VarSelLCM

Variable Selection for Model-Based Clustering of Continuous, Count, Categorical or Mixed-Type Data Set with Missing Values.

Description:

- *Authors*: **Matthieu Marbac** and **Mohammed Sedki**.
- *License*: GPL-2.
- *Download VarSelLCM 2.0.1 (beta version for mixed-type data)*: link.

References:

- Marbac, M. and Sedki, M. (2017), Variable selection for model-based clustering using the integrated complete-data likelihood, Statistics and Computing, Volume 27, Issue 4, pp 1049–1063.
- Marbac, M. and Patin, E. and Sedki, M. (2018). Variable selection for mixed data clustering: Application in human population genomics. Arxiv 1703.02293.

Introduction:
VarSelLCM permits model selection (number of component and variable selection) for clustering.

It can analyze continuous, categorical, integer and mixed-data. Moreover, data with missing values can be analyze without any pre-processing.

A Shiny application is implemented to easily interpret the clustering results.

Imputation function is implemented. Thus, *VarSelLCM* can be used as an imputation approach based on mixture model.

Tool functions (*summary, print* and *plot*) facilitate the result interpretation.

Overview of the VarSelLCM functions

This section performs the whole analysis of the *Heart* data set . It uses all the functions implemented in the package *VarSelLCM* and can be used as a tutorial.

Clustering is performed with two clusters by doing a variable selection. Model selection is done with BIC.

Loadings

```
library(VarSelLCM)
```

```
# Data loading
data("heart")
# Add a missing value (just to show that it works!)
heart[1,1] <- NA
# Clustering with variable selection and a number of cluster betwee 1 and 4
# Model selection is BIC (to use MICL, the option must be specified)
out <- VarSelCluster(heart[,-13], 2)
# Summary of the best model
summary(out)
```

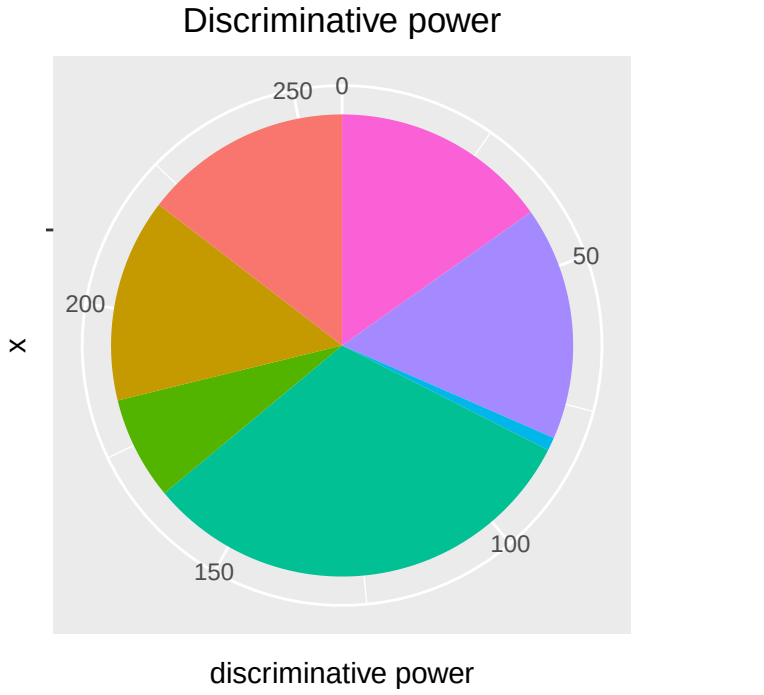
```
Data set:
   Number of individuals: 270
   Number of continuous variables: 3
   Number of count variables: 1
   Percentile of missing values for the integer variables: 0.37
   Number of categorical variables: 8

Model:
```

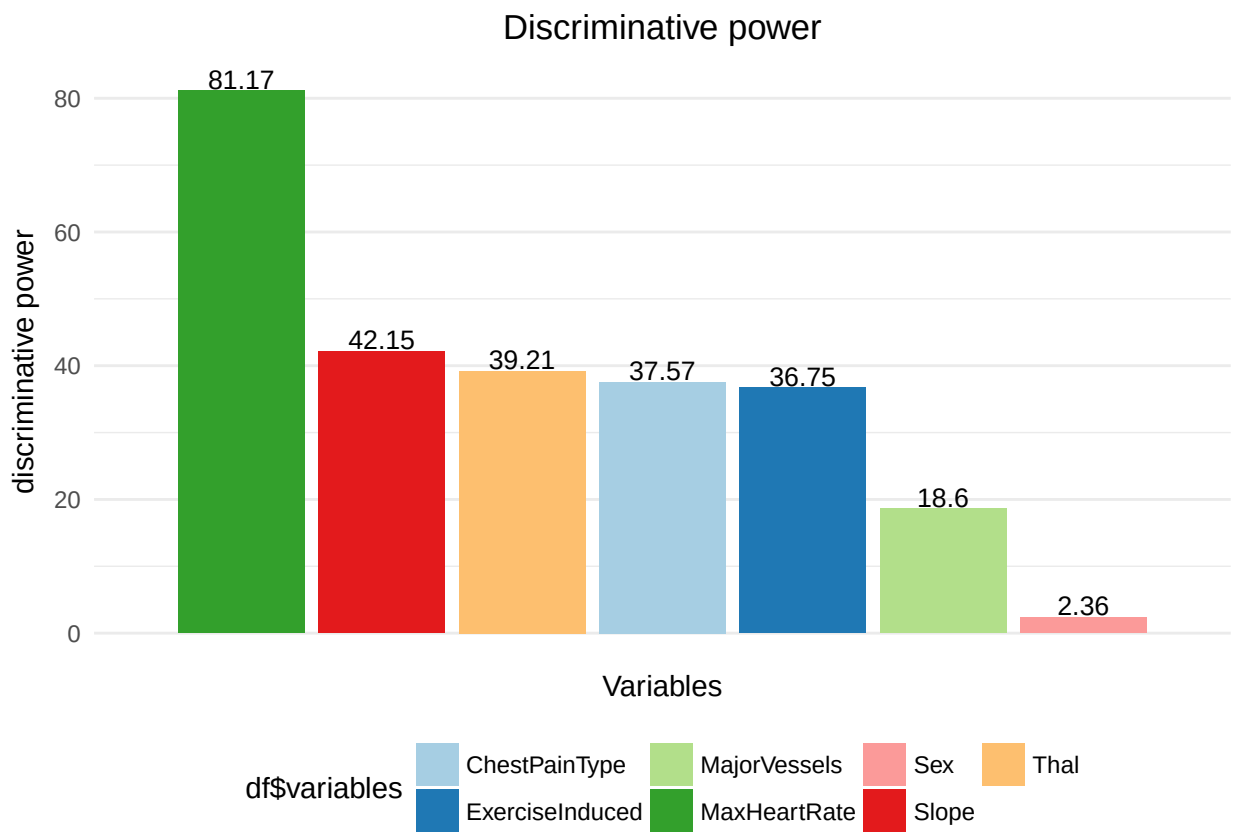
```
   Number of components: 2
   Model selection has been performed according to the BIC  criterion
   Variable selection has been performed, 8  ( 66.67 % ) of the variables are relevant for clustering

Information Criteria:
   loglike: -6403.136
   AIC:     -6441.136
   BIC:     -6509.506
   ICL:     -6638.116
```

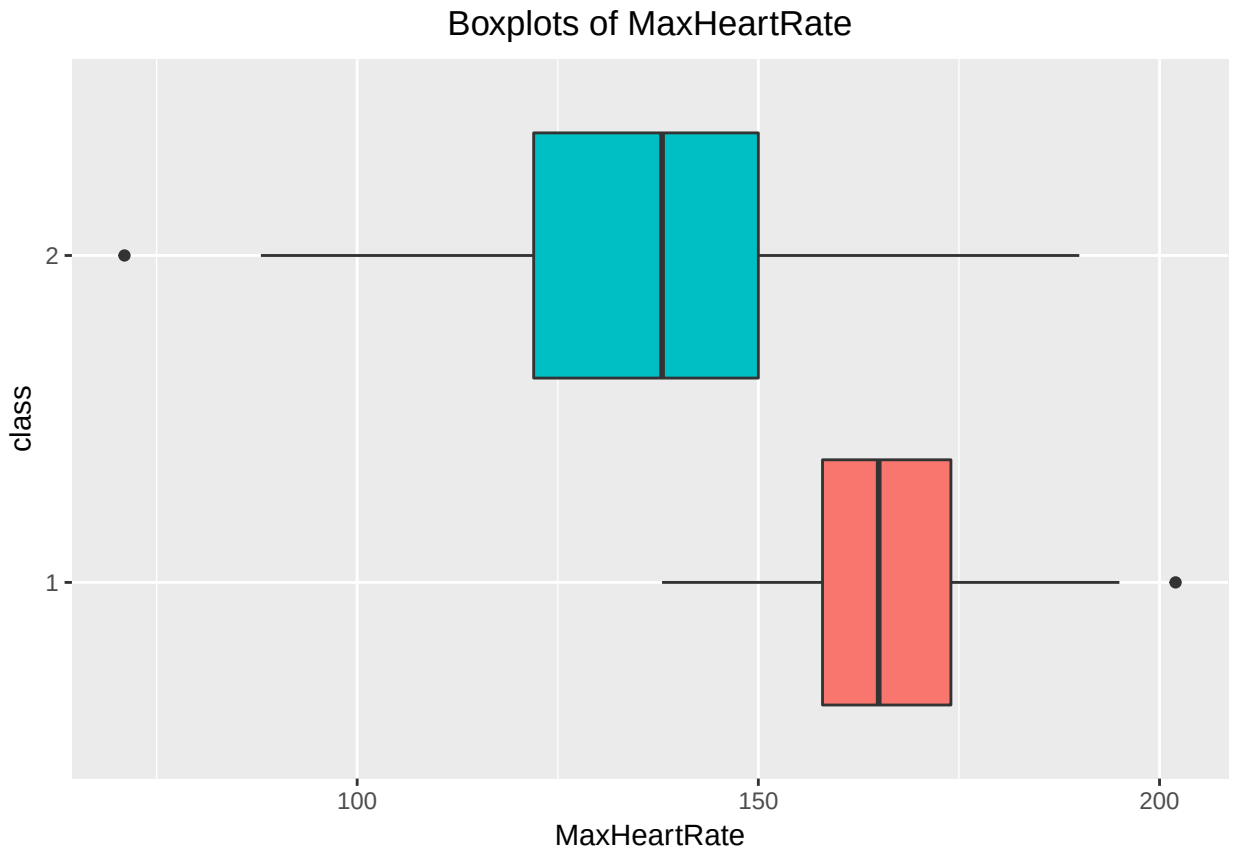
```r
# Discriminative power (pie)
plot(out, type="pie")
```



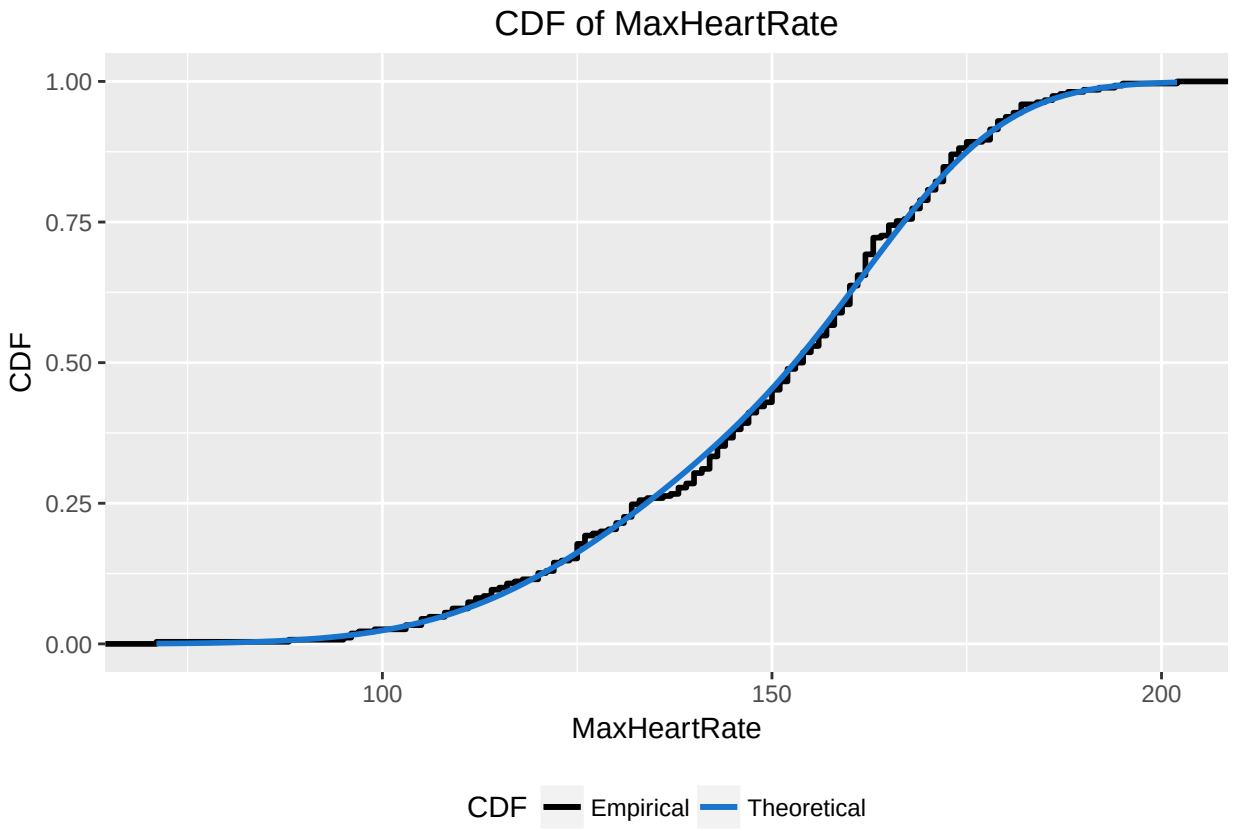

```r
# Discriminative power (bar)
plot(out, type="bar")
```



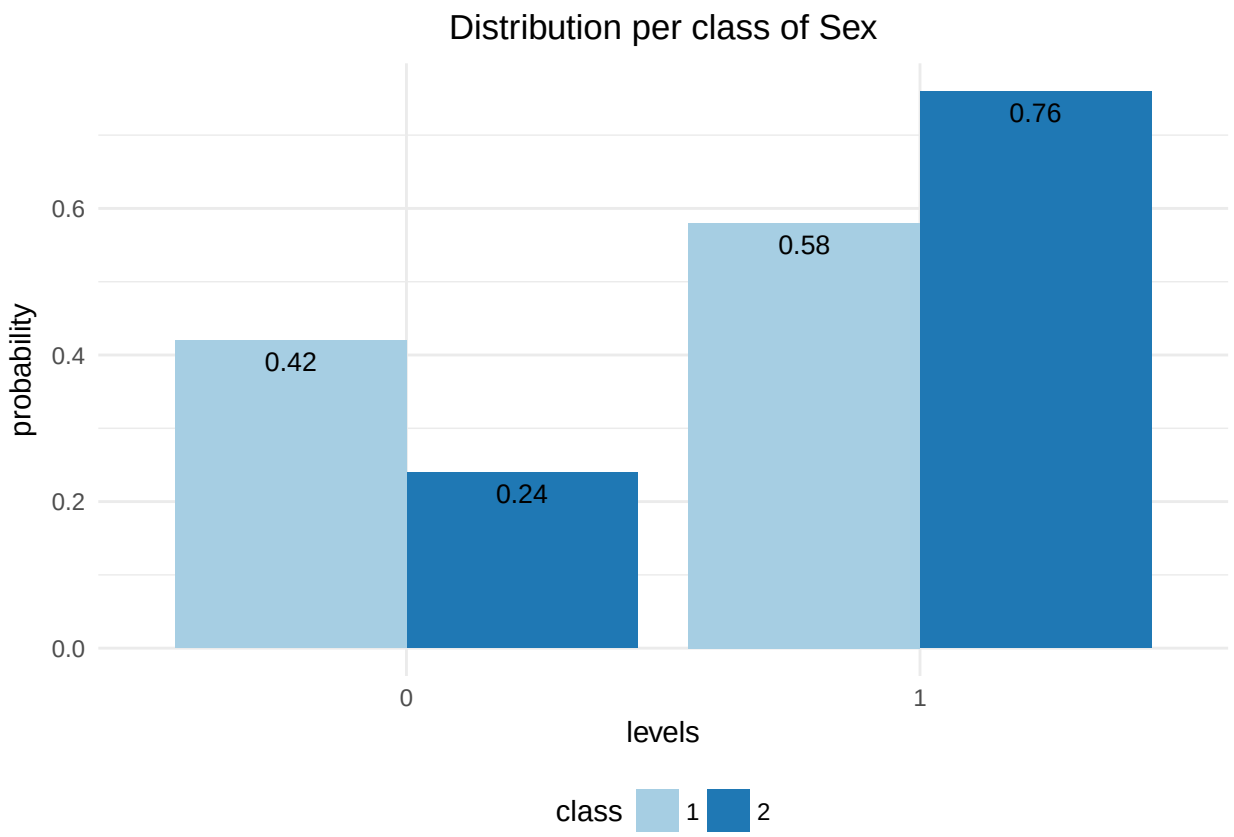

```
# Boxplot for continuous (or interger) variable
plot(out, y="MaxHeartRate", type="boxplot")
```




```
# Empirical and theoretical distributions (to check that clustering is pertinent)
plot(out, y="MaxHeartRate", type="cdf")
```

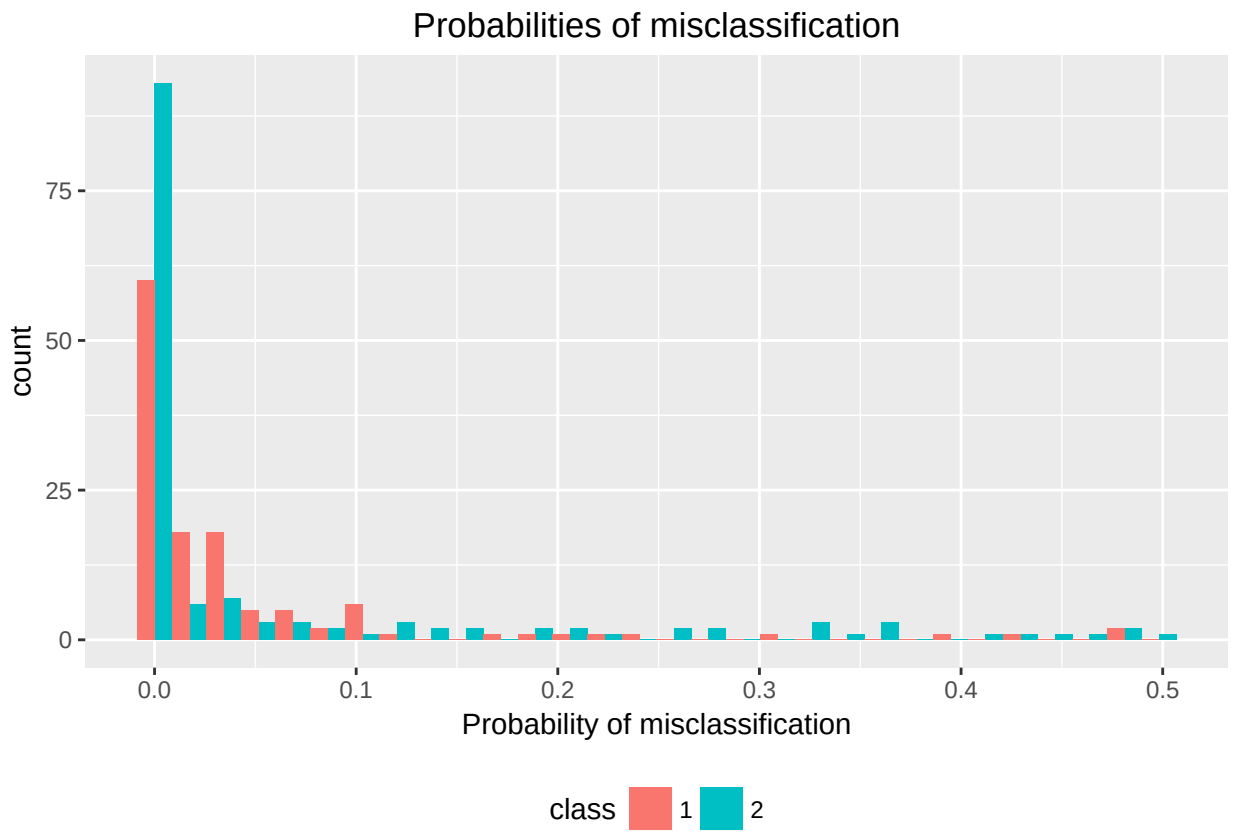



```
# Summary of categorical variable
plot(out, y="Sex")
```




```
# Summary of the probabilities of missclassification
plot(out, type="probs-class")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Start the shiny application
 VarSelShiny(out)
```