# Latest developments of the RSDA 2.0: An R package for Symbolic Data Analysis

## SDA-2017
## Ljubljana, Slovenia

*Oldemar Rodríguez*

*University of Costa Rica*

- To install the package from CRAN
  - ✓ install.packages("RSDA",dependencies=TRUE)

- To load the package
  - ✓ library(RSDA)

# To get help about the package

- ✓ ?RSDA

RSDA {RSDA} R Documentation

## R to Symbolic Data Analysis

### Description

This work is framed inside the Symbolic Data Analysis (SDA). The objective of this work is to implement in R to the symbolic case certain techniques of the automatic classification, as well as some lineal models. These implementations will always be made following two fundamental principles in Symbolic Data Analysis like they are: Classic Data Analysis should always be a case particular case of the Symbolic Data Analysis and both, the exit as the input in an Symbolic Data Analysis should be symbolic. We implement for variables of type interval the mean, the median, the mean of the extreme values, the standard deviation, the deviation quartil, the dispersion boxes and the correlation also three new methods are also presented to carry out the lineal regression for variables of type interval. We also implement in this R package the method of Principal Components Analysis in two senses: First, we propose three ways to project the interval variables in the circle of correlations in such way that is reflected the variation or the inexactness of the variables. Second, we propose an algorithm to make the Principal Components Analysis for variables of type histogram. We implement a method for multidimensional scaling of interval data, denominated INTERSCAL.

### Details

| | |
|---|---|
| Package: | RSDA |
| Type: | Package |
| Version: | 2.0 |
| Date: | 2015-10-01 |
| License: | GPL (>=2) |

Most of the function of the package stars from a symbolic data table that can be store in a CSV file withe follwing forma: In the first row the labels $C means that follows a continuous variable, $I means an interval variable, $H means a histogram variables and $S means set variable. In the first row each labels should be follow of a name to variable and to the case of histogram a set variables types the names of the modalities (categories) . In data rows for continuous variables we have just one value, for interval variables we have the minimum and the maximum of the interval, for histogram variables we have the number of modalities and then the probability of each modality and for set variables we have the cardinality of the set and next the elements of the set.

### Author(s)

Oldemar Rodriguez Rojas
Maintainer: Oldemar Rodriguez Rojas <oldemar.rodriguez@ucr.ac.cr>

### References

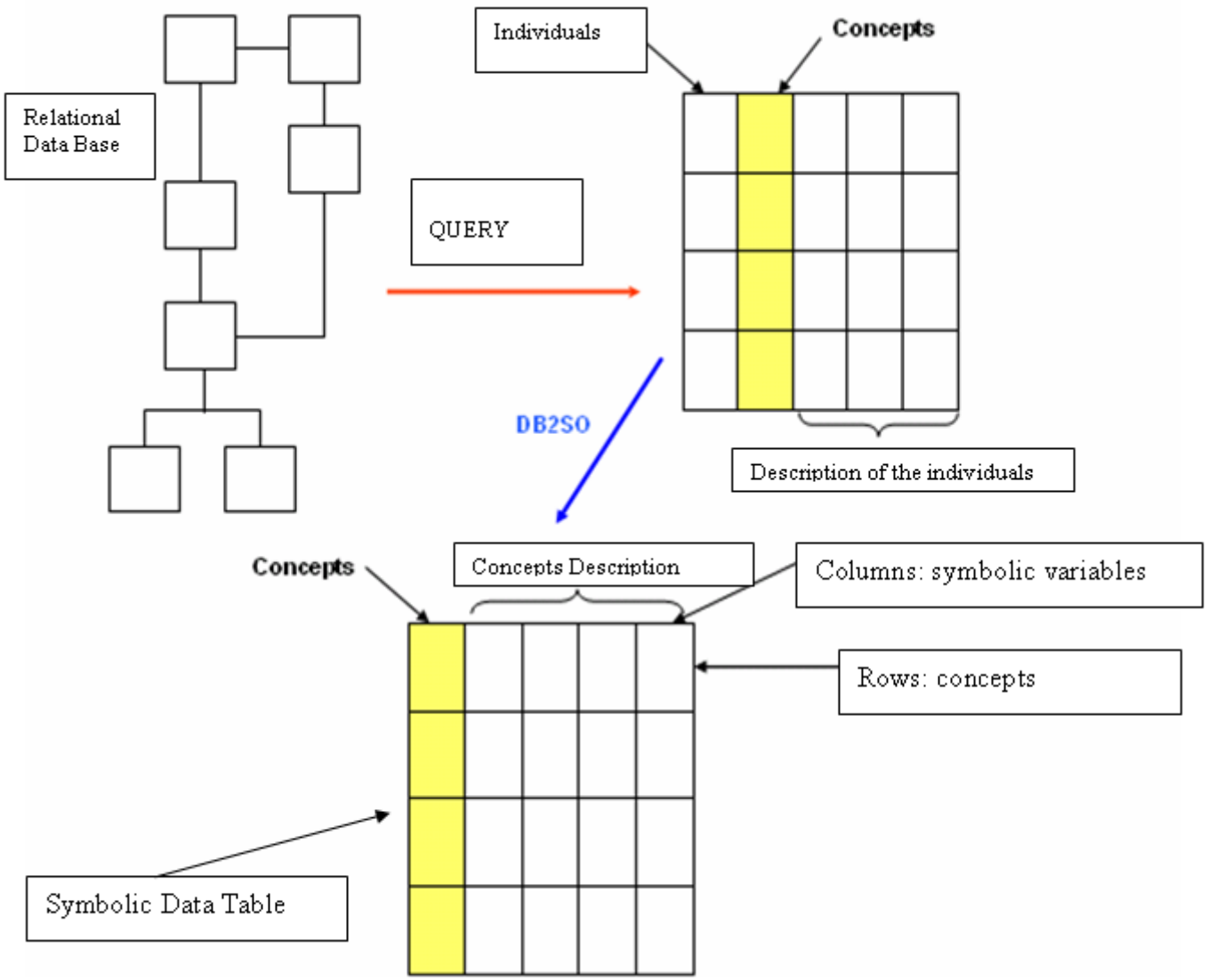
Billard L. and Diday E. (2006). Symbolic data analysis: Conceptual statistics and data mining. Wiley, Chichester.

Billard L., Douzal-Chouakria A. and Diday E. (2011) Symbolic Principal Components For Interval-Valued Observations, Statistical Analysis and Data Mining. 4 (2), 229-246. Wiley.

# What is new in RSDA 2.0?

- Symbolic Data Frame
- New distances between symbolic objects
- Radar (star) plots for interval variables
- Histogram distribution for interval variables
- New PCA methods (optimized, tops, principals curves)
- Now RSDA is in GitHub to collaborate to RSDA

# How to generate a symbolic data table from a classic data table?



# How to generate a symbolic data table from a classic data table?

Classic Data Table

Cases

Millions …

| *Id-trx* | *Trans Type* | *Commerce Loc* | *Amount* | *Card Num* |
|---|---|---|---|---|
| *3457* | *36* | *Curridabat* | *2,500.00* | *1000* |
| *1251* | *28* | *San Pedro* | *1,750.00* | *1001* |
| *3245* | *39* | *Grecia* | *2,400.00* | *1000* |
| *7635* | *35* | *San Pedro* | *1,900.00* | *1001* |
| *3245* | *35* | *Alajuela* | *1,850.00* | *1001* |
| *5367* | *27* | *Alajuela* | *1,900.00* | *1002* |
| *6486* | *34* | *Heredia* | *1,600.00* | *1002* |

Symbolic Data Table

Concepts

Hundreds

| *Card Num* | *Trans Type* | *Commerce Loc* | Amount | Salary |
|---|---|---|---|---|
| *1000* | 36(1/2),39(1/2) | *{Curr-50%,Gre-50%}* | [2.4,2.5] | 255.4 |
| *1001* | 28(1/3),35(2/3) | *{SP-66%,Al-33%}* | [1.75,1.9] | 122,2 |
| *1002* | 27(1/2),34(1/2) | *{Al-50%,Her-50%}* | [1.6,1.9] | 534,5 |

# How to generate a symbolic data table from a classic data table?

Classical description of Schools

| Schools | Town | Nb of pupils | Kind | Level |
|---|---|---|---|---|
| Jaurès | Paris | 320 | Public | 1 |
| Condorcet | Paris | 450 | Public | 3 |
| Chevreul | Lyon | 200 | Public | 2 |
| St Hélène | Lyon | 380 | Private | 3 |
| St Sernin | Toulouse | 290 | Public | 1 |
| St Hilaire | Toulouse | 210 | Private | 2 |

Symbolic description of the towns by the schools variables

| Town | Nb of pupils | Kind | Level |
|---|---|---|---|
| Paris | [320, 450] | (100%)Public | {1, 3} |
| Lyon | [200, 380] | (50%)Public , (50%)Private | {2, 3} |
| Toulouse | [210, 290] | (50%)Public , (50%)Private | {1, 2} |

# How to generate a symbolic data table from a classic data table?

Table 2.3 Credit card dataset.

| $i$ | Name | Month | Food | Social | Travel | Gas | Clothes |
|---|---|---|---|---|---|---|---|
| 1 | Jon | February | 23.65 | 14.56 | 218.02 | 16.79 | 45.61 |
| 2 | Leigh | May | 28.47 | 8.99 | 141.60 | 21.74 | 86.04 |
| 3 | Leigh | July | 30.86 | 9.55 | 193.14 | 24.26 | 95.68 |
| 4 | Tom | July | 24.13 | 15.97 | 190.40 | 35.71 | 20.02 |
| 5 | Jon | April | 23.40 | 11.61 | 179.38 | 23.73 | 48.89 |
| 6 | Jon | November | 23.11 | 16.71 | 178.78 | 20.55 | 47.96 |
| 7 | Leigh | September | 32.14 | 12.34 | 165.65 | 17.62 | 66.40 |
| 8 | Leigh | August | 25.92 | 20.78 | 201.18 | 32.97 | 70.96 |
| 9 | Leigh | November | 31.52 | 16.62 | 177.50 | 20.95 | 71.18 |
| 10 | Jon | November | 23.11 | 14.41 | 179.86 | 20.53 | 51.49 |

Table 2.4 Credit card use by person-months.

| Name – Month | Food | Social | Travel | Gas | Clothes |
|---|---|---|---|---|---|
| Jon – January | [20.81, 29.38] | [9.74, 18.86] | [192.33, 205.23] | [13.01, 24.42] | [44.28, 53.82] |
| Jon – February | [21.44, 27.58] | [10.86, 18.01] | [214.98, 229.63] | [16.08, 22.86] | [50.51, 63.57] |
| ⋮ ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Tom – January | [23.28, 30.00] | [8.67, 18.31] | [193.53, 206.53] | [26.28, 35.61] | [15.51, 25.66] |
| Tom – February | [20.61, 28.66] | [10.66, 17.20] | [195.53, 203.83] | [25.43, 34.18] | [12.99, 24.88] |
| ⋮ ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Leigh – January | [25.59, 35.33] | [7.07, 19.00] | [194.12, 207.05] | [17.75, 23.07] | [61.47, 75.43] |
| Leigh – February | [31.30, 40.80] | [9.05, 24.44] | [212.76, 227.43] | [13.81, 25.08] | [71.63, 85.58] |
| ⋮ ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

# Center and Range Method (CRM)

- Then the *center* least square estimate of β is computed by:

$$\hat{\boldsymbol{\beta}}^{c} = ((\mathbf{X}^{c})^{T}\mathbf{X}^{c})^{-1}(\mathbf{X}^{c})^{T}\mathbf{y}^{c}$$

- Then the *range* least square estimate of β is computed by:

$$\hat{\boldsymbol{\beta}}^{r} = ((\mathbf{X}^{r})^{T}\mathbf{X}^{r})^{-1}(\mathbf{X}^{r})^{T}\mathbf{y}^{r}$$

- The response value is fitted as follows: $\hat{y} = [\hat{y_L}, \hat{y_U}]$

$$\boxed{\hat{y}_L = \hat{y}^c - \hat{y}^r} \text{ and } \boxed{\hat{y}_U = \hat{y}^c + \hat{y}^r,}$$

$$\text{where, } \hat{y}^c = (\tilde{\mathbf{x}}^c)^T\hat{\boldsymbol{\beta}}^{\mathbf{c}}, \hat{y}^r = (\tilde{\mathbf{x}}^r)^T\hat{\boldsymbol{\beta}}^{\mathbf{r}}$$

# Ridge Center and Range Method

- Then the *center* least square estimate of β is computed by:

$$\hat{\beta}^{c^{\mathbf{ridge}}} = \left((\mathbf{X}^c)^T(\mathbf{X}^c) + \lambda\mathbf{I}\right)^{-1}(\mathbf{X}^c)^T\mathbf{y}^c$$

- Then the *range* least square estimate of β is computed by:

$$\hat{\beta}^{r^{\mathbf{ridge}}} = \left((\mathbf{X}^r)^T(\mathbf{X}^r) + \lambda\mathbf{I}\right)^{-1}(\mathbf{X}^r)^T\mathbf{y}^r$$

- The response value is fitted as follows: $\hat{y} = [\hat{y}_L, \hat{y}_U]$

$$\boxed{\hat{y}_L = \hat{y}^c - \hat{y}^r} \quad \text{and} \quad \boxed{\hat{y}_U = \hat{y}^c + \hat{y}^r,}$$

$$\text{where } \hat{y}^c = (\tilde{x}^c)^T\hat{\beta}^{c^{\mathbf{ridge}}} \quad \text{and} \quad \hat{y}^r = (\tilde{x}^c)\hat{\beta}^{r^{\mathbf{ridge}}}$$

# Lasso Center and Range Method

- Then the *center* least square estimate of β is obtained solving the following optimization problem:

$$\hat{\beta}^{c^{\text{lasso}}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i^c - \beta_0 - \sum_{j=1}^{p} x_{ij}^c \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Then the *range* least square estimate of β is obtained solving the following optimization problem:

$$\hat{\beta}^{r^{\text{lasso}}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left( (y_i^r - \beta_0 - \sum_{j=1}^{p} x_{ij}^r \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

$\hat{\beta}_j^{c^{\text{lasso}}} \neq 0$ $\qquad$ $\hat{\beta}_j^{c^{\text{lasso}}} \neq 0$

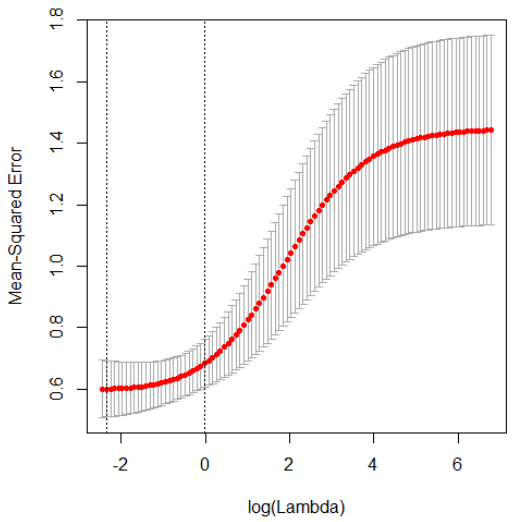
# Lasso Center and Range Method

- Then the response value is fitted as follows: $\hat{y} = [\hat{y_L}, \hat{y_U}]$

$$\boxed{\hat{y}_L = \hat{y}^c - \hat{y}^r} \quad \text{and} \quad \boxed{\hat{y}_U = \hat{y}^c + \hat{y}^r,}$$

$$\text{where } \hat{y}^c = (\tilde{x}^c)^T \hat{\beta}^{c\,\text{lasso}} \quad \text{and} \quad \hat{y}^r = (\tilde{x}^c)\hat{\beta}^{r\,\text{lasso}}$$

- The optimal value of λ is computed using ***Cross-Validation***:



# Elastic Net Center Method (ECM)

- The center least square estimate of $\beta$ is obtained solving the following optimization problem:

$$\hat{\beta}^{\text{NET}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i^c - \beta_0 - \sum_{j=1}^{p} x_{ij}^c \beta_j)^2 + \lambda \left( \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j| \right) \right\}$$

- Then the response value is fitted as follows: $\hat{y} = [\hat{y}_L, \hat{y}_U]$

$$\hat{y}_L = (\mathbf{x}_L)^T \hat{\beta}^{\mathbf{NET}} \quad \text{and} \quad \hat{y}_U = (\mathbf{x}_U)^T \hat{\beta}^{\mathbf{NET}}$$

where $(\mathbf{x}_L)^T = (1, a_1, \ldots, a_p)$, $(\mathbf{x}_U)^T = (1, b_1, \ldots, b_p)$.

# Elastic Net Center and Range Method

- Then the *center* least square estimate of $\beta$ is obtained solving the following optimization problem:

$$\hat{\beta}^{c^{\text{NET}}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i^c - \beta_0 - \sum_{j=1}^{p} x_{ij}^c \beta_j)^2 + \lambda \left( \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j| \right) \right\}$$

- Then the *range* least square estimate of $\beta$ is obtained solving the following optimization problem:

$$\hat{\beta}^{r^{\text{NET}}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i^r - \beta_0 - \sum_{j=1}^{p} x_{ij}^r \beta_j)^2 + \lambda \left( \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j| \right) \right\}$$

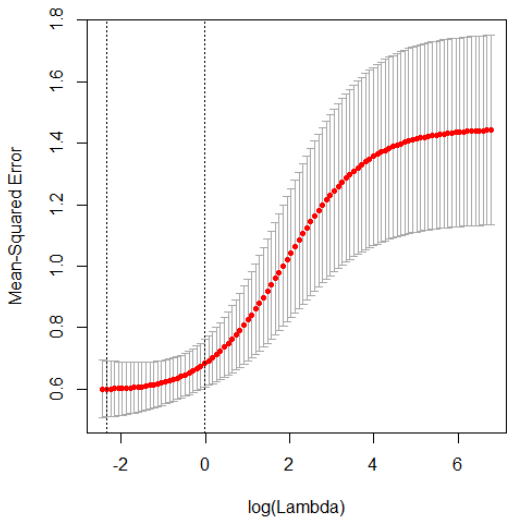
# Elastic Net Center and Range Method

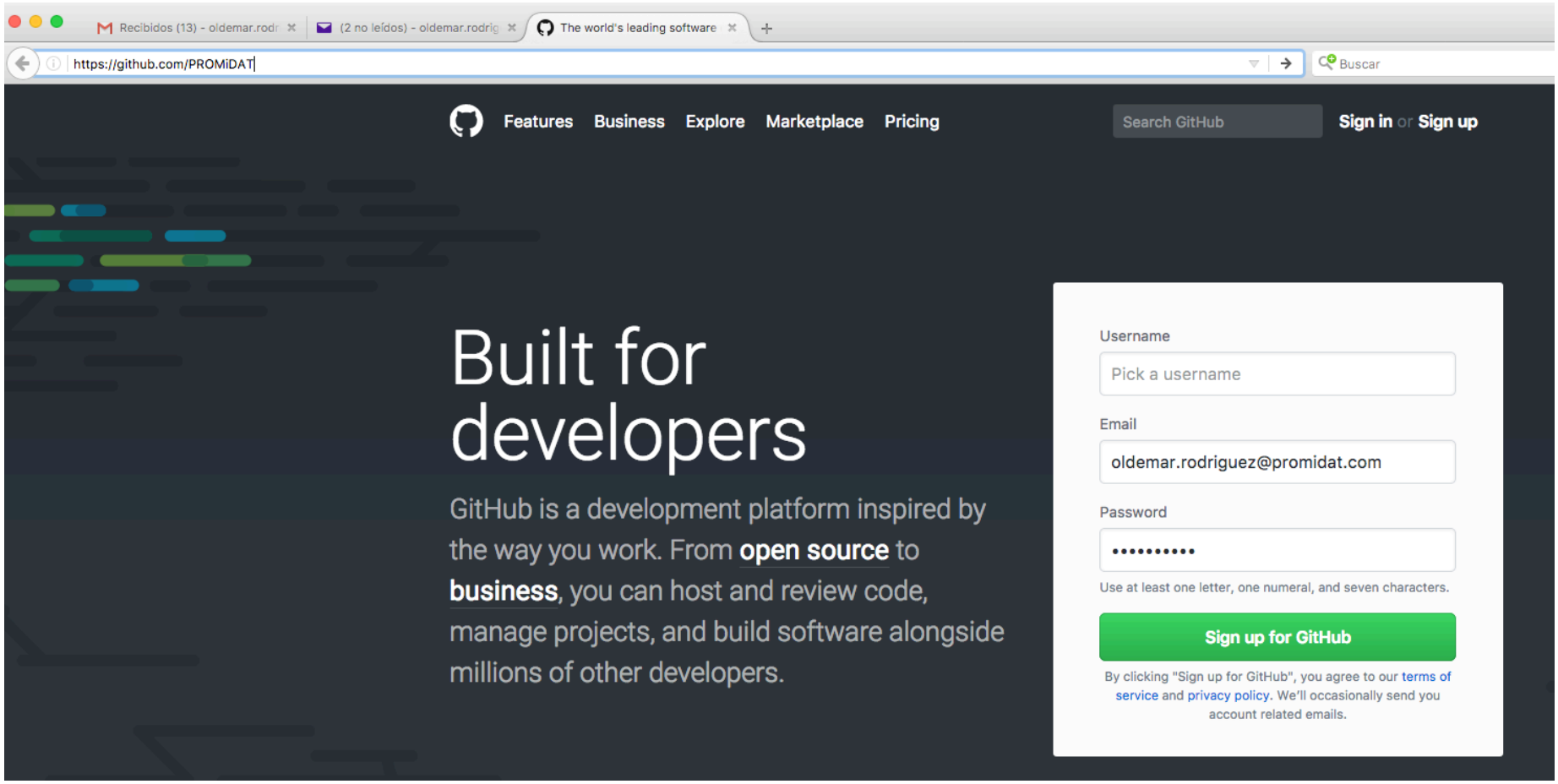
- Then the response value is fitted as follows: $\hat{y} = [\hat{y_L}, \hat{y_U}]$

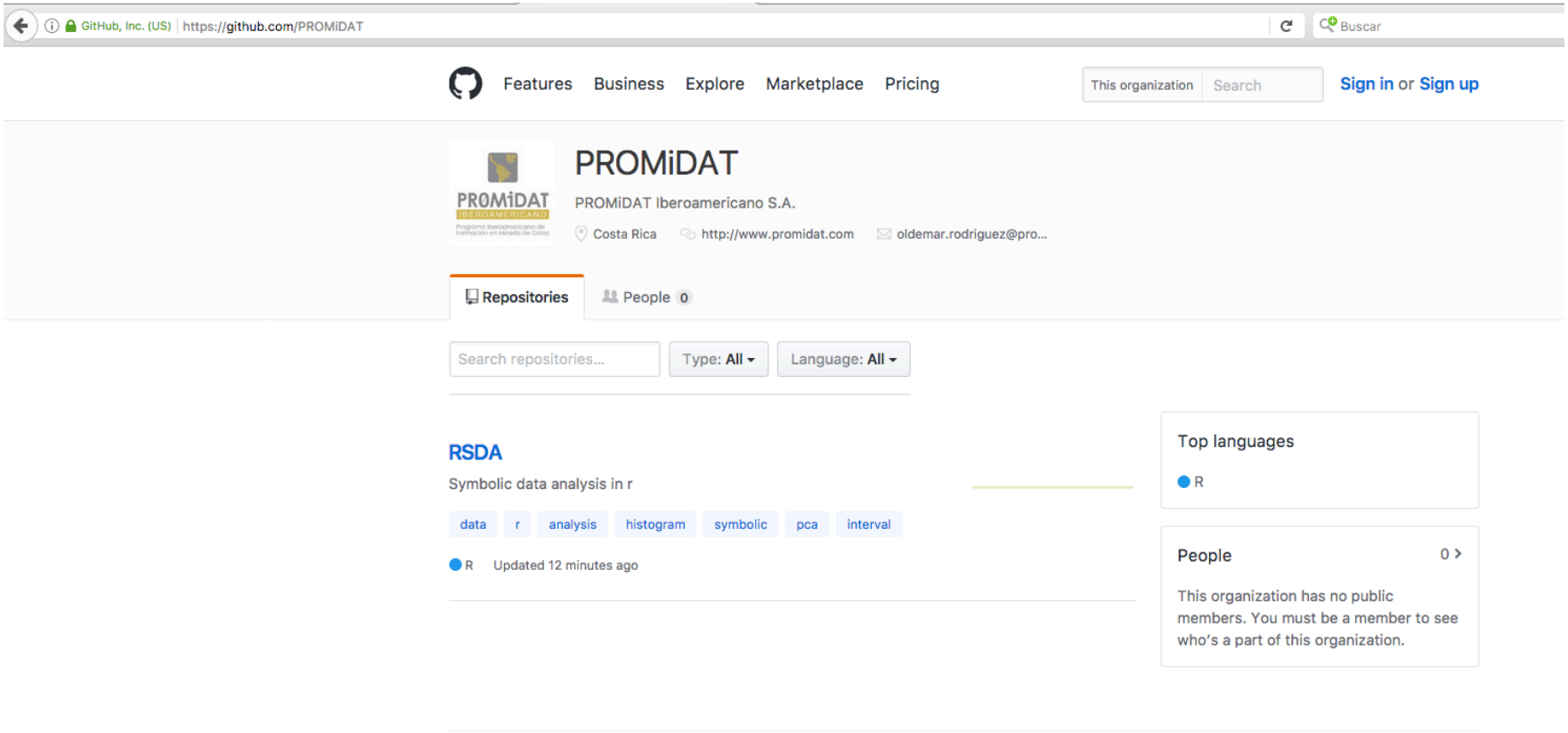
$$\boxed{\hat{y}_L = \hat{y}^c - \hat{y}^r} \quad \text{and} \quad \boxed{\hat{y}_U = \hat{y}^c + \hat{y}^r,}$$

$$\text{where } \hat{y}^c = (\tilde{x}^c)^T \hat{\beta}^{c^{\mathbf{NET}}} \quad \text{and} \quad \hat{y}^r = (\tilde{x}^c) \hat{\beta}^{r^{\mathbf{NET}}}$$

- The optimal value of λ is computed using ***Cross-Validation***:




Recibidos (13) - oldemar.rodr
(2 no leídos) - oldemar.rodrig
The world's leading software
https://github.com/PROMiDAT
Buscar
Features
Business
Explore
Marketplace
Pricing
Search GitHub
Sign in or Sign up
Built for developers
GitHub is a development platform inspired by the way you work. From open source to business, you can host and review code, manage projects, and build software alongside millions of other developers.
Username
Pick a username
Email
oldemar.rodriguez@promidat.com
Password
Use at least one letter, one numeral, and seven characters.
Sign up for GitHub
By clicking "Sign up for GitHub", you agree to our terms of service and privacy policy. We'll occasionally send you account related emails.


GitHub, Inc. (US) https://github.com/PROMiDAT
Buscar
Features Business Explore Marketplace Pricing
This organization Search
Sign in or Sign up
PROMiDAT
PROMiDAT Iberoamericano S.A.
Costa Rica
http://www.promidat.com
oldemar.rodriguez@pro...
Repositories
People 0
Search repositories...
Type: All
Language: All
RSDA
Symbolic data analysis in r
data
r
analysis
histogram
symbolic
pca
interval
R
Updated 12 minutes ago
Top languages
R
People
0 >
This organization has no public members. You must be a member to see who's a part of this organization.

README.md

# RSDA

Symbolic Data Analysis (SDA) was proposed by professor Edwin Diday in 1987, the main purpose of SDA is to substitute the set of rows (cases) in the data table for a concept (second order statistical unit). This package implements, to the symbolic case, certain techniques of automatic classification, as well as some linear models.

## Installation

You can install RSDA from github with:

```
# install.packages("devtools")
devtools::install_github("PROMiDAT/RSDA")
```

## Examples

```
data(ex1_db2so)
ex1_db2so
```

```
        state sex county group age
1     Florida   M      2     6   3
2  California   F      4     3   4
3       Texas   M     12     3   4
4     Florida   F      2     3   4
5       Texas   M      4     6   4
6       Texas   F      2     3   3
7     Florida   M      6     3   4
```

You can get this presentation at:

www.oldemarrodriguez.com

oldemar.rodriguez@ucr.ac.cr

Thank you…….