A Clusterwise Center and Range Regression Model for Interval-Valued Data

Francisco de A. T. de Carvalho, Danilo N. Queiroz
(CIn/UFPE-Brazil)

Gilbert Saporta (CNAM-France)

Outline

- Introduction
- Interval-Valued Data
- Clusterwise Regression Model
  - Algorithm
  - Prediction
  - “Goodness of fit” measures
- Applications
  - Car interval-valued data set
- Final Remarks
- References

Introduction

- Clusterwise linear regression is a useful technique when heterogeneity is present in the data.
- It has been proposed as a way to identify both the partition of the data and the relevant regression models, one for each cluster
- Some references: Spaeth (1979), Wayne et al (1988) Hennig (2000), Plaia (2001), Caporossi and Hansen (2007)
- Aim: to adapt clusterwise regression to interval-valued data

Interval-Value Data - I

- Interval-valued data arise in practical situations such as
  - recording monthly interval temperatures in meteorological stations
  - daily interval stock prices
  - or from the aggregation of huge data-bases into a reduced number of groups.
- Interval-valued data has been very much considered in Symbolic Data Analysis
- Book references: Bock and Diday (2000), Billard and Diday (2006), Diday and Noirhome (2008)

Interval-Value Data - II

	Pulse Rate	Systolic pressure	Diastolic pressure
1	[60, 72]	[90,130]	[70,90]
2	[70,112]	[110,142]	[80,108]
3	[54,72]	[90,100]	[50,70]
4	[70,100]	[130,160]	[80,110]
5	[63,75]	[60,100]	[140,150]
6	[44,68]	[90,100]	[50,70]

Each object i is described by a vector of intervals

Interval-Valued Data Analysis Tools are very much required

Clusterwise Regression Model - I

- The present clusterwise regression model is based
  - On the dynamic clustering algorithm (Diday and Simon (1976))
  - Center and range linear regression model (Lima Neto and De Carvalho (2008)
- $E = \{1, \ldots, n\}$: set of observations described by *p+1* interval-valued variables;
- Obervation $i \in E$ is described by a vector of intervals

$\mathbf{e}_i = (w_{i1}, \cdots, w_{ip}, z_i)$ where

$w_{ij} = [w_{ij}^L, w_{ij}^U]$ and $z_i = [z_i^L, z_i^U]$ *(i=1,…,n; j=1,…,p)*

Clusterwise Regression Model - II

- Obervation $i \in E$ is also described by a vector of bi-variate quantitative vectors

$$\mathbf{t}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{ip}, \mathbf{y}_i)$$ where

$$\mathbf{x}_{ij} = \begin{pmatrix} x_{ij}^c \\ x_{ij}^r \end{pmatrix} \qquad x_{ij}^c = \frac{w_{ij}^U + w_{ij}^L}{2} \qquad x_{ij}^r = \frac{w_{ij}^U - w_{ij}^L}{2}$$

$$\mathbf{y}_{ij} = \begin{pmatrix} y_{ij}^c \\ y_{ij}^r \end{pmatrix} \qquad y_i^c = \frac{z_i^U + z_i^L}{2} \qquad y_i^r = \frac{z_i^U - z_i^L}{2}$$

$$(i = 1, \ldots, n;\ j = 1, \ldots, p)$$

Clusterwise Regression Model - III

- Aim:
  - Obtain a partition of E into K clusters $P_1, \ldots, P_K$, each cluster P_k $(k=1,\ldots,K)$ being represented by a prototype (model), by (locally) optimizing an adequacy criterion
- Particularity of the method:
  - The prototype of each cluster is given by a linear regression between dependent and independent interval-valued variables

$$\mathbf{y}_{i(k)} = \boldsymbol{\beta}_{0(k)} + \sum_{j=1}^{p} \boldsymbol{\beta}_{j(k)} \mathbf{x}_{ij} + \boldsymbol{\varepsilon}_{i(k)} \quad (\forall i \in P_k)$$

Clusterwise Regression Model - IV

$$\boldsymbol{\beta}_{0(k)} = \begin{pmatrix} \beta^c_{0(k)} \\ \beta^r_{0(k)} \end{pmatrix} \qquad \boldsymbol{\beta}_{j(k)} = \begin{pmatrix} \beta^c_{j(k)} & 0 \\ 0 & \beta^r_{j(k)} \end{pmatrix}$$

$$\boldsymbol{\varepsilon}_{i(k)} = \begin{pmatrix} \varepsilon^c_{i(k)} \\ \varepsilon^r_{i(k)} \end{pmatrix} = \begin{pmatrix} y^c_i - (\beta^c_{0(k)} + \sum_{j=1}^{p} \beta^c_{j(k)} x^c_{ij}) \\ y^r_i - (\beta^r_{0(k)} + \sum_{j=1}^{p} \beta^r_{j(k)} x^r_{ij}) \end{pmatrix} \quad (\forall i \in P_k)$$

- Adequacy criterion

$$J = \sum_{k=1}^{K} \sum_{i \in P_k} (\boldsymbol{\varepsilon}_{i(k)})^T (\boldsymbol{\varepsilon}_{i(k)}) = \sum_{k=1}^{K} \sum_{i \in P_k} [(\varepsilon^c_{i(k)})^2 + (\varepsilon^r_{i(k)})^2]$$

Algorithm - I

- Initialization
  - Fixe the number *K* $(2 \leq K << n)$ of clusters;
  - Set t=0;
  - Randomly obtain $P^{(0)} = (P_1^{(0)}, \ldots, P_K^{(0)})$
- Step 1: determination of the best prototypes
  - Set t = t + 1;
  - The partition $P^{(t-1)} = (P_1^{(t-1)}, \ldots, P_K^{(t-1)})$ is fixed

Algorithm - II

- The prototype

$$\left(\hat{\mathbf{y}}_{i(k)}\right)^{(t)} = \begin{pmatrix} \left(\hat{y}^{c}_{i(k)}\right)^{(t)} \\ \left(\hat{y}^{r}_{i(k)}\right)^{(t)} \end{pmatrix} = \begin{pmatrix} \left(\hat{\beta}^{c}_{0(k)}\right)^{(t)} + \sum_{j=1}^{p} \left(\hat{\beta}^{c}_{j(k)}\right)^{(t)} x^{c}_{ij} \\ \left(\hat{\beta}^{r}_{0(k)}\right)^{(t)} + \sum_{j=1}^{p} \left(\hat{\beta}^{r}_{j(k)}\right)^{(t)} x^{r}_{ij} \end{pmatrix} \quad (\forall i \in P_k)$$

of cluster P_k *(k=1,…,K)*, which minimizes *J*, has the least squares estimates of the parameters given by the solution of the system

$$\left(\hat{\boldsymbol{\beta}}\right)^{(t)} = \left(\left(\hat{\beta}^{c}_{0(k)}\right)^{(t)}, \ldots, \left(\hat{\beta}^{c}_{p(k)}\right)^{(t)}, \left(\hat{\beta}^{r}_{0(k)}\right)^{(t)}, \ldots, \left(\hat{\beta}^{r}_{p(k)}\right)^{(t)}\right) = \left(\mathbf{A}^{(t)}\right)^{-1} \mathbf{b}^{(t)}$$

where

Algorithm-III

$$\mathbf{A} = \begin{pmatrix} |P_k| & \sum_{i\in P_k} x_{i1}^c & \cdots & \sum_{i\in P_k} x_{ip}^c & 0 & 0 & \cdots & 0 \\ \sum_{i\in P_k} x_{i1}^c & \sum_{i\in P_k} (x_{i1}^c)^2 & \cdots & \sum_{i\in P_k} x_{ip}^c x_{i1}^c & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i\in P_k} x_{ip}^c & \sum_{i\in P_k} x_{i1}^c x_{ip}^c & \cdots & \sum_{i\in P_k} (x_{ip}^c)^2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & |P_k| & \sum_{i\in P_k} x_{i1}^r & \cdots & \sum_{i\in P_k} x_{ip}^r \\ 0 & 0 & \cdots & 0 & \sum_{i\in P_k} x_{i1}^r & \sum_{i\in P_k} (x_{i1}^r)^2 & \cdots & \sum_{i\in P_k} x_{ip}^r x_{i1}^r \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \sum_{i\in P_k} x_{ip}^r & \sum_{i\in P_k} x_{i1}^r x_{ip}^r & \cdots & \sum_{i\in P_k} (x_{ip}^r)^2 \end{pmatrix}$$

$$\mathbf{b} = \left(\sum_{i\in P_k} y_i^c, \sum_{i\in P_k} y_i^c x_{i1}^c, \ldots, \sum_{i\in P_k} y_i^c x_{ip}^c, \sum_{i\in P_k} y_i^r, \sum_{i\in P_k} y_i^r x_{i1}^r, \ldots, \sum_{i\in P_k} y_i^r x_{ip}^r \right)^T$$

Algorithm - IV

- Step 2: definition of the best partition

$$P_k = \left\{ i \in E : (\boldsymbol{\varepsilon}_{i(k)})^T (\boldsymbol{\varepsilon}_{i(k)}) \leq (\boldsymbol{\varepsilon}_{i(h)})^T (\boldsymbol{\varepsilon}_{i(h)}),\, h = 1, \ldots, K \right\}$$

- Stop criterion. Repeat steps 1 and 2 until the criterion J converges

Prediction

- A new observation $\mathbf{e} = (w_1, \cdots, w_p, z)$ is described by the vector of bivariate quantitative vectors $\mathbf{t} = (\mathbf{x}_1, \cdots, \mathbf{x}_p, \mathbf{y})$
- Prediction of the interval $z = [z^L, z^U]$ from the estimated bivariate vectors $\hat{\mathbf{y}}_{(k)} = (\hat{y}^c_{(k)}, \hat{y}^r_{(k)})^T$ (k=1,…,K)

$$\hat{z}_{(k)} = [\hat{z}^L_{(k)}, \hat{z}^U_{(k)}] \text{ with } \hat{z}^L_{(k)} = \hat{y}^c_{(k)} - \hat{y}^r_{(k)} \text{ and } \hat{z}^U_{(k)} = \hat{y}^c_{(k)} + \hat{y}^r_{(k)}$$

where

$$\hat{y}^c_{(k)} = \hat{\beta}^c_{0(k)} + \sum_{\text{j=1}}^{\text{p}} \hat{\beta}^c_{j(k)} x^c_j \text{ and } \hat{y}^r_{(k)} == \hat{\beta}^r_{0(k)} + \sum_{\text{j=1}}^{\text{p}} \hat{\beta}^r_{j(k)} x^r_j$$

"Goodness-of-fit" measures - I

- Determination coefficients

$$R^2_{c(k)} = \frac{\sum_{i \in P_k} (\hat{y}^c_{i(k)} - \overline{y}_{c(k)})^2}{\sum_{i \in P_k} (y^c_i - \overline{y}_{c(k)})^2} \text{ with } \overline{y}_{c(k)} = \frac{\sum_{i \in P_k} y^c_i}{n_k}$$

$$R^2_{r(k)} = \frac{\sum_{i \in P_k} (\hat{y}^r_{i(k)} - \overline{y}_{r(k)})^2}{\sum_{i \in P_k} (y^r_i - \overline{y}_{r(k)})^2} \text{ with } \overline{y}_{r(k)} = \frac{\sum_{i \in P_k} y^r_i}{n_k}$$

- Lower and upper boundaries root-mean-square error

$$RMSE_L = \sqrt{\frac{\sum_{i=1}^{n} (z^L_i - \hat{z}^L_i)^2}{n}} \qquad RMSE_U = \sqrt{\frac{\sum_{i=1}^{n} (z^U_i - \hat{z}^U_i)^2}{n}}$$

Application: car interval-valued data set - I

- 33 car models described by 2 interval-valued variables: price *y* and engine capacity *x*
- http://www.info.fundp.ac.be/asso/index.html

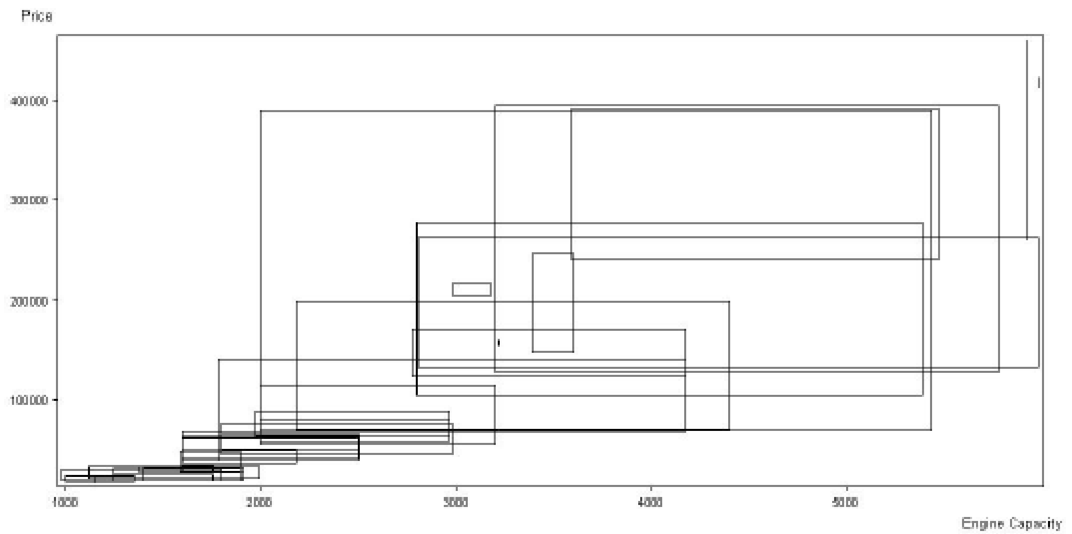

Fig. 1. The car interval-valued data set.

Application: car interval-valued data set - II

- Aim: predict *Price (y)* from *Engine Capacity (x)* through linear regression models
- Both variables – *Price* and *Engine Capacity* –, have been considered for clustering purposes
- The algorithm has been performed on this data set in order to obtain a partition into $K = \{1, 2, 3\}$ clusters
- For a fixed *K*, the algorithm is run 100 times and the best result according to the adequacy criterion is selected.

Application: car interval-valued data set - III

- Regression equations

$K-partition$	cluster k	"Center Model"	"Range Model"
1	1	$\hat{y}^c_{(1)} = -98840.9 + 79.2\,x^c_1$	$\hat{y}^r_{(1)} = -341.4 + 60.9\,x^r_1$
2	1	$\hat{y}^c_{(1)} = -63462.2 + 59.6\,x^c_1$	$\hat{y}^r_{(1)} = -4560.1 + 47.1\,x^r_1$
	2	$\hat{y}^c_{(2)} = -22836.5 + 68.8\,x^c_1$	$\hat{y}^r_{(2)} = 34563.6 + 68.6\,x^r_1$
3	1	$\hat{y}^c_{(1)} = -77422.1 + 82.0\,x^c_1$	$\hat{y}^r_{(1)} = 2229.7 + 92.2\,x^r_1$
	2	$\hat{y}^c_{(2)} = -58484.1 + 71.1\,x^c_1$	$\hat{y}^r_{(2)} = 101952.9 - 546.7\,x^r_1$
	3	$\hat{y}^c_{(3)} = -73362.1 + 62.0\,x^c_1$	$\hat{y}^r_{(3)} = -9755.9 + 53.2\,x^r_1$

- Determination coefficients

K-partition	1	2		3		
cluster k	1	1	2	1	2	3
$R^2_{c(k)}$	0.93	0.95	0.91	0.97	0.99	0.98
$R^2_{r(k)}$	0.53	0.79	0.66	0.98	0.98	0.83

Application: car interval-valued data set - IV

- Predictions: the estimates of the *K* regression models are combined according to the “stacked regressions” approach (Breiman (1996))
- Stacked regressions: uses cross validation data and least squares under non-negativity constraints for forming linear combinations of different predictors
- These predictions are combined to obtain the predictions for the observations belonging to the test data set
- $RMSE_L$ and $RMSE_U$ are computed from the predicted values on the test data sets

Application: car interval-valued data set - V

- This process is repeated 100 times and it is calculated the average and standard deviation of the $RMSE_L$ and $RMSE_U$ measures

K-partition	1	2	3
$RMSE_L$	96649.28 (13812.49)	90417.42 (13538.22)	94993.75 (11376.24)
$RMSE_U$	143416.6 (17294.02)	135471.4 (17027.49)	137825.9 (14243.29)

- 2 regression models given by the 2-cluster partition give the best preditive model through the "stacked regressions" approach

Concluding Remarks

- It was introduced a clusterwise regression model for interval-valued data.
- It combines the dynamic clustering algorithm with the center and range regression model for interval-valued
- Aim: to identify both the partition of the data and the relevant regression models (one for each cluster).
- Experiments with a car interval-valued data set showed the interest of this approach

Bibliography

[1] BILLARD, L. and DIDAY, E. (2007): Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley-Interscience, San Francisco.

[2] BREIMAN, L. (1996): Stacked Regressions. Machine Learning 24, 49-64.

[3] DIDAY, E. and SIMON, J.C. (1976): Clustering analysis. In: K.S. Fu (Eds.): Digital Pattern Classication. Springer, Berlin, 47–94.

[4] HENNIG, C. (2000): Identifiability of models for clusterwise linear regression. J. Classication 17 (2), 273-296.

[5] LIMA NETO, E. A. and DE CARVALHO, F.A.T. (2008): Centre and Range method for fitting a linear regression model to symbolic interval data. Computational Statistics and Data Analysis, 52 (3): 1500-1515.

[6] SPAETH, H. (1979): Clusterwise Linear Regression. Computing 22 (4), 367-373.

[7] WAYNE, S., DESARBO, W.S. and CRON, W.L. (1988): A maximum likelihood methodology for clusterwise linear regression. J. Classication 5 (2), 249-282.

Thank you