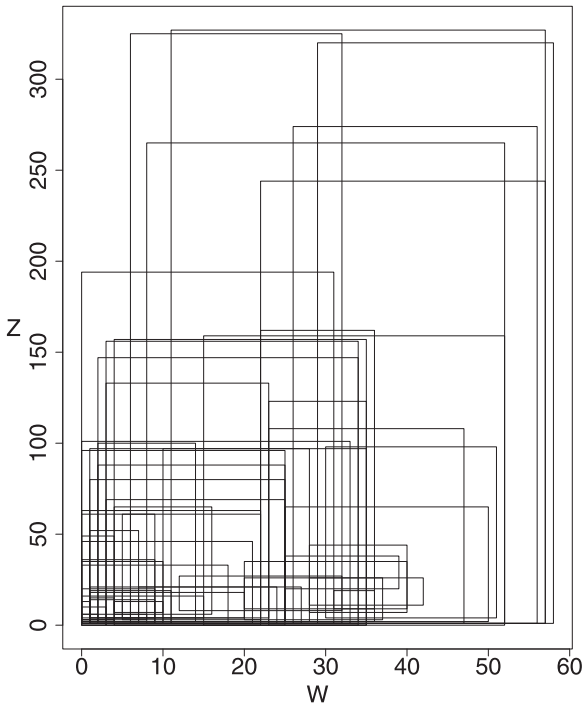

Fig. 5. Interval variables for the unemployment in Portugal data set.

Table 17
Unemployment interval-valued data set: percentage change (%) in the parameter estimates after suppression of the outliers observations.

Method	Center		Range	
	$\hat{\beta}_0^c$	$\hat{\beta}_1^c$	$\hat{\beta}_0^r$	$\hat{\beta}_1^r$
CRM	41.39	33.96	53.20	22.77
iETKRR$_{(S3,\ S2)}$	7.97	18.48	22.70	8.21
IRR	18.04	24.86	34.64	10.09
IQR	204.75	46.08	47.48	15.44
SSLR$_{(2,\ 2)}$	27.59	34.24	35.21	11.62

4.4. Unemployment interval-valued data set

This interval data set brings information about the unemployment in Portugal based on the log-time of unemployment (Z) and the time that people have worked previously (W). The data set presents 58 classes of individuals grouped according to gender, region, age and education. For a particular class, each interval-valued variable (Z and W) represents the minimum and the maximum values observed for the set individuals belonging to this class. The complete interval data set is available in Ref. [13].

Fig. 5 presents the scatter plots of the interval-valued data. According with the interval outlier criterion, we identify 5 outliers in this data set, which represents 8.62% of the whole data set. The observations 12 and 17 were considered outliers in midpoint and range, while the observation 20 presented a discrepancy for the midpoint and the observation 57 a discrepancy for the range.

Table 17 exhibited the performance of the robust methods in comparison with the non-robust method CRM. The method iETKRR$_{(S3,\ S2)}$ presented the lower percentage of change in parameter estimates for the center and range models, respectively. The convergence of the parameter estimates occurred after 31 iterations and 1 negative range was estimated by method iETKRR, representing just 1.72% of the overall data set. In this case, we used the rule proposed by Ref. [52] on this observation. Remark that the other robust methods presented a high percentage of change for the intercept and for the slope parameter of the explanatory variable W, mainly in the center model. The IQR method exhibited the worst performance between the robust regression methods in this interval-valued data set. Particularly, CRM method outperform the IQR method in the model for the centers.

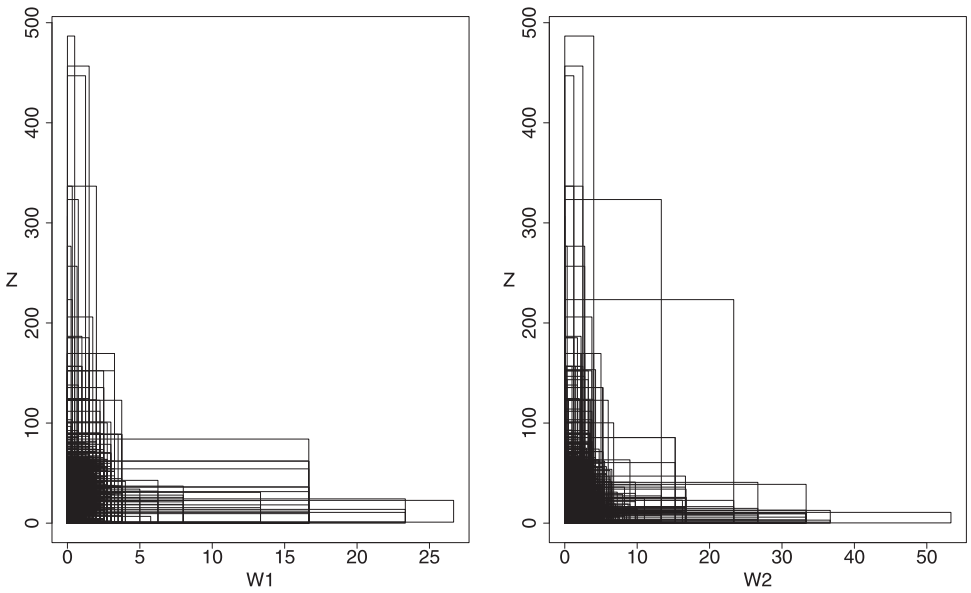

Fig. 6. Interval variables for the scientific production interval data set.

Table 18
Scientific production interval-valued data set: percentage change (%) in the parameter estimates after suppression of the outliers observations.

Method	Center			Range		
	$\hat{\beta}_0^c$	$\hat{\beta}_1^c$	$\hat{\beta}_2^c$	$\hat{\beta}_0^r$	$\hat{\beta}_1^r$	$\hat{\beta}_2^r$
CRM	20.93	58.10	6.02	24.29	58.77	8.07
iETKRR$_{(S2,\,S1)}$	3.50	0.51	1.32	0.43	0.28	0.02
IRR	1.30	1.43	3.33	1.08	0.82	0.63
IQR	7.69	9.52	12.50	6.67	8.89	4.99
SSLR$_{(2,\,2)}$	3.13	2.38	9.09	3.03	1.41	4.81

4.5. Scientific production interval-valued data set

This interval data set brings information about the scientific production in 430 universities and research institutes in Brazil. The original database contains information about 141,260 Brazilian researchers, each one described by continuous numerical variable representing averages of production values computed in three years (2006, 2007 and 2008) for each researcher.

Ref. [42] built an interval-valued data set grouping the original database according to institute and sub-area of knowledge variables. Thus, the resulting interval-valued data set has 5630 items and describes scientific production of the institutes according to the subject area of knowledge.

Ref. [21] considered 3 interval-valued variable and proposed a quantile regression model for interval-valued variables. The explanatory variables are chosen using a priori expert knowledge and they are: NPhd (PhD guidelines finished) and NMaster (Master guidelines finished). These independent variables were considered to explain the response variable "number of work published" (NPub – Work Publications).

Fig. 6 presents the scatter plots for the interval-valued variables. According to the interval outlier criterion of Ref. [19], were identified a total of 153 outliers being 9 outliers on midpoints, 12 outliers on ranges and 132 outliers in the midpoint and range. The total number of outliers represents 2.71% of the complete interval-valued data set.

Table 18 provides the performance of the robust methods in comparison with the non-robust method CRM. The method iETKRR$_{(S2,\,S1)}$ presented the lower percentage of change in parameter estimates for the center and range models, respectively, except for the parameter $\hat{\beta}_0^c$. The convergence of the parameter estimates occurred after 33 iterations and no negative range was estimated by method iETKRR. The IRR method presented the second best performance while the method IQR presented the worst performance between the robust regression methods for interval-valued variables.

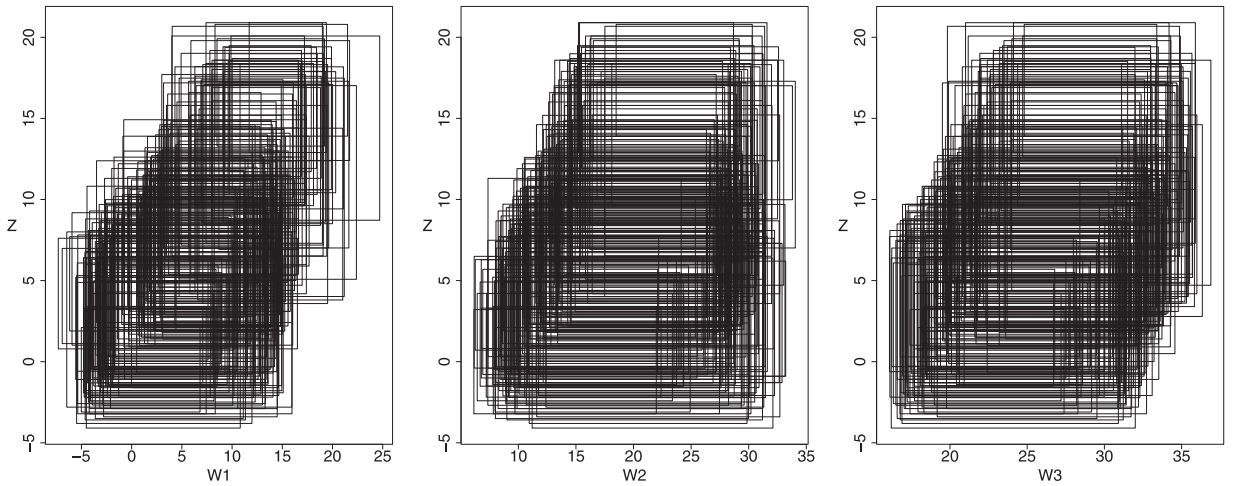

Fig. 7. Interval variables for the temperature in the East of China.

Table 19
Temperature in the East of China interval-valued data set: percentage change (%) in the parameter estimates after suppression of the outliers observations.

Method	Center				Range			
	$\hat{\beta}_0^c$	$\hat{\beta}_1^c$	$\hat{\beta}_2^c$	$\hat{\beta}_3^c$	$\hat{\beta}_0^r$	$\hat{\beta}_1^r$	$\hat{\beta}_2^r$	$\hat{\beta}_3^r$
CRM	16.36	1.09	16.81	1.77	12.82	51.36	20.23	10.35
iETKRR$_{(S3,\,S2)}$	9.07	5.53	14.92	1.59	1.92	1.54	4.82	1.03
IRR	20.42	0.82	12.50	7.14	7.37	44.72	77.31	13.97
IQR	4.63	1.69	0.36	2.65	4.97	108.41	98.86	3.63
SSLR$_{(2,\,2)}$	8.09	0.59	3.94	3.37	1.85	38.52	138.90	11.15

4.6. Temperature in the East of China interval-valued data set

This data set brings the record of the intervals of temperature (Celsius scale) in the East of China in the four quarters, Q_1–Q_4, between the years of 1974 and 1988. The aim is to predict the interval of temperature in the quarter Q_4 (Z) based on the intervals of temperature in the previous quarters Q_1 (W_1), Q_2 (W_2) and Q_3 (W_3). The data set consists of 225 interval-valued observations and can be accessed in the R library MAINT.DATA.

Fig. 7 presents the scatter plots of the interval-valued data. We identified a total of 21 outliers being 12 outliers on midpoints, 11 outliers on ranges and 2 outliers in the midpoint and range. The total number of outliers represents 9.33% of the complete interval-valued data set.

Table 19 exhibits the performance of the robust methods in comparison with the non-robust method CRM. The SSLR$_{(2,\,2)}$ method presented the best performance for the midpoints of the intervals. The method iETKRR$_{(S3,\,S2)}$ presented an intermediate performance in the midpoints. However, the new approach exhibited the lowest percentage of change of the parameter estimates in the range models. The convergence of the parameter estimation algorithm occurred after 62 iterations and no negative range was estimated by the method iETKRR. The robust regression methods IRR and IQR did not presented a good overall performance in this data set.

4.7. Cars interval-valued data set

The last real interval-valued data set presents the record of four car features (Price, Engine Capacity, Top Speed and Acceleration) for 27 cars models. The aim is to predict the Price (Z) based on Engine Capacity Q_1 (W_1), Top Speed (W_2) and Acceleration (W_3). The full data set is available in the R library MAINT.DATA.

Fig. 8 presents the scatter plots of the interval-valued data. We identified a total of 5 outliers being 2 outliers on midpoints and 3 outliers on ranges. This is the real data set with the higher percentage of outliers, corresponding to 18.51% of the total of observations.

The comparative performance between the robust regression methods is presented in Table 20. The method iETKRR$_{(S3,\,S2)}$ presented a good performance for the midpoint and ranges, followed by the methods IRR and SSLR. The convergence of the parameter estimation algorithm occurred after 25 iterations and 1 negative range was estimated by the method iETKRR, representing just 3.70% of the overall data set. In this case, we used the rule proposed by Ref. [52]. The IQR method presented the worst performance for this data set.

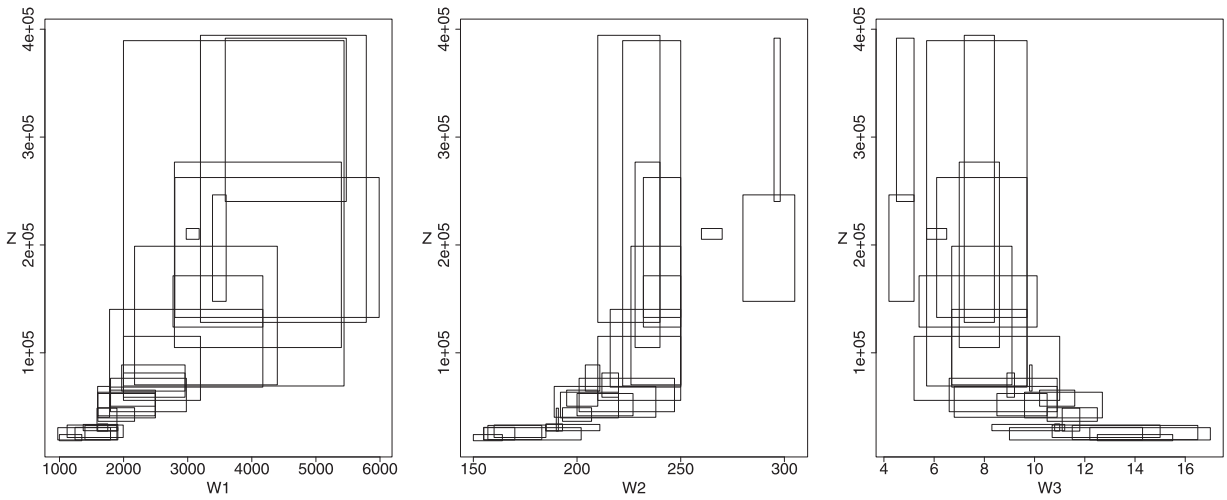

Fig. 8. Interval variables for Cars data set.

Table 20
Cars interval-valued data set: percentage change (%) in the parameter estimates after suppression of the outliers observations.

Method	Center				Range			
	$\hat{\beta}_0^c$	$\hat{\beta}_1^c$	$\hat{\beta}_2^c$	$\hat{\beta}_3^c$	$\hat{\beta}_0^r$	$\hat{\beta}_1^r$	$\hat{\beta}_2^r$	$\hat{\beta}_3^r$
CRM	174.32	12.57	232.02	257.77	45.18	6.65	19.62	2.30
iETKRR$_{(S3,\ S2)}$	182.22	1.30	51.82	56.36	65.96	1.26	74.43	38.03
IRR	144.43	1.91	31.65	64.88	44.90	5.82	382.93	55.38
IQR	196.45	5.62	205.91	30.49	9.62	23.96	118.66	90.66
SSLR$_{(2,\ 2)}$	100.74	1.80	73.98	79.18	92.53	1.16	45.89	110.75

Table 21
Overall performance of the method based on percentage of change's rank.

Method	Taobao	Mushroom	Soccer	Unempl.	Scientif. prod.	Temp.	Cars	Avg. rank
CRM	5	5	4	4.5	5	5	4	4.6
iETKRR	3	1	1	1	1	2	1	1.4
IRR	1	2	3	2	2	3.5	2.5	2.3
IQR	4	4	5	4.5	4	3.5	5	4.3
SSLR	2	3	2	3	3	1	2.5	2.4

4.8. iETKRR overall performance on the real data sets

The performance of the method iETKRR on the seven real interval-valued data sets demonstrates that the proposed approach represents a competitive method when compared with the other robust regression methods presents in literature. Table 21 ranks the methods for each data set based on the percentage of change of the parameter estimates (rank 1 for the lowest percentage of change). As expected, the CRM approach is the most sensitive method to the presence of outliers. The method IQR exhibited the worst performance between the robust regression methods for interval-valued variables. The methods IRR and SSLR have outperformed the previous approaches and have demonstrated a similar performance. The method iETKRR had the best performance based on the average of the rank, corroborating with the results presented in the simulation section. Moreover, the use of different hyper-parameter estimates demonstrated be useful due to the different types of outliers present in the real data sets.

5. Concluding remarks

A robust linear regression method for interval-valued variables based on exponential-type kernel functions was proposed in this paper. The proposed method provides a new objective function that is suitable to manage interval-valued data because it is able to take into account the informations provided either by the center and the radius of the intervals or the lower and upper boundaries of the intervals. Moreover, it allows to combine different hyper-parameter estimators either on the center and on the radius of the intervals or on the lower and upper boundaries of the intervals. Thus, it provides more flexibility and robustness to the proposed model to treat the different types of outliers present in interval-valued data sets.

The Exponential-Type Kernel Robust Regression method for interval-valued variables (iETKRR) of this paper allows the use of exponential-type kernel functions to penalize bad fitted interval-valued observations. The weights given to the midpoints and ranges of each interval-valued observation are updated at each iteration of the parameter estimation algorithm. We used the Gaussian kernel in the parameter estimation algorithm due its mathematical properties.

The proposed parameter estimation algorithm is based on a re-weighted iterative least square process, being necessary initial values of the vector of parameters for the midpoints and ranges of the intervals as well as to the width hyper-parameter estimators of the exponential-type kernel functions. We considered the parameter estimates of CRM approach as initial values.

Besides, a simulation study with synthetic data sets provided insights about the appropriateness of some width hyper-parameter estimators to X-space outliers, Y-space outliers as well as leverage interval-valued points in order to improve the performance of the proposed approach in terms of bias and MSE of the parameter estimates. Moreover, we observed that the iETKRR algorithm provides parameter estimates close to the global minimum that optimizes the objective function, in the majority of scenarios and configurations, and that the fixed maximum number of iterations was never hit.

A comparative study between the iETKRR method and other robust regression approaches for interval-valued variables have demonstrated the usefulness of the proposed method. We have considered synthetic data sets with X-space outliers, Y-space outliers and leverage outlier points, in a Monte Carlo simulation framework with 10,000 replications, different sample sizes and increasing percentage of outliers, in a total of 138 different settings.

The results presented in the experimental section suggested that the iETKRR method presented a satisfactory performance in comparison with the robust regression methods for interval-valued variables (IRR, IQR and SSLR). For the majority of the configuration settings, the iETKRR method outperformed the previous approaches or demonstrated to be a strong competitor model.

The iETKRR exhibited the best precision (MSE) for the parameter estimates and the lowest computational time in the scenario where just clean observations were considered. In scenarios with Y-Space outliers, the proposed approach demonstrated to be a strong competitor model with up to 15% of outliers, exhibiting the best performance for some specific configurations. Concerning the presence of X-Space outliers, the iETKRR method had the best performance with up to 10% of outliers and presented a competitive performance for 15%. In the scenarios with the presence of leverage interval-valued observations, the iETKRR method outperformed the other approaches in all configurations.

The use of different hyper-parameter estimators demonstrated to be useful for the iETKRR approach due to the different types of outliers considered in the experimental analysis as well in the real interval-valued data sets. The new hyper-parameter estimator S4 improved the performance of the iETKRR method in 5 different scenarios when *p.out* was greater than 10%. Besides, concerning benchmark real data sets, it was observed that the iETKRR method presented the best overall performance in comparison with the other robust regression methods for interval-valued variables (IRR, IQR and SSLR). Finally, despite of the usefulness of the iETKRR model, we believe that this one can be further improved if it is able to take into account simultaneously the information of the center and the radius. This is still a challenge and we intend to tackle this topic in a future work.

Acknowledgments

The authors are grateful to the anonymous referees and the Associate Editor for their careful revision, valuable suggestions, and comments which improved this paper. The authors would like to thank CAPES (National Foundation for Post-Graduated Programs, Brazil) and CNPq (National Council for Scientific and Technological Development, Brazil) for their financial support. The second author would like to thank also FACEPE (Reseach Agency from the State of Pernambuco, Brazil).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ins.2018.05.008.

References

[1] J. Ahn, M. Peng, C. Park, Y. Jeon, A resampling approach for interval-valued data regression, Stat. Anal. Data Min. 5 (4) (2012) 336–348.
[2] L. Billard, E. Diday, Regression analysis for interval-valued data, in: Proceedings of the Seventh Conference of the International Federation of Classification Societies on Data Analysis, Classification and Related Methods, 2000, pp. 369–374.
[3] L. Billard, E. Diday, From the statistics of data to the statistics of knowledge: symbolic data analysis, J. Am. Stat. Assoc. 98 (462) (2003) 470–487.
[4] H.H. Bock, E. Diday, editors. Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Heidelberg, 2000.
[5] P. Brito, A.P.D. Silva, Modeling interval data with normal and skew-normal distributions, J. Appl. Stat. 39 (2012) 157–170.
[6] B. Caputo, K. SIM, F. Furesjo, A. Mola, Appearance-based object recognition using svms: which kernel should I use? in: Proceedings of NIPS Workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer Vision, 2002.
[7] F.A.T.D. Carvalho, E.A.L. Neto, M.R.P. Ferreira, A robust regression method based on exponential-type kernel functions, Neurocomputing 234 (2017) 58–74.
[8] E.P.-T. Chang, L. Stanley, A generalized fuzzy weighted least-squares regression, Fuzzy Sets Syst. 82 (1996) 289–298.
[9] S.H. Choi, J.J. Buckley, Least absolute deviation estimator in fuzzy regression, Soft Comput. 12 (2008) 257–263.
[10] R. Coppi, P. D'Urso, P. Giordani, A. Santoro, Least squares estimation of a linear regression model with lr fuzzy response, Comput. Stat. Data Anal. 51 (2006) 267–286.
[11] N. Cristianini, J. Shawe-Taylor, Introduction to Support Vector Machines, Cambridge University Press, Cambridge, 2000.

[12] P. Diamond, Fuzzy least squares, Inf. Sci. 46 (1988) 141–157.
[13] S. Dias, P. Brito, Off the beaten track: a new linear model for interval data, Eur. J. Oper. Res. 258 (3) (2017) 1118–1130.
[14] M.A.O. Domingues, R.M.C.R. de Souza, F.J.A. Cysneiros, A robust method for linear regression of symbolic interval data, Pattern Recognit. Lett. 31 (2010) 1991–1996.
[15] P. D'Urso, Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data, Comput. Stat. Data Anal. 42 (2003) 47–72.
[16] P. D'Urso, T. Gastaldi, A least-squares approach to fuzzy linear regression analysis, Comput. Stat. Data Anal. 34 (2000) 427–440.
[17] P. D'Urso, R. Massari, Weighted least squares and least median squares estimation for the fuzzy linear regression analysis, Metron 71 (2013) 279–306.
[18] P. D'Urso, R. Massari, A. Santoro, Robust fuzzy regression analysis, Inf. Sci. 181 (2011) 4154–4174.
[19] R.A.A. Fagundes, R.M.C.R. de Souza, F.J.A. Cysneiros, Robust regression with application to symbolic interval data, Eng. Appl. Artif. Intell. 26 (2013) 563–573.
[20] R.A.A. Fagundes, R.M.C.R. de Souza, F.J.A. Cysneiros, Interval kernel regression, Neurocomputing 128 (2014) 371–388.
[21] R.A.A. Fagundes, R.M.C.R. de Souza, Y.M.G. Soares, Quantile regression of interval-valued data, in: Proceedings of 23rd International Conference on Pattern Recognition, 2016.
[22] H.-W. Ge, Shi-Tongwang, Dependency between degree of fit and input noise in fuzzy linear regression using non-symmetric fuzzy triangular coefficients, Fuzzy Sets Syst. 158 (2007) 2189–2202.
[23] P. Giordani, Lasso-constrained regression analysis for interval-valued data, Adv. Data Anal. Classif. 9 (1) (2015) 5–19.
[24] G. González-Rivera, W. Lin, Constrained regression for interval-valued data, J. Bus. Econ. Stat. 31 (4) (2013) 473–490.
[25] P. Hao, J. Guo, Constrained center and range joint model for interval-valued symbolic data regression, Comput. Stat. Data Anal. 116 (2017) 106–138.
[26] P.J. Huber, Robust regression: asymptotic, conjectures and monte carlo, Ann. Stat. 1 (1973) 799–991.
[27] P.J. Huber, Robust Statistics, John Wiley and Sons Inc., New York, 1981.
[28] C. Hurvich, J. Simonoff, C. Tsai, Smoothing parameter selection in nonparametric regression using an improved akaike information criterion, J. R. Stat. Soc. B 60 (1998) 271–293.
[29] Y. Jeon, J. Ahn, C. Park, A nonparametric kernel approach to interval-valued data analysis, Technometrics 57 (4) (2015) 566–575.
[30] C. Lim, Interval-valued data regression using nonparametric additive models, J. Korean Stat. Soc. 45 (3) (2017) 358–370.
[31] R.A. Maronna, R.D. Martin, V.J. Yohai, Robust Statistics: Theory and Methods, John Wiley and Sons Inc., Chichester, 2006.
[32] J. Mercer, Functions of positive and negative type and their connection with the theory of integral equations, Philos. Trans. R. Soc. A 209 (1909) 441–458.
[33] Y. Miin-Shen, L. Tzu-Shun, Fuzzy least-squares linear regression analysis for fuzzy input-output data, Fuzzy Sets Syst. 126 (2002) 389–399.
[34] M. Modarres, E. Nasrabadi, M. Nasrabadi, Fuzzy linear regression models with least square errors, Appl. Math. Comput. 163 (2005) 977–989.
[35] K.R. Mueller, S. Mika, G.R. Raetsch, K. Tsuda, B. Schloelkopf, An introduction to kernel-based learning algorithms, IEEE Trans. Neural Netw. 12 (2001) 181–202.
[36] E. Nasrabadi, S.H.M. Ghatee, An lp-based approach to outliers detection in fuzzy regression analysis, Int. J. Unc. Fuzz. Knowl. Based Syst. 15 (2007) 441–456.
[37] E.A.L. Neto, U.U. Anjos, Regression model for interval-valued variables based on copulas, J. Appl. Stat. 42 (2015) 2010–2029.
[38] E.A.L. Neto, F.A.T.D. Carvalho, Centre and range method for fitting a linear regression model to symbolic interval data, Comput. Stat. Data Anal. 52 (3) (2008) 1500–1515.
[39] E.A.L. Neto, F.A.T.D. Carvalho, Constrained linear regression models for symbolic interval-valued variable, Comput. Stat. Data Anal. 54 (2010) 333–347.
[40] E.A.L. Neto, G.M. Cordeiro, F.A.T.D. Carvalho, Bivariate symbolic regression models for interval-valued variables, J. Stat. Comput. Simul. 81 (2011) 1727–1744.
[41] G. Peters, Fuzzy linear regression with fuzzy intervals, Fuzzy Sets Syst. 63 (1994) 45–55.
[42] B.A. Pimentel, R.M.C.R. de Souza, A weighted multivariate fuzzy c-means method in interval-valued scientific production data, Expert Syst. Appl. 41 (2014) 3223–3236.
[43] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016.
[44] P. Rousseeuw, V. Yohai, Robust Regression by Means of S-Estimators, Springer US, New York, NY, pp. 256–272.
[45] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, John Wiley and Sons Inc., New York, 1987.
[46] X. Ruoning, L. Chulin, Multidimensional least-squares fitting with a fuzzy model, Comput. Stat. Data Anal. 119 (2001) 215–223.
[47] B. Scholkoepf, A. Smola, Learning with Kernels, MIT Press, Cambridge MA, 2002.
[48] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.
[49] Z.G. Su, P.H. Wang, Y.G. Li, Z.K. Zhou, Parameter estimation from interval-valued data using the expectation-maximization algorithm, J. Appl. Stat. 85 (2015) 320–338.
[50] H. Tanaka, S. Uejima, K. Asai, Linear regression analysis with fuzzy model, IEEE Trans. Syst. Man Cybern. 12 (1982) 903–907.
[51] Y. Wei, S. Wang, H. Wang, Interval-valued data regression using partial linear model, J. Stat. Comput. Simul. 87 (16) (2017) 3175–3194.
[52] W. Xu, Symbolic Data Analysis: Interval-valued Data Regression, University of Georgia, Athens, 2010 Ph.D. thesis.
[53] V.J. Yohai, High breakdown point and high efficiency robust estimates for regression, Ann. Stat. 15 (1973) 642–656.