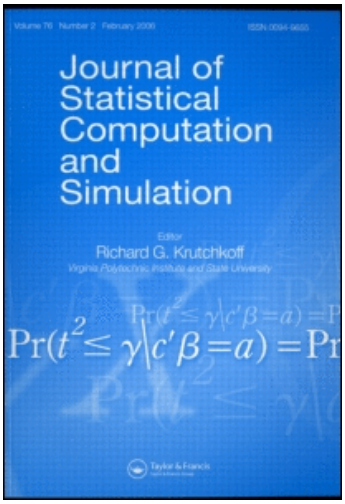

Journal of Statistical Computation and Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713650378

Bivariate symbolic regression models for interval-valued variables

Eufrásio de A. Lima Neto[a]; Gauss M. Cordeiro[b]; Francisco de A. T. de Carvalho[c]
[a] Departamento de Estatística, Universidade Federal da Paraíba, João Pessoa (PB), Brazil [b] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco, Recife (PE), Brazil [c] Centro de Informática, Universidade Federal de Pernambuco, Recife (PE), Brazil

First published on: 07 April 2011

To cite this Article Lima Neto, Eufrásio de A. , Cordeiro, Gauss M. and de Carvalho, Francisco de A. T.(2011) 'Bivariate symbolic regression models for interval-valued variables', Journal of Statistical Computation and Simulation,, First published on: 07 April 2011 (iFirst)
To link to this Article: DOI: 10.1080/00949655.2010.500470
URL: http://dx.doi.org/10.1080/00949655.2010.500470

Journal of Statistical Computation and Simulation
iFirst, 2011, 1–18

Taylor & Francis
Taylor & Francis Group

Bivariate symbolic regression models for interval-valued variables

Eufrásio de A. Lima Neto[a]*, Gauss M. Cordeiro[b] and Francisco de A.T. de Carvalho[c]

[a]*Departamento de Estatística, Universidade Federal da Paraíba, Centro de Ciências Exatas e da Natureza, Cidade Universitária, s/n, CEP 58051-900 João Pessoa (PB), Brazil;* [b]*Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco, Rua Dom Manoel de Medeiros s/n, Dois Irmãos, CEP 52171-900 Recife (PE), Brazil;* [c]*Centro de Informática, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire, s/n, Cidade Universitária, CEP 50740-540 Recife (PE), Brazil*

(*Received 29 September 2009; final version received 8 June 2010*)

Interval-valued variables have become very common in data analysis. Up until now, symbolic regression mostly approaches this type of data from an optimization point of view, considering neither the probabilistic aspects of the models nor the nonlinear relationships between the interval response and the interval predictors. In this article, we formulate interval-valued variables as bivariate random vectors and introduce the bivariate symbolic regression model based on the generalized linear models theory which provides much-needed exibility in practice. Important inferential aspects are investigated. Applications to synthetic and real data illustrate the usefulness of the proposed approach.

Keywords: bivariate symbolic regression method; generalized linear model; deviance; interval-valued data; residual analysis; symbolic data analysis

1. Introduction

Symbolic data analysis (SDA) has been introduced as a domain related to multivariate analysis, pattern recognition and artificial intelligence in order to introduce new methods and to extend classical data analysis techniques and statistical methods to symbolic data [1–3]. In SDA, a variable can assume as a value an interval from a set of real numbers, a set of categories, an ordered list of categories or even a histogram. These new variables could take into account the variability and/or uncertainty presented in the data.

Interval-valued variables have been mainly studied in the area of SDA, where very often an object represents a group of individuals and the variables used to describe it need to assume a value which express the variability inherent to the description of a group. Moreover, interval-valued data arise in practical situations such as recording monthly interval temperatures in meteorological stations, daily interval stock prices, among others. Another source of interval data is the aggregation of huge databases into a reduced number of groups, the properties of which are described

*Corresponding author. Email: eufrasio@de.ufpb.br

ISSN 0094-9655 print/ISSN 1563-5163 online

DOI: 10.1080/00949655.2010.500470
http://www.informaworld.com

by symbolic interval-valued variables. Therefore, tools for interval-valued data analysis are very much required.

Nowadays, different approaches have been introduced to analyse interval-valued data. In the field of SDA, several suitable tools for managing interval-valued data have been discussed in the literature. Bertrand and Goupil [4] and Billard and Diday [5] introduced central tendency and dispersion measures suitable for interval-valued data. De Carvalho [6] proposed histograms for interval-valued data. Concerning factorial methods, Cazes *et al*. [7] and Lauro and Palumbo [8] proposed principal component analysis methods suitable for interval-valued data. Palumbo and Verde [9] and Lauro *et al*. [10] generalized factorial discriminant analysis to interval-valued data. Concerning interval-valued time series, Maia *et al*. [11] have introduced autoregressive integrated moving average (ARIMA), artificial neural network (ANN) as well as a hybrid methodology that combines both ARIMA and ANN models in order to forecast interval-valued time series. Other contributions in the SDA field were proposed by Groenen *et al*. [12] and Ichino *et al*. [13], among others.

SDA provides a number of clustering methods for symbolic data. These methods differ with regard to the type of symbolic data considered, their cluster structures and/or the clustering criteria adopted [14–18]. More recently, De Carvalho [19] introduced adaptive and nonadaptive fuzzy c-means clustering methods for partitioning interval-valued data as well as (fuzzy) cluster and partition interpretation tools.

In the regression analysis of quantitative data, the items are usually represented as a vector of quantitative measurements [20–22]. The generalized linear models (GLMs) represent a major synthesis of regression models by allowing a wide range of types of response data and explanatory variables to be handled in a single unifying framework. These models are based on the exponential family of distributions and represent a very important regression tool due to their flexibility and applicability in practical situations [23]. However, due to recent advances in information technologies, it is now common to record interval-valued data. Concerning the regression analysis, Billard and Diday [24] presented the first approach to fit a linear regression model on interval-valued data sets. Their approach consists of fitting a linear regression model on the midpoint of the interval values assumed by the variables in the learning set and to apply this model on the lower and upper boundaries of the interval values of the explanatory variables to predict, respectively, the lower and upper boundaries of the interval values of the dependent variable. Lima Neto and De Carvalho [25] improved the former approach by presenting a new method based on two linear regression models, the first regression model over the midpoints of the intervals and the second one over the ranges, which reconstruct the boundaries of the interval-values of the dependent variable in a more efficient way when compared with Billard and Diday's method. Alfonso *et al*. [26] extended the regression model methodology in order to include taxonomy variables as predictor variables and to take into account hierarchical structures between symbolic variables. Maia and De Carvalho [27] introduced a least absolute deviation regression model, based on a new optimization criterion, suitable for managing interval-valued data.

Despite some recent valuable contributions in regression models for interval-valued data, the usual regression models attack the problem from an optimization point of view and do not take into account the random nature of the response variable. In this way, the lack of a probabilistic distribution for the response interval-valued variable has made the use of inference techniques over the parameter estimates impossible like, for example, hypothesis tests, residual analysis and diagnostic measures. Another important aspect about the actual symbolic regression methods is that they do not consider nonlinear relationships between the response interval-valued variable and the explanatory interval-valued variables.

In this article, we consider an interval-valued variable assumed to be a bivariate random vector having a joint probability density function belonging to the bivariate exponential family of distributions. Special cases of the bivariate exponential family are the bivariate Gaussian and

the bivariate gamma distributions [28]. The symbolic regression models proposed by Billard and Diday [24] and Lima Neto and De Carvalho [25] do not guarantee mathematical coherence for the predicted values of the interval boundaries ($\hat{y}_L \leq \hat{y}_U$). In order to tackle this problem, we propose bivariate symbolic regression models (BSRM) based on the GLM framework with the random component having the bivariate exponential family of distributions and link functions that guarantee mathematical coherence for the interval boundaries. Further, we propose goodness-of-fit measures, a new definition of residuals for interval and inference techniques such as residual analysis and diagnostic measures. Alternative ways are discussed for estimating the coefficient of correlation ρ and the dispersion parameter ϕ (see [29] for more details concerning the relevance of these problems).

The article is organized as follows. Section 2 reviews the bivariate exponential family of distributions and some of its properties. Section 3 introduces a BSRM based on the GLM framework with different vectors of linear parameters. We also discuss the Fisher scoring method to estimate the linear parameters of the model and alternatives methods to estimate the dispersion parameter and the coefficient of correlation. We also provide the goodness-of-fit statistics, residuals and leverage measures. Section 4 introduces another BSRM model with a common vector of linear parameters. Section 5 presents a comparative study with the methods proposed by Billard and Diday [24] and Lima Neto and De Carvalho [25] in terms of synthetic and real interval-valued data sets. Section 6 ends with some concluding remarks.

2. Probabilistic background

Let $Y = \{y_1, \ldots, y_n\}$ be a set of observations that represents a random sample of the interval-valued variable Y. Each observation $y_i = [y_{Li}, y_{Ui}] \in Y$ is defined as an interval $y \in \Im = \{[y_L, y_U] : y_L, y_U \in \Re, y_L \leq y_U\}$ and represents the observed value of the interval variable Y. An interval of real values is an infinity list of values and it is difficult to take into account the whole information inside it. Despite the loss of information, we consider an interval-valued variable Y as a two-dimensional or a bivariate quantitative feature vector $\mathbf{y}_i = (y_{1i}, y_{2i})$, where the variables Y_1 and Y_2 are one-dimensional random variables representing, for example, the lower and upper boundaries or the midpoint and half-range of the intervals or any other pair of interval features possible to be represented.

Consider that the joint density probability function of the bivariate quantitative feature vector $\mathbf{y}_i = (y_{1i}, y_{2i})$ belongs to the bivariate exponential family of distributions [28,29] defined by

$$f(\mathbf{y}; \boldsymbol{\theta}) = \exp[\phi^{-1}\{y_1\theta_1 + y_2\theta_2 - b(\theta_1, \theta_2, \rho)\} + c(y_1, y_2, \rho, \phi)], \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is the vector of canonical parameters, ϕ is a common dispersion parameter and ρ is a constant correlation parameter between these two random variables. We assume that the functions $b(\cdot, \cdot, \cdot)$ and $c(\cdot, \cdot, \cdot, \cdot)$ are known. The function $b(\cdot, \cdot, \cdot)$ is the cumulant generating function of Equation (1) and the mean and variance of the bivariate random vector $\mathbf{Y} = (Y_1, Y_2)$ can be obtained from well-known equations of natural exponential families.

The bivariate exponential family of distributions has important properties for the expected value, variance–covariance matrix and a direct relation with the GLM framework. All these characteristics will be considered here to fit a symbolic regression model with probabilistic assumptions.

Let $\mathbf{y}_i = (y_{1i}, y_{2i})$ for $i = 1, \ldots, n$ be independent observations taken from Equation (1). The log-likelihood function for the ith observation can be written as

$$l_i = l_i(\boldsymbol{\theta}, \phi, \rho) = \phi^{-1}\{y_{1i}\theta_1 + y_{2i}\theta_2 - b(\theta_1, \theta_2, \rho)\} + c(y_{1i}, y_{2i}, \rho, \phi). \tag{2}$$

The mean of the random variable Y_1 can be easily obtained from the well-known relationship

$$E\left(\frac{\partial l}{\partial \theta_1}\right) = E\left(\frac{y_1 - \partial b/\partial \theta_1}{\phi}\right) = \frac{\mu_1 - \partial b/\partial \theta_1}{\phi} = 0,$$

and then $\mu_1 = \partial b/\partial \theta_1 = b^{(1)}$, where $b = b(\theta_1, \theta_2, \rho)$ and from now on, the superscripts (1) and (2) indicate derivatives with respect to the canonical parameters θ_1 and θ_2, respectively. Analogously, the mean of Y_2 is $\mu_2 = \partial b/\partial \theta_2 = b^{(2)}$.

The mean parameters μ_1 and μ_2 are functions of the canonical parameters θ_1 and θ_2 and of the correlation parameter ρ. Hence, the canonical parameters are also functions of the mean parameters μ_1 and μ_2 and ρ, say

$$\theta_i = q_i(\mu_1, \mu_2, \rho)$$

for i={1,2}. The variance of the random variable Y_1 is obtained from the identity

$$E\left(\frac{\partial^2 l}{\partial \theta_1^2}\right) + E\left(\frac{\partial l}{\partial \theta_1}\right)^2 = -\frac{\partial^2 b/\partial \theta_1^2}{\phi} + \frac{\mathrm{Var}(Y_1)}{\phi^2} = 0$$

and then

$$-\frac{b^{(11)}}{\phi} + \frac{\mathrm{Var}(Y_1)}{\phi^2} = 0,$$

which gives

$$\mathrm{Var}(Y_1) = \phi b^{(11)}.$$

In the same way, the variance of the random variable Y_2 is $\mathrm{Var}(Y_2) = \phi b^{(22)}$.

The covariance between the random variables Y_1 and Y_2 follows from the regularity condition

$$E\left(\frac{\partial^2 l}{\partial \theta_1 \partial \theta_2}\right) + E\left(\frac{\partial l}{\partial \theta_1}\frac{\partial l}{\partial \theta_2}\right) = 0$$

and then

$$\mathrm{Cov}(Y_1, Y_2) = \phi b^{(12)}.$$

Hence, the covariance matrix of $\mathbf{Y}$ can be written as

$$\Sigma = \phi \begin{bmatrix} b^{(11)} & b^{(12)} \\ b^{(21)} & b^{(22)} \end{bmatrix} = \phi \begin{bmatrix} V^{(1)} & V^{(12)} \\ V^{(21)} & V^{(2)} \end{bmatrix},$$

where $V^{(j)}$ is the variance function of the random variable Y_j and $V^{(jk)}$ is the covariance function between the random variables Y_j and Y_k, for $j, k = 1, 2$.

3. Bivariate symbolic regression model 1

Initially, we consider an interval-valued random variable Y represented by a pair of intervals features (Y_1, Y_2) such as the lower and upper bounds (Y^L, Y^U) or the midpoint and half-range (Y^m, Y^r) or any other pair of interval features. Then, we consider that this pair of interval features (Y_1, Y_2) belongs to a bivariate joint distribution. The choice of the bivariate exponential family of distributions allows us to extend the GLM framework for the case of interval-valued variables. Methods based on the midpoint and range of intervals have been discussed in the literature. D'Urso and Gastaldi [30] suggested that the dependence between the centre and the range is often

encountered in real-world applications. Particularly, in the SDA field, Lauro and Palumbo [8] discussed that some principal component approaches are based on midpoint of intervals, and Maia and De Carvalho [27] proposed a time series model for interval-valued data, taking into account the midpoint and range of the intervals. Thus, the BSRM presented here can be easily defined in terms of any pair of interval features (Y_1,Y_2) – for example, the lower and upper boundaries of the interval-valued variable (Section 5).

Let $E = \{e_1, \ldots, e_n\}$ be a set of examples that are described by $p+1$ interval-valued variables Y, $T_1, \ldots, T_p$. The interval-valued variable Y is a dependent variable and it is related to a set of interval-valued variables T_j ($j = 1, 2, \ldots, p$), known as independent variables. Each example $e_i \in E$ ($i = 1, \ldots, n$) is denoted by an interval quantitative feature vector $(\mathbf{t}_i, y_i)$, with $\mathbf{t}_i = (t_{i1}, \ldots, t_{ip})$, where $t_{ij} = [a_{ij}, b_{ij}] \in \Im = \{[a, b] : a, b \in \Re, a \leq b\} (j = 1, \ldots, p)$ and $y_i = [y_{Li}, y_{Ui}] \in \Im$ are the observed values of T_j and Y, respectively.

Now, let Y_1, X_{1j} and Y_2, X_{2j} ($j = 1, 2, \ldots, p$) be the quantitative variables that represent the lower and upper bounds or the midpoints and the half-range of the intervals defined by the interval-valued variables Y and T_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 Y and T_j, each example $e_i \in E$ ($i = 1, \ldots, n$) will be denoted 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 $x_{ij}^L = a_{ij}$, $x_{ij}^U = b_{ij}$, $y_i^L = y_{Li}$ and $y_i^U = y_{Ui}$ 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} ($j = 1, 2, \ldots, p$) represent, respectively, the midpoint and half-range of the interval variables Y and T_j, each example $e_i \in E$ ($i = 1, \ldots, n$) will be denoted by the vectors $(\mathbf{x}_i^m, y_i^m)$ and $(\mathbf{x}_i^r, y_i^r)$, with $\mathbf{x}_i^m = (x_{i1}^m, \ldots, x_{ip}^m)$ and $\mathbf{x}_i^r = (x_{i1}^r, \ldots, x_{ip}^r)$, where $x_{ij}^m = (a_{ij} + b_{ij})/2$, $x_{ij}^r = (b_{ij} - a_{ij})/2$, $y_i^m = (y_{Li} + y_{Ui})/2$ and $y_i^r = (y_{Ui} - y_{Li})/2$ are the observed values of the variables X_j^m, X_j^r, Y^m and Y^r, respectively.

Following the GLM framework, we consider a BSRM1 with probabilistic support defined by two components (a random component and a systematic component) to model interval-valued data. The random component considers the bivariate random vector

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix},$$

having the bivariate exponential family (1). In the systematic component, the explanatory variables X_{1j} and X_{2j} ($j = 1, 2, \ldots, p$) are responsible for the variability of Y_1 and Y_2, respectively, and they are defined by

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}_1) = \mathbf{X}_1\boldsymbol{\beta}_1 \quad \text{and} \quad \boldsymbol{\eta}_2 = g_2(\boldsymbol{\mu}_2) = \mathbf{X}_2\boldsymbol{\beta}_2, \tag{3}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are known model matrices formed by the observed values of the variables X_{1j} and X_{2j}, respectively, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of parameters to be estimated, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are the linear predictors, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean of the response variables Y_1 and Y_2, respectively, with $\boldsymbol{\eta}_1 = (\eta_{1_1}, \ldots, \eta_{1_n})^{\mathrm{T}}$, $\boldsymbol{\mu}_1 = (\mu_{1_1}, \ldots, \mu_{1_n})^{\mathrm{T}}$ and $\boldsymbol{\beta}_1 = (\beta_{1_0}, \ldots, \beta_{1_p})^{\mathrm{T}}$. In the same way, we have $\boldsymbol{\eta}_2 = (\eta_{2_1}, \ldots, \eta_{2_n})^{\mathrm{T}}$, $\boldsymbol{\mu}_2 = (\mu_{2_1}, \ldots, \mu_{2_n})^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (\beta_{2_0}, \ldots, \beta_{2_p})^{\mathrm{T}}$. Here, $g_1(\boldsymbol{\mu}_1)$ and $g_2(\boldsymbol{\mu}_2)$ are well-known link functions that connect the mean of the response variables Y_1 and Y_2 with the explanatory variables X_{1j} and X_{2j} ($j = 1, \ldots, p$), respectively.

In the BSRM1, it is possible to choose different link functions. A few functions available for the BSRM are *identity*, *logarithmic*, *inverse*, *power*, among others. However, some link functions have particular properties and may be preferred in some situations. For example, if we consider

the half-range of intervals in the random component, the logarithmic link function will guarantee positiveness for the predicted values of $\hat{y}_i^r$ ($\hat{y}_i^r > 0$) and this result implies that $\hat{y}_i^L \leq \hat{y}_i^U$, $i = 1, \ldots, n$.

3.1. *Parameter estimation*

The maximum-likelihood method will be used as a theoretical basis for parameter estimation in the BSRM1. For maximizing the log likelihood, we first assume that ρ is fixed and then obtain the likelihood equations for estimating $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Both vectors can be estimated without the knowledge of ϕ. In principle, ϕ could also be estimated by maximum likelihood although there may be practical difficulties associated with this method for some bivariate distributions in Equation (1). Next, we give a simple way to estimate ϕ based on the deviance of the model.

An algorithm for estimating these vectors of linear parameters can be developed from the scoring method. Differentiating the total log likelihood (2) yields the score functions for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$

$$U(\beta_{1j}) = \frac{\partial l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)}{\partial \beta_{1j}} = \frac{1}{\phi}\sum_{i=1}^{n}(y_{1i} - b_i^{(1)})\frac{\partial \theta_1}{\partial \beta_{1j}} = \frac{1}{\phi}\sum_{i=1}^{n}(y_{1i} - \mu_{1i})\frac{1}{V_i^{(1)} g_1'(\mu_{1i})}x_{1_{ij}}$$

and

$$U(\beta_{2j}) = \frac{\partial l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)}{\partial \beta_{2j}} = \frac{1}{\phi}\sum_{i=1}^{n}(y_{2i} - b_i^{(2)})\frac{\partial \theta_2}{\partial \beta_{2j}} = \frac{1}{\phi}\sum_{i=1}^{n}(y_{2r} - \mu_{2i})\frac{1}{V_i^{(2)} g_2'(\mu_{2i})}x_{2_{ij}}.$$

In matrix notation, the score functions are

$$\mathbf{U}(\boldsymbol{\beta}_1) = (\mathbf{X}_1)^{\mathrm{T}}\mathbf{W}_1\mathbf{z}_1 \quad \text{and} \quad \mathbf{U}(\boldsymbol{\beta}_2) = (\mathbf{X}_2)^{\mathrm{T}}\mathbf{W}_2\mathbf{z}_2, \tag{4}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are diagonal weighted matrices with corresponding elements

$$w_{1i} = [V_i^{(1)} g_1'(\mu_{1i})^2]^{-1} \quad \text{and} \quad w_{2i} = [V_i^{(2)} g_2'(\mu_{2i})^2]^{-1},$$

and $\mathbf{z}_1$ and $\mathbf{z}_2$ are modified dependent variables related to $\mathbf{y}_1$ and $\mathbf{y}_2$ given by

$$\mathbf{z}_1 = \mathbf{G}_1(\mathbf{y}_1 - \boldsymbol{\mu}_1) \quad \text{and} \quad \mathbf{z}_2 = \mathbf{G}_2(\mathbf{y}_2 - \boldsymbol{\mu}_2),$$

where $\mathbf{G}_1 = \mathrm{diag}\{g_1'(\mu_{1_1}), \ldots, g_1'(\mu_{1_n})\}$ and $\mathbf{G}_2 = \mathrm{diag}\{g_2'(\mu_{2_1}), \ldots, g_2'(\mu_{2_n})\}$ are $n \times n$ diagonal matrices.

The expected value of the svth component of the information matrix for the vector of parameters $\boldsymbol{\beta}_1$ is

$$\kappa_{1(s,v)} = -E\left[\frac{\partial^2 \boldsymbol{\beta}_1}{\partial \beta_{1s}\partial \beta_{1v}}\right] = E\left[\frac{\partial l(\boldsymbol{\beta}_1)}{\partial \beta_{1s}}\frac{\partial l(\boldsymbol{\beta}_1)}{\partial \beta_{1v}}\right] = E[U(\beta_{1s})U(\beta_{1v})].$$

In matrix notation, the information matrices for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be written as

$$\mathbf{K}_1 = \phi^{-1}(\mathbf{X}_1)^{\mathrm{T}}\mathbf{W}_1\mathbf{X}_1 \quad \text{and} \quad \mathbf{K}_2 = \phi^{-1}(\mathbf{X}_2)^{\mathrm{T}}\mathbf{W}_2\mathbf{X}_2.$$

From the information matrices and Equations (4), we can use the scoring method to obtain the conditional maximum-likelihood estimates (MLEs) of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for a given ρ. We have

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + (\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(k)}\mathbf{z}^{(k)}, \tag{5}$$

where

$$\boldsymbol{\beta}^{(k+1)} = \begin{bmatrix} \boldsymbol{\beta}_1^{(k+1)} \\ \boldsymbol{\beta}_2^{(k+1)} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix},$$
$$\mathbf{W}^{(k)} = \begin{bmatrix} \mathbf{W}_1^{(k)} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2^{(k)} \end{bmatrix} \quad \text{and} \quad \mathbf{z}^{(k)} = \begin{bmatrix} \mathbf{z}_1^{(k)} \\ \mathbf{z}_2^{(k)} \end{bmatrix}.$$

Equations (5) have the same form of the estimating equations for the GLMs. In general terms, we regress the modified dependent variable $\mathbf{z}^{(k)}$ on the local model matrix $\mathbf{X}$ by taking $\mathbf{W}^{(k)}$ as a modified weighted matrix. At $k = 1$, an initial approximation $\boldsymbol{\beta}^{(1)}$ could be used to evaluate $\mathbf{W}^{(1)}$ and $\mathbf{z}^{(1)}$ from which Equations (5) yield the next estimate $\boldsymbol{\beta}^{(2)}$. Hence, we update $\mathbf{W}^{(2)}$ and $\mathbf{z}^{(2)}$, and so the iterations continue until the convergence is achieved and then the conditional MLE $\hat{\boldsymbol{\beta}}$ is obtained. In general, the convergence speed is fast, but it strongly depends on the choice of the initial value $\boldsymbol{\beta}^{(1)}$.

3.2. *Goodness-of-fit measures*

Conditioned on the parameter ρ, the discrepancy of a BSRM1 can be defined as twice the difference between the maximum log likelihood achievable and that achieved for the model under investigation. The discrepancy is known as the deviance of the current model and has the form of a genuine GLM deviance, since it is a function of the data only and of the MLEs $\hat{\mu}_{1i}$ and $\hat{\mu}_{2i}$, for $i = 1, \ldots, n$, which are calculated from the data. Hence, the deviance conditioned on ρ, say $D(\rho)$, can be written as

$$D(\rho) = 2\sum_{i=1}^{n}\{y_{1i}[q_1(y_{1i}, \rho) - q_1(\hat{\mu}_{1i}, \rho)] + y_{2i}[q_2(y_{2i}, \rho) - q_2(\hat{\mu}_{2i}, \rho)] + [b(q_1(\hat{\mu}_{1i}, \rho), q_2(\hat{\mu}_{2i}, \rho), \rho) - b(q_1(y_{1i}, \rho), q_2(y_{2i}, \rho), \rho)]\}. \quad (6)$$

Another measure of discrepancy of easy interpretation of a BSRM1 is the generalized Pearson X^2 statistic (defined conditioned on the parameter ρ) by

$$X^2(\rho) = \sum_{i=1}^{n}\left[\frac{(y_{1i} - \hat{\mu}_{1i})^2}{V(\hat{\mu}_{1i})} + \frac{(y_{2i} - \hat{\mu}_{2i})^2}{V(\hat{\mu}_{2i})}\right].$$

Maximum-likelihood estimation of the dispersion parameter ϕ directly is a more difficult problem than the estimation of $\boldsymbol{\beta}$ and the complexity depends entirely on the functional form of the function $c(y_1, y_2, \rho, \phi)$. For some BSRMs, the MLE of the dispersion parameter could be very complicated, but we can use a method of deviance estimator to obtain a consistent estimate of the dispersion parameter ϕ from the estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ obtained from Equation (5), with dimensions $p_1 \times 1$ and $p_2 \times 1$, respectively. The deviance can be approximated by a χ^2_ν distribution with $\nu = 2n - (p_1 + p_2)$ degrees of freedom, which leads to a simple estimate of ϕ

$$\tilde{\phi} = \frac{D(\rho)}{2n - (p_1 + p_2)}, \quad (7)$$

based on the fact that the deviance can be approximated by a chi-squared distribution.

Substituting the estimates $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$ and $\tilde{\phi}$ in Equation (2) yields the profile log likelihood for the parameter ρ

$$l_p(\rho) = \phi^{-1} \sum_{i=1}^{n} \{y_{1i}\hat{\theta}_1 + y_{2i}\hat{\theta}_2 - b(\hat{\theta}_1, \hat{\theta}_2, \rho)\} + \sum_{i=1}^{n} c(y_{1i}, y_{2i}, \rho, \tilde{\phi}). \tag{8}$$

In the next step, the procedure calculates the profile log likelihood $l_p(\rho)$ in Equation (8) for a trial series of values of $\rho \in [-1, 1]$ and determines numerically the value of the estimate $\hat{\rho}$ that maximizes $l_p(\rho)$. Once the estimate $\hat{\rho}$ is obtained, it can be substituted into the algorithm (5) to produce the new conditional estimate $\hat{\boldsymbol{\beta}}$ and then substituting in Equation (7) to obtain a new estimate $\tilde{\phi}$. The new values of $\hat{\boldsymbol{\beta}}$ and $\tilde{\phi}$ can update $\hat{\rho}$, and so the iterations continue until convergence is observed. In other words, we use a see-saw algorithm for maximizing the log likelihood. Holding ρ as a prior constant at the current estimate, the estimate of $\hat{\boldsymbol{\beta}}$ is obtained from Equation (5) and the estimate of $\tilde{\phi}$ comes from Equation (7). Holding $\hat{\boldsymbol{\beta}}$ and $\tilde{\phi}$ fixed at the current estimates, the estimate of ρ is obtained by maximizing Equation (8). Cycling between holding both $\hat{\boldsymbol{\beta}}$ and $\tilde{\phi}$ fixed and holding ρ fixed will lead to the unconditional estimates of $\boldsymbol{\beta}$, ϕ and ρ. The joint iterative process of estimating these parameters can be carried out by standard statistical software such as MATLAB, S-PLUS, R and SAS.

3.3. *Residuals and diagnostic measures*

In this section, we present some residual definitions and diagnostic measures that are useful for making inferences about the response distribution, identify outliers, among others aspects. An important contribution in this section is a unique residual definition for an interval-valued data. Usually, the residuals are defined separately for each boundary of the interval.

The projection matrix $\mathbf{H}$ for the BSRM1 takes the form

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{1/2}, \tag{9}$$

which is equivalent to replacing $\mathbf{X}$ by $\mathbf{W}^{1/2}\mathbf{X}$, which effectively allows for the change in variance with the mean. Here,

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix}.$$

The well-known measure of leverage is given by the diagonal elements of the projection matrix. The leverage measures of the vector $\mathbf{y}_i = (y_{1i}, y_{2i})$ can be represented by the diagonal elements $h_{1_{ii}}$ and $h_{2_{ii}}$ of the matrices $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively. However, the matrix $\mathbf{H}$ depends on the explanatory variables, the link and variance functions, making the interpretation of these measures much more difficult. We have $\sum_i h_{1_{ii}} = p_1$ and $\sum_i h_{2_{ii}} = p_2$. Hence, the interval-valued observations for the response variable Y with high leverage are indicated by $h_i = (h_{1_{ii}} + h_{2_{ii}})$ greater than $2(p_1 + p_2)/n$. An index plot of each h_i versus i with this lower limit could be an useful informal tool for looking at leverage.

Some residual measures commonly used in the GLM theory can be easily extended to BSRM1. The residual related to the ith vector of observations (y_{1i}, y_{2i}) can be composed by two parts: the residual for the observation y_{1i} and the residual for the observation y_{2i}. The Pearson residual is given by

$$r_{1i}^{P} = \frac{y_{1i} - \hat{\mu}_{1i}}{\sqrt{\widehat{V}_{1i}}} \quad \text{and} \quad r_{2i}^{P} = \frac{y_{2i} - \hat{\mu}_{2i}}{\sqrt{\widehat{V}_{2i}}}. \tag{10}$$

The Studentized Pearson residual which has a constant variance when $\phi \to 0$ can be expressed as

$$r_{1i}^{SP} = \frac{y_{1i} - \hat{\mu}_{1i}}{\sqrt{V(\hat{\mu}_{1i})(1 - \hat{h}_{1_{ii}})}} \quad \text{and} \quad r_{2i}^{SP} = \frac{y_{2i} - \hat{\mu}_{2i}}{\sqrt{V(\hat{\mu}_{2i})(1 - \hat{h}_{2_{ii}})}}. \tag{11}$$

Notice that the Pearson and Studentized Pearson residuals denote the discrepancy between the observed and predicted values for each boundary of the interval, separately.

On the other hand, the deviance residual can be interpreted as a joint residual measure or a global residual measure for the vector of observations (y_{1i}, y_{2i}). The ith deviance residual is defined in terms of the square root of the contribution of the ith observation to the deviance (6). Conditioned on ρ, the deviance residual can be expressed as

$$r_i^D = \text{sign}[(y_{1i} - \hat{\mu}_{1i}) + (y_{2i} - \hat{\mu}_{2i})]\sqrt{d_i}, \tag{12}$$

where

$$\begin{aligned} d_i = 2\{&y_{1i}[q_1(y_{1i}, \rho) - q_1(\hat{\mu}_{1i}, \rho)] + y_{2i}[q_2(y_{2i}, \rho) - q_2(\hat{\mu}_{2i}, \rho)] \\ &+ b(q_1(\hat{\mu}_{1i}, \rho), q_2(\hat{\mu}_{2i}, \rho), \rho) - b(q_1(y_{1i}, \rho), q_2(y_{2i}, \rho), \rho)\}. \end{aligned}$$

An overall measure of the influence of the ith vector of observations on the parameter estimates is the scaled-likelihood distance that can be defined similar to the GLMs [31] as

$$\text{LD}_i = \frac{2}{p}\{l(\hat{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}_{(i)})\}, \tag{13}$$

where $\hat{\boldsymbol{\beta}}$ is the estimate with all vector points and $\hat{\boldsymbol{\beta}}_{(i)}$ is obtained when the ith vector is deleted. Following [32], we can expand Equation (13) in Taylor series to rewrite this expression in terms of the generalized Cook distance. Hence, the generalized Cook distance for the BSRM1 in order to measure the influence of the ith vector (y_{1i}, y_{2i}) can be defined by

$$D_i = D_{1i} + D_{2i}, \tag{14}$$

where

$$D_{1i} = \frac{h_{1_{ii}}}{p_1(1 - h_{1_{ii}})}(r_{1i}^P)^2 \quad \text{and} \quad D_{2i} = \frac{h_{2_{ii}}}{p_2(1 - h_{2_{ii}})}(r_{2i}^P)^2.$$

Values of D_i higher than $\chi^2_{p_1+p_2,\alpha}/(p_1 + p_2)$ can be considered influential, where $\chi^2_{p_1+p_2,\alpha}$ is the upper $100(1 - \alpha)\%$ point of the $\chi^2_{p_1+p_2}$ distribution.

4. Bivariate symbolic regression model 2

The BSRM2 takes the random component following Equation (1) and considers the same vector of linear parameters in the systematic component given by

$$\boldsymbol{\eta}_1 = g_1(\boldsymbol{\mu}_1) = \mathbf{X}_1\boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\eta}_2 = g_2(\boldsymbol{\mu}_2) = \mathbf{X}_2\boldsymbol{\beta}, \tag{15}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are known model matrices, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\text{T}}$ is the vector of parameters to be estimated, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are the linear predictors and $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean of the response variables Y_1 and Y_2, respectively, with the same dimensions as defined in Section 3.

The jth component of the score function $\mathbf{U}(\boldsymbol{\beta})$ for the common vector of parameters $\boldsymbol{\beta}$ is given by

$$U(\beta_j) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\left[(y_{1i} - b_i^{(1)})\frac{\partial\theta_1}{\partial\mu_1}\frac{\partial\mu_1}{\partial\eta_1}\frac{\partial\eta_1}{\partial\beta_j} + (y_{2i} - b_i^{(2)})\frac{\partial\theta_2}{\partial\mu_2}\frac{\partial\mu_2}{\partial\eta_2}\frac{\partial\eta_2}{\partial\beta_j}\right],$$
$$= \frac{1}{\phi}\sum_{i=1}^{n}\left[(y_{1i} - \mu_{1i})\frac{x_{1_{ij}}}{V_i^{(1)}g_1'(\mu_{1i})} + (y_{2i} - \mu_{2i})\frac{x_{2_{ij}}}{V_i^{(2)}g_2'(\mu_{2i})}\right].$$

In matrix notation, the score functions reduce to

$$\mathbf{U}(\boldsymbol{\beta}) = (\mathbf{X}_1)^{\mathrm{T}}\mathbf{W}_1\mathbf{z}_1 + (\mathbf{X}_2)^{\mathrm{T}}\mathbf{W}_2\mathbf{z}_2, \tag{16}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are diagonal weighted matrices, and $\mathbf{z}_1$ and $\mathbf{z}_2$ are modified dependent variables related to $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively, as defined in Section 3.1. The information matrix for $\boldsymbol{\beta}$ is

$$\mathbf{K} = \frac{1}{\phi}[(\mathbf{X}_1)^{\mathrm{T}}\mathbf{W}_1\mathbf{X}_1 + (\mathbf{X}_2)^{\mathrm{T}}\mathbf{W}_2\mathbf{X}_2]. \tag{17}$$

The scoring method to obtain the MLE of $\boldsymbol{\beta}$ becomes

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + (\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{W}^{(k)}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{W}^{(k)}\mathbf{z}^{(k)}, \tag{18}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}, \quad \mathbf{W}^{(k)} = \begin{bmatrix}\mathbf{W}_1^{(k)} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_2^{(k)}\end{bmatrix} \quad \text{and} \quad \mathbf{z}^{(k)} = \begin{bmatrix}\mathbf{z}_1^{(k)}\\ \mathbf{z}_2^{(k)}\end{bmatrix}.$$

4.1. *Model checking and residuals*

The formulae derived in Section 3 for the deviance of a BSRM1 and to estimate the dispersion parameter ϕ can be applied for the systematic component (15). Further, the computational procedure to estimate ρ and the see-saw algorithm are still valid here. However, the projection matrix $\mathbf{H}$ now reduces to

$$\tilde{\mathbf{H}} = \mathbf{W}^{1/2}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{W}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{W}^{1/2}, \tag{19}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix} \quad \text{and} \quad \mathbf{W} = \begin{bmatrix}\mathbf{W}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_2\end{bmatrix}.$$

The leverage measure for the ith vector (y_{1i}, y_{2i}) is defined by the element $\tilde{h}_i = (\tilde{h}_{1_{ii}} + \tilde{h}_{2_{ii}})$, where $\tilde{h}_{1_{ii}}$ and $\tilde{h}_{2_{ii}}$ are obtained directly from the matrix $\tilde{\mathbf{H}}$. We have $\sum_i \tilde{h}_i = \sum_i(\tilde{h}_{1_{ii}} + \tilde{h}_{2_{ii}}) = p$. Hence, we can consider an interval-valued observation with a high leverage measure when $\tilde{h}_i$ is greater than $2p/n$. Unlike the expression (9), the projection matrix $\tilde{\mathbf{H}}$ in Equation (19) is not a block-diagonal matrix.

The formulae for the Pearson, Studentized Pearson and deviance residuals given in Section 3.3 continue to hold here by replacing h by $\tilde{h}$. The same occurs with the influence measure D_i. However, we now consider influential observations for which the values of D_i are higher than those of $\chi^2_{2p,\alpha}/(2p)$.

5. Numerical study and applications

To show the usefulness of the proposed models, experiments with synthetic interval-valued data sets are now presented. The prediction performance of the BSRMs will be compared with the methods proposed by Billard and Diday [24] and Lima Neto and De Carvalho [25] for these interval data sets. The performance assessment of these approaches will be compared in terms of the *root mean-square error* for each interval boundaries in a K-fold cross-validation scheme. A statistical nonparametric hypothesis test will be applied to measure the statistical difference between the models fitted to interval-valued data sets. Finally, the models will be fitted to real interval-valued data and their performance will be compared.

5.1. *Synthetic interval-valued data sets*

Consider standard quantitative data sets with $n = 300$ points. Each data set is partitioned into 10 disjoint subsets. The learning set is composed by nine disjoint subsets and the test set by the last one. Each data point belonging to the standard data set is a seed for an interval data. Thus, the interval-valued data sets are obtained from these standard data sets.

The construction of the standard data sets and the corresponding interval-valued data sets are carried out in the following steps:

(s_1) Let us suppose that each explanatory variable X_j^m ($j = \{1, 3\}$) is normally distributed with mean 30 and standard deviation 1; at each iteration, n values of each variable X_j^m are randomly selected. At this step, the midpoint of the explanatory interval-valued variables are generated.

(s_2) Let us suppose that each explanatory variable X_j^r ($j = \{1, 3\}$) is normally distributed with mean 10 and standard deviation 0.5; at each iteration, n values of each variable X_j^r are randomly selected. At this step, the ranges of the explanatory interval-valued variables are generated.

(s_3) The linear predictor for the midpoints of intervals is defined in terms of the explanatory variables X_j^m according to $\eta_i^m = \beta_0^m + \sum_{j=1}^{\{1,3\}} \beta_j^m x_{ij}^m$, where $\beta_j^m \sim U[0.9, 1.1]$. In the same way, the linear predictor for the ranges of the intervals is defined in terms of the explanatory variables X_j^r according to $\eta_i^r = \beta_0^r + \sum_{j=1}^{\{1,3\}} \beta_j^r x_{ij}^r$, where $\beta_j^r \sim U[0.9, 1.1]$.

(s_4) Let us suppose that the link functions that connect the mean of the response variables with the linear predictors are identity functions for the midpoint ($\boldsymbol{\mu}^m = \boldsymbol{\eta}^m$) and the range ($\boldsymbol{\mu}^r = \boldsymbol{\eta}^r$).

(s_5) Consider that each component (y_i^m, y_i^r) of the random vector (Y^m, Y^r) belongs to the bivariate normal distribution with mean $\mu = (\eta_i^m, \eta_i^r)^{\mathrm{T}}$ and variance–covariance matrix $\boldsymbol{\Sigma}$ given by $\sigma^m = \sqrt{\sigma^2_{\eta^m}}$, $\sigma^r = \sqrt{\sigma^2_{\eta^r}}$ and $\mathrm{cov}(Y^m, Y^r) = \rho\sigma^m\sigma^r$.

(s_6) The interval-valued data set is partitioned into learning (nine disjoint subsets) and test (one disjoint subset) data sets in a 10-fold cross-validation scheme.

5.2. *Experimental evaluation*

The performance assessment of the BSRMs and the methods proposed by Billard and Diday [24] and Lima Neto and De Carvalho [25] will be based on the following measures, evaluated in the test interval-valued data sets: the lower boundary root mean-square error (RMSE_L) and the upper boundary root mean-square error (RMSE_U). These measures, calculated from the observed values $y_i = [y_{Li}, y_{Ui}]$ of Y and their corresponding predicted values $\hat{y}_i = [\hat{y}_{Li}, \hat{y}_{Ui}]$, are defined by

$$\mathrm{RMSE}_L = \sqrt{\frac{\sum_{i=1}^n (y_{Li} - \hat{y}_{Li})^2}{n}} \quad \text{and} \quad \mathrm{RMSE}_U = \sqrt{\frac{\sum_{i=1}^n (y_{Ui} - \hat{y}_{Ui})^2}{n}}. \tag{20}$$

These measures are determined for each method in a 10-fold cross-validation scheme with 10 replications, for each number of explanatory variables ($p = \{1,3\}$) and values for the correlation coefficient ($\rho = \{0.00, 0.50, 0.75, 0.95\}$). At each fold and replication, we fit a symbolic regression model in the training interval-valued data set by considering the BSRMs and the methods proposed by Billard and Diday [24] and Lima Neto and De Carvalho [25]. Thus, the fitted regression models are used to predict the interval values of the variable Y in the test interval-valued data set and these measures are calculated. For each measure, the Mann–Whitney nonparametric statistical test is applied to compare the prediction performance between the BSRMs and the methods proposed by Billard and Diday [24] and Lima Neto and De Carvalho [25].

For any two compared methods (*A and B*, in this order) concerning the RMSE_L and RMSE_U measures (*the higher the measures, the worse the method*), the null and alternative hypotheses are structured as: H_0 : Method A = Method B versus H_1: Method A $\leq$ Method B. We consider a significance level of 1% for rejection of the null hypothesis H_0.

5.3. *Prediction performance for the BSRM methods*

Table 1 presents the prediction performance measures RMSE_L and RMSE_U, in the test data sets, for the BSRM1 against the methods proposed by Billard and Diday [24] and Lima Neto and De Carvalho [25], denoted by CM and CRM, respectively, in terms of the 10-fold cross-validation scheme with 10 replications. The fitted BSRM1 considers the lower limit (Y^L) and the upper limit (Y^U) of the intervals in the random component. The results show the superiority of the BSRM1 over the CM since the p-values are inside of the rejection area of H_0, at the significance level of 1%, for all configurations considered. On the other hand, the comparison between the BSRM1 and the CRM methods gives evidence that the new approach present the same prediction performance due to the nonrejection of the null hypothesis. However, for the BSRM1, it is possible to use inference techniques over the parameter estimates, apply model diagnostic techniques and make the residual analysis. These results indicate that the new models present advantages in relation to the CM and CRM methods.

Table 2 compares the prediction performance of the BSRM2 against the CM and CRM methods. For the BSRM2, we consider the midpoint (Y^m) and the half-range (Y^r) of the interval-valued data sets in the random component of the model, e.g. $\mathbf{Y} = (Y^m, Y^r)$. The results show the superiority of the BSRM2 over the CM method due to the rejection of the null hypothesis H_0 for different numbers of explanatory variables and different values of the correlation coefficient. In the comparison between the BSRM2 and the CRM methods, the experiments show that both approaches present

Table 1. Comparison of the prediction performance, in the test data sets, between the BSRM1 and the CM and CRM methods in a 10-fold cross-validation scheme.

			CM		CMR	
n	p	ρ	RMSE_L	RMSE_U	RMSE_L	RMSE_U
300	1	0.00	6.10×10^{-10}	2.16×10^{-9}	1.0000	0.9523
		0.50	2.20×10^{-16}	9.93×10^{-16}	0.9756	0.9290
		0.75	4.61×10^{-15}	1.70×10^{-9}	0.9737	0.9659
		0.95	2.20×10^{-16}	3.47×10^{-8}	0.8422	0.9542
	3	0.00	2.07×10^{-5}	1.55×10^{-5}	0.9756	0.9776
		0.50	3.54×10^{-9}	1.26×10^{-5}	0.9328	0.9406
		0.75	1.57×10^{-15}	2.01×10^{-7}	0.9386	0.9659
		0.95	2.20×10^{-16}	2.18×10^{-11}	0.7797	0.8670

P-values for the Mann–Whitney statistical test.

Table 2. Comparison of the prediction performance, in the test data sets, between the BSRM2 and the CM and CRM methods in a 10-fold cross-validation scheme.

			CM		CMR	
n	p	ρ	RMSE_L	RMSE_U	RMSE_L	RMSE_U
300	1	0	7.09×10^{-8}	3.24×10^{-8}	0.9581	0.9990
		0.5	9.79×10^{-13}	2.83×10^{-9}	0.9717	0.9815
		0.75	5.06×10^{-12}	7.00×10^{-7}	0.9523	0.9386
		0.95	1.42×10^{-13}	5.51×10^{-7}	0.9250	0.8940
	3	0	2.20×10^{-16}	2.20×10^{-16}	0.9678	0.8940
		0.5	7.29×10^{-8}	9.57×10^{-7}	0.7815	0.8155
		0.75	1.76×10^{-15}	1.87×10^{-8}	0.8960	0.8960
		0.95	7.23×10^{-13}	1.47×10^{-5}	0.8536	0.9756

P-values for the Mann–Whitney statistical test.

the same prediction performance. However, the BSRM2 allows the use of inference techniques and it presents advantages in relation to the CM and CRM methods.

5.4. *Applications to real interval-valued data sets*

The models developed in Sections 3 and 4 are now applied to two real interval-valued data sets to show their usefulness in practical applications. We calculate the MLEs of the parameters and the goodness-of-fit statistics, residuals and diagnostic measures. The selection of the bivariate distribution for the random component, the link functions and the use of the intervals bounds (lower and upper limits) or the midpoint and half-range in the bivariate random vector $\mathbf{Y} = (Y_1, Y_2)$ is merely exploratory. This type of model selection also occurs in the GLM theory when we choose the distribution and link function and, if necessary, transformations for the response variable or explanatory variables. Thus, the use of exploratory methods can suggest some relevant information about these questions.

5.4.1. *Mushroom interval-valued data set*

The first data set gives the values of the pileus cap width (Y), stipe length (T_1) and stipe thickness (T_2) for 23 mushroom species. Our aim is to predict the interval values of the dependent variable Y in terms of the explanatory variables T_j ($j = 1, 2$). The data set given in Table 3 was obtained from [1]. We assume the random component $\mathbf{Y} = (Y_1, Y_2)$ structured in terms of the midpoint Y^m

Table 3. Mushroom interval-valued data set.

Species	Y	T_1	T_2	Species	Y	T_1	T_2
1	[3.0–8.0]	[4.0–9.0]	[0.50–2.50]	13	[3.5–8.0]	[4.0–10.0]	[1.00–2.00]
2	[6.0–21.0]	[4.0–14.0]	[1.00–3.50]	14	[7.0–14.0]	[8.0–14.0]	[1.50–2.50]
3	[4.0–8.0]	[5.0–11.0]	[1.00–2.00]	15	[8.0–20.0]	[9.0–19.0]	[3.00–5.00]
4	[6.0–7.0]	[4.0–7.0]	[3.00–4.50]	16	[2.5–4.0]	[2.5–4.5]	[0.40–0.70]
5	[5.0–12.0]	[2.0–5.0]	[1.50–2.50]	17	[7.0–19.0]	[8.0–15.0]	[2.00–3.50]
6	[5.0–15.0]	[4.0–10.0]	[2.00–4.00]	18	[5.0–15.0]	[6.0–15.0]	[2.50–3.50]
7	[4.0–11.0]	[3.0–7.0]	[0.40–1.00]	19	[8.0–12.0]	[6.0–12.0]	[1.50–2.00]
8	[5.0–10.0]	[3.0–6.0]	[1.00–2.00]	20	[2.0–6.0]	[3.0–7.0]	[0.40–0.80]
9	[2.5–4.0]	[3.0–5.0]	[0.40–0.70]	21	[6.0–12.0]	[6.0–12.0]	[1.50–2.00]
10	[2.5–6.0]	[1.5–3.5]	[1.00–1.50]	22	[6.0–12.0]	[6.0–16.0]	[1.00–2.00]
11	[1.5–2.5]	[3.0–6.0]	[0.25–0.35]	23	[5.0–17.0]	[4.0–14.0]	[1.00–3.50]
12	[4.0–15.0]	[4.0–15.0]	[1.50–2.50]				

and the half-range Y^r of the interval-valued variable Y, respectively. We also assume the bivariate normal [33] and bivariate gamma [28] distributions for the response variable Y, which are natural members of the bivariate exponential family.

We adopt the *identity* link function for the bivariate normal model and the *log* link function for the bivariate gamma model. The algorithm to obtain the estimates of the linear parameters and the scalars ϕ and ρ and the calculations to yield the goodness-of-fit statistics, residuals and diagnostic measures for both BSRMs was implemented through the software R (http://www.r-project.org). The results of the fitted BSRMs are given in Table 4.

Table 4. Fitted BSRMs to the mushroom interval-valued data set.

Model	Error (link)	Normal (identity)	Gamma (log)
$BSRM1$			
	$\hat{\phi}$	1.4120	0.1605066
	$\hat{\rho}$	0.574	0.090
	D	56.48	6.42
	X^2	59.26	5.76
	$\hat{\boldsymbol{\beta}}$	$\hat{\beta}_m = (0.964, 0.619, 1.335)$	$\hat{\beta}_m = (1.119, 0.074, 0.192)$
		$\hat{\beta}_r = (-0.219, 0.781, 2.199)$	$\hat{\beta}_r = (-0.041, 0.258, 0.616)$
	Iteration	5	18
$BSRM2$			
	$\hat{\phi}$	1.3808	0.1926977
	$\hat{\rho}$	0.611	0.015
	D	55.23	7.71
	X^2	58.13	4.89
	$\hat{\boldsymbol{\beta}}$	$\hat{\beta} = (0.573, 0.659, 1.391)$	$\hat{\beta} = (0.648, 0.118, 0.256)$
	Iteration	4	17

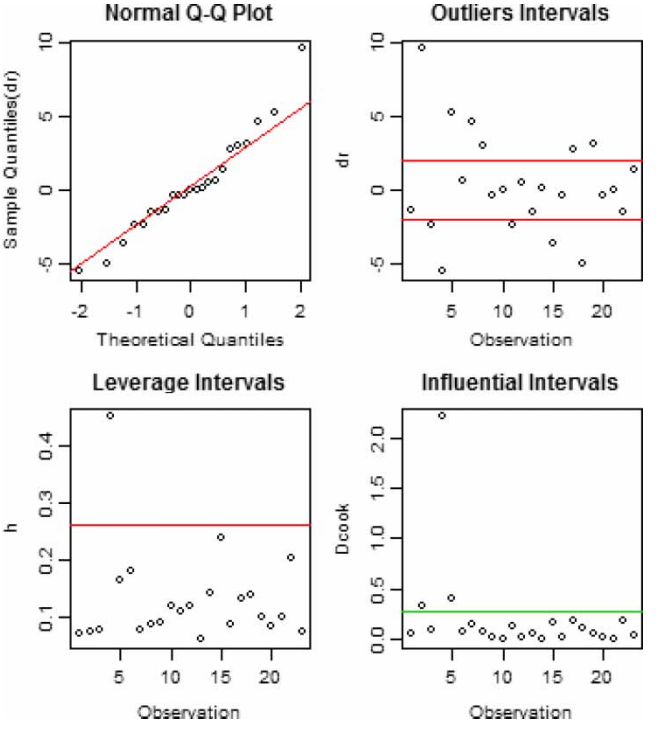

Figure 1. Mushroom data set: residual analysis for the BSRM2 with bivariate normal distribution.

Table 5. Comparison between the symbolic regression methods.

Method	RMSE_L	RMSE_U
BSRM2 (normal)	1.167	2.589
BSRM2 (gamma)	3.774	3.722
BSRM1 (normal)	1.197	2.601
BSRM1 (gamma)	1.179	2.640
CM	1.333	2.721
CRM	1.194	2.501

The BSRM2 with bivariate normal distribution provides the best fit to these data based on the analysis of the residuals, goodness-of-fit statistics and correlation coefficient estimates. Figure 1 shows the residual analysis for this model through a qq-plot for the deviance residual and the identification of outliers, leverage and influential interval-valued observations.

Table 5 compares the BSRM against the CM and CRM methods. Based on the RMSE_L and RMSE_U measures, the BSRM2 with bivariate normal distribution presents the best performance between the bivariate symbolic regression methods, whereas the BSRM2 with bivariate gamma distribution presents the worst performance. The results also indicate that the BSRM2 with bivariate normal distribution outperforms the CM method and presents a very similar performance in comparison with the CRM method. These results support the conclusions obtained in the analysis of the synthetic interval-valued data sets (Section 5.3).

5.4.2. *Soccer interval data set*

This data set gives the records of the weight (Y), height (T_1) and age (T_2) for 531 soccer players grouped into 20 teams of the French Football Professional Championship. We use the BSRMs to predict the dependent variable Y from the explanatory variables T_j ($j = 1, 2$). Table 6 gives the data which can be accessed for free at http://www.ceremade.dauphine.fr/ touati/foot2.htm.

We consider the bivariate random vector in terms of the lower bound Y^L and upper bound Y^U of the interval-valued variable Y. For both BSRMs, we assume the bivariate normal distribution [33] with *identity* link function and the bivariate gamma distribution [28] with *log* link function. The algorithm to obtain the estimates of all parameters and the calculations of the goodness-of-fit statistics, residuals and diagnostic measures was implemented through the software R. Table 7 gives the results for the fitted models.

Figure 2 provides the residual analysis for the BSRM1 with bivariate gamma distribution. In special, the qq-plot shows that the deviance residuals are normally distributed. Further, there are no outliers, leverage and influential interval-valued observations for these data.

Table 6. Soccer interval data set.

Team	Y	T_1	T_2	Team	Y	T_1	T_2
A	[58–85]	[164–192]	[21–35]	K	[62–86]	[164–191]	[18–34]
B	[67–84]	[171–190]	[20–30]	L	[62–80]	[168–189]	[19–35]
C	[65–88]	[170–186]	[18–36]	M	[63–85]	[167–190]	[18–31]
D	[60–83]	[162–188]	[19–31]	N	[65–95]	[168–196]	[20–35]
E	[60–84]	[170–189]	[18–34]	O	[63–83]	[170–187]	[18–35]
F	[67–83]	[173–190]	[18–36]	P	[60–87]	[170–197]	[18–37]
G	[69–90]	[176–193]	[19–34]	Q	[67–85]	[168–190]	[18–32]
H	[65–85]	[170–193]	[19–31]	R	[62–83]	[169–192]	[18–35]
I	[63–84]	[168–188]	[18–34]	S	[63–84]	[172–192]	[18–33]
J	[58–88]	[167–197]	[19–35]	T	[63–85]	[169–194]	[20–34]

Table 7. Fitted BSRMs to the soccer interval-valued data set.

Model	Error (link)	Normal (identity)	Gamma (log)
$BSRM1$			
	$\hat{\phi}$	5.7425	0.0009582
	$\hat{\rho}$	0.376	0.050
	D	195.24	0.0325781
	X^2	200.16	1.193861
	$\hat{\boldsymbol{\beta}}$	$\hat{\beta}_L = (-46.285, 0.632, 0.143)$	$\hat{\beta}_L = (2.406, 0.010, 0.0022)$
		$\hat{\beta}_U = (-20.811, 0.527, 0.147)$	$\hat{\beta}_U = (3.211, 0.0061, 0.0018)$
	Iteration	3	89
$BSRM2$			
	$\hat{\phi}$	5.3724	0.0015035
	$\hat{\rho}$	0.482	0.020
	D	182.66	0.0511194
	X^2	190.18	0.0514307
	$\hat{\boldsymbol{\beta}}$	$\hat{\beta} = (-60.403, 0.683, 0.447)$	$\hat{\beta} = (2.430, 0.0095, 0.0057)$
	Iteration	4	86

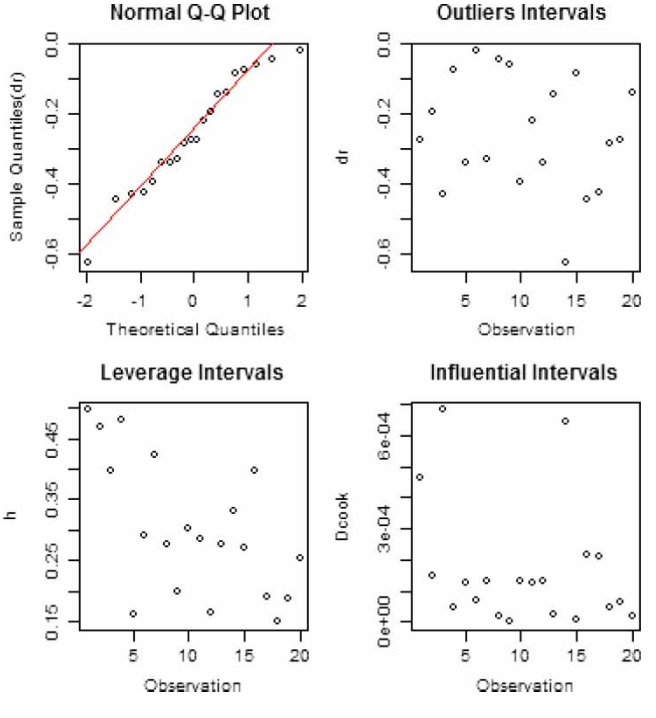

Figure 2. Soccer data set: residual analysis for the BSRM1 with bivariate gamma distribution.

Table 8 compares the BSRM with the CM and CRM methods. The results show that all the BSRM outperform the CM method and that the BSRM methods present a very similar performance to the CRM method. These results support the analysis of the synthetic interval-valued data sets (Section 5.3). The BSRM1 with bivariate gamma distribution (*log* link function) provides the best performance, between the bivariate symbolic regression methods, under the measures RMSE_L and RMSE_U.

Table 8. Comparison between the symbolic regression methods

Method	RMSE_L	RMSE_U
BSRM1 (gamma)	2.232	2.570
BSRM1 (normal)	2.241	2.575
BSRM2 (gamma)	2.245	2.835
BSRM2 (normal)	2.261	2.697
CM	7.540	7.681
CRM	1.946	2.661

6. Concluding remarks

An interval of real values is an infinity list of values and it is difficult to take into account the role information inside it. Despite the loss of information, we represent an interval-valued variable Y as a two-dimensional or a bivariate quantitative feature vector $\mathbf{y}_i = (y_{1i}, y_{2i})$, where the variables Y_1 and Y_2 are one-dimensional random variables. We propose a new class of models so-called the BSRMs which can be useful in statistical analysis of interval-valued data. This new class of models closely follows the framework of the GLMs. We assume that the joint distribution of the response interval-valued variable $\mathbf{Y} = (Y_1, Y_2)$ belongs to the bivariate exponential family of distributions. This family extends the GLM theory for the case of interval-valued variables. The random component of the BSRMs can be represented in terms of the midpoint and half-range of the intervals or the lower and upper boundaries of the intervals or in terms of any other pair of interval features. We show that the BSRMs are an important tool for solving problems related to SDA since they can fit interval-valued data well.

The mean of the response variable is related to a set of known explanatory interval-valued variables through a link function. We define a profile log-likelihood function to estimate the coefficient of correlation ρ between Y_1 and Y_2. We give a see-saw algorithm for maximum-likelihood estimation of the model parameters. In addition, we define goodness-of-fit statistics, deviance measures, residuals and leverage measures for the proposed models. For the first time, as for as we know, we discuss a joint residual measure for interval-valued observations.

We fit the BSRMs to synthetic interval-valued data and to real interval-valued data to compare with the symbolic regression methods proposed by [24,25] denoted by CM and CRM, respectively. The results show that the new models outperform the CM method. In comparison with the CRM method, the BSRM presents a similar performance. However, the new models have advantages in relation to the CM and CRM methods as far as inference techniques are concerned and indicate that they can be very useful in analysing interval-valued data in the SDA field.

Acknowledgements

We thank the anonymous referees for their helpful suggestions and comments. We are also grateful to the Brazilian agencies CNPq, CAPES and FACEPE for their financial support.

References

[1] L. Billard and E. Diday, *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley, New York, 2006.
[2] H.-H. Bock and E. Diday, *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Heidelberg, 2000.
[3] E. Diday and M. Fraiture-Noirhomme, *Symbolic Data Analysis and the SODAS Software*, Wiley-Interscience, Chichester, 2008.
[4] P. Bertrand and F. Goupil, *Descriptive statistics for symbolic data*, in *Analysis of Symbolic Data*, H.-H Bock and E. Diday, eds., Springer, Heidelberg, 2000, pp. 106–124.

[5] L. Billard and E. Diday, *From the statistics of data to the statistics of knowledge: Symbolic data analysis*, J. Am. Statist. Assoc. 98 (2003), pp. 470–487.
[6] F.A.T. De Carvalho, *Histograms in symbolic data analysis*, Ann. Oper. Res. 55 (1995), pp. 229–322.
[7] P. Cazes, A. Chouakria, E. Diday, and S. Schektman, *Extension de l'analyse en composantes principales à des données de type intervalle*, Rev. Statist. Appl. 24 (1997), pp. 5–24.
[8] N.C. Lauro and F. Palumbo, *Principal component analysis of interval data: A symbolic data analysis approach*, Comput. Stat. 15 (2000), pp. 73–87.
[9] F. Palumbo and R. Verde, *Non-symmetrical factorial discriminant analysis for symbolic objects*, Appl. Stoch. Models Bus. Ind. 15 (2000), pp. 419–427.
[10] N.C. Lauro, R. Verde, and F. Palumbo, *Factorial discriminant analysis on symbolic objects*, in *Analysis of Symbolic Data*, H.-H Bock and E. Diday, eds., Springer, Heidelberg, 2000, pp. 212–233.
[11] A.L.S. Maia, F.A.T De Carvalho, and T.B. Ludermir, *Forecasting models for interval-valued time series*, Neurocomputing 71 (2008), pp. 3344–3352.
[12] P. Groenen, S. Winsberg, O. Rodrigues, and E. Diday, *Multidimensional scaling of interval dissimilarities*, Comput. Statist. Data Anal. 51 (2006), pp. 360–378.
[13] M. Ichino, H. Yaguchi, and E. Diday, *A fuzzy symbolic pattern classifier*, in *Ordinal and Symbolic Data Analysis*, E. Diday, Y. Lechevallier and O. Opitz, eds., Springer, Berlin, 1996, pp. 92–102.
[14] M. Chavent, *A monothetic clustering method*, Pattern Recogn. Lett. 19 (1998), pp. 989–996.
[15] F.A.T. De Carvalho, R.M.C.R. Souza, M. Chavent, and Y. Lechevallier, *Adaptive Hausdorff distances and dynamic clustering of symbolic data*, Pattern Recogn. Lett. 27 (2006), pp. 167–179.
[16] R.M.C.R. De Souza, and F.A.T. De Carvalho, *Clustering of interval data based on city-block distances*, Pattern Recogn. Lett. 25 (2004), pp. 353–365.
[17] K.C. Gowda and E. Diday, *Symbolic clustering using a new dissimilarity measure*, Pattern Recogn. 24 (1991), pp. 567–578.
[18] D.S. Guru, B.B. Kiranagi, and P. Nagabhushan, *Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns*, Pattern Recogn. Lett. 25 (2004), pp. 1203–1213.
[19] F.A.T. De Carvalho, *Fuzzy c-means clustering methods for symbolic interval data*, Pattern Recogn. Lett. 28 (2007), pp. 423–437.
[20] N.R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley, New York, 1981.
[21] D.C. Montgomery and E.A. Peck, *Introduction to Linear Regression Analysis*, John Wiley, New York, 1982.
[22] H. Scheffé, *The Analysis of Variance*, John Wiley, New York, 1959.
[23] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed., Chapman & Hall, London, 1989.
[24] L. Billard and E. Diday, *Regression analysis for interval-valued data*, in *Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies, (IFCS-2000) University of Namur, Belgium, 11–14 July 2000*, Springer-Verlag, Heidelberg, 2000, pp. 369–374.
[25] E.A. Lima Neto and F.A.T. De Carvalho, *Centre and range method to fitting a linear regression model on symbolic interval data*, Comput. Stat. Data Anal. 52 (2008), pp. 1500–1515.
[26] F. Alfonso, L. Billard, and E. Diday, *Symbolic linear regression with taxonomies*, in *Classification, Clustering and Data Mining Applications*, D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul, eds., Springer-Verlag, Berlin, 2004, pp. 429–437.
[27] A.L.S. Maia and F.A.T. De Carvalho, *Fitting a least absolute deviation regression model on symbolic interval data*, in *Lecture Notes in Artificial Intelligence: Proceedings of the Ninth Brazilian Symposium on Artificial Intelligence*, Springer-Verlag, Berlin, 2008, pp. 207–216.
[28] M. Iwasaki and H. Tsubaki, *A new bivariate distribution in natural exponential family*, Metrika 61 (2005), pp. 323–336.
[29] M. Iwasaki and H. Tsubaki, *A bivariate generalized linear model with an application to meteorological data analysis*, Statist. Methodol. 2 (2005), pp. 175–190.
[30] P. D'Urso and T. Gastaldi, *A least-square approach to fuzzy linear regression analysis*, Comput. Stat. Data Anal. 34 (2000), pp. 427–440.
[31] R.D. Cook and S. Weisberg, *Residual and Influence in Regression*, Springer, Berlin, 1982.
[32] A.C. Davison and E.J. Snell, *Residuals and diagnostics*, Chapman & Hall, London, 1991.
[33] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed., Prentice Hall, Englewood Cliffs, 2007.