

VarSelLCM

Variable Selection for Model-Based Clustering of Mixed-Type Data Set with Missing Values.

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., Patin, E. and Sedki, M. (2018), Variable selection for mixed data clustering: Application in human population genomics, *Journal of Classification*, to appear.

Introduction

VarSelLCM permits a full model selection (detection of the relevant features for clustering and selection of the number of clusters) in model-based clustering, according to classical information criteria (BIC, MICL or AIC).

Data to analyzed can be composed of continuous, integer and/or categorical features. Moreover, missing values are managed, without any pre-processing, by the model used to cluster with the assumption that values are missing completely at random. Thus, VarSelLCM can also be used for data imputation via mixture models.

An R-Shiny application is implemented to easily interpret the clustering results

Clustering for cytological diagnosis

This section performs the whole analysis of the Golub data set. The first column indicates the type of leukemia. Clustering is made on the other (continuous) features.

```
library(VarSelLCM)
# please install the package multtest to get the data
# source("https://bioconductor.org/biocLite.R")
# biocLite("multtest")
data(golub, package = "multtest")

# one row = one individual
golub <- t(golub)
```

Clustering is done with and without variable selection. Here ICL and MICL criteria are used because the number of observations is less than the number of features (thus, BIC is not relevant).

```
# Please indicate the number of cores you wan to use for parallelisation
nb.CPU <- 4
# clustering without variable selection (about than 10/20 sec on 4 CPU)
res.noselec <- VarSelCluster(golub, 2, crit.varsel = "ICL",
                             vbleSelec = FALSE,
                             nbcores = nb.CPU)

# clustering with variable selection (about than 10/20 sec on 4 CPU)
res.selec <- VarSelCluster(golub, 2, crit.varsel = "MICL",
                           vbleSelec = TRUE,
                           nbcores = nb.CPU)
```

Summary of the results

```
summary(res.noselec)
```

Data set:

```
Number of individuals: 38
Number of continuous variables: 3051
```

Model:

```
Number of components: 2
Model selection has been performed according to the ICL criterion
```

Information Criteria:

```
loglike: -71047.52
AIC:     -83252.52
BIC:     -93245.89
ICL:     -114089.7
```

```
summary(res.selec)
```

Data set:

```
Number of individuals: 38
Number of continuous variables: 3051
```

Model:

```
Number of components: 2
Model selection has been performed according to the MICL criterion
Variable selection has been performed, 553 ( 18.13 % ) of the variables are relevant for clustering
```

Information Criteria:

```
loglike: -77235.87
AIC:     -84444.87
BIC:     -90347.55
ICL:     -103858.8
MICL:    -103858.8
Best values has been found 3 times
```

Comparison of the estimated partitions with the Adjusted Rand Index

```
# Comparison of the Adjusted Rand Index (variable selection increases the partition accuracy)
ARI(res.noselec@partitions@zMAP, golub.cl)
```

```
[1] 0.6992238
```

```
ARI(res.selec@partitions@zMAP, golub.cl)
```

```
[1] 0.7927409
```

Mixed-type data analysis

This section performs the whole analysis of the *Heart* data set. *Warning continuous features must be stored in numeric, integer features must be stored in integer and categorical features must be stored in factor.*

```
library(VarSelLCM)
# Data loading
data("heart")
head(heart)
```

	Age	Sex	ChestPainType	RestBloodPressure	SerumCholestoral
1	70	1	4	130	322
2	67	0	3	115	564
3	57	1	2	124	261
4	64	1	4	128	263
5	74	0	2	120	269
6	65	1	4	120	177

	FastingBloodSugar	ResElectrocardiographic	MaxHeartRate	ExerciseInduced
1	0	2	109	0
2	0	2	160	0
3	0	0	141	0
4	0	0	105	1
5	0	2	121	1
6	0	0	140	0

	Slope	MajorVessels	Thal	Class
1	2	3	3	2
2	2	0	7	1
3	1	0	7	2
4	2	1	7	1
5	1	1	3	1
6	1	0	7	1

Clustering is performed with variable selection. Model selection is done with BIC because $n \gg d$. The number of components is between 1 and 4. Do not hesitate to use parallelisation (here only two cores are used).

```
# Add a missing value artificially (just to show that it works!)
heart[1,1] <- NA
# Clustering with variable selection and a number of cluster between 1 and 4
# Model selection is BIC (to use MICL, the option must be specified)
out <- VarSelCluster(heart[, -13], 1:4, nbcores = 2)
```

Now, all the results can be analyzed by the Shiny application...

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

... but this analysis can also be done on R.

To get a summary of the selected model.

```
# 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
```

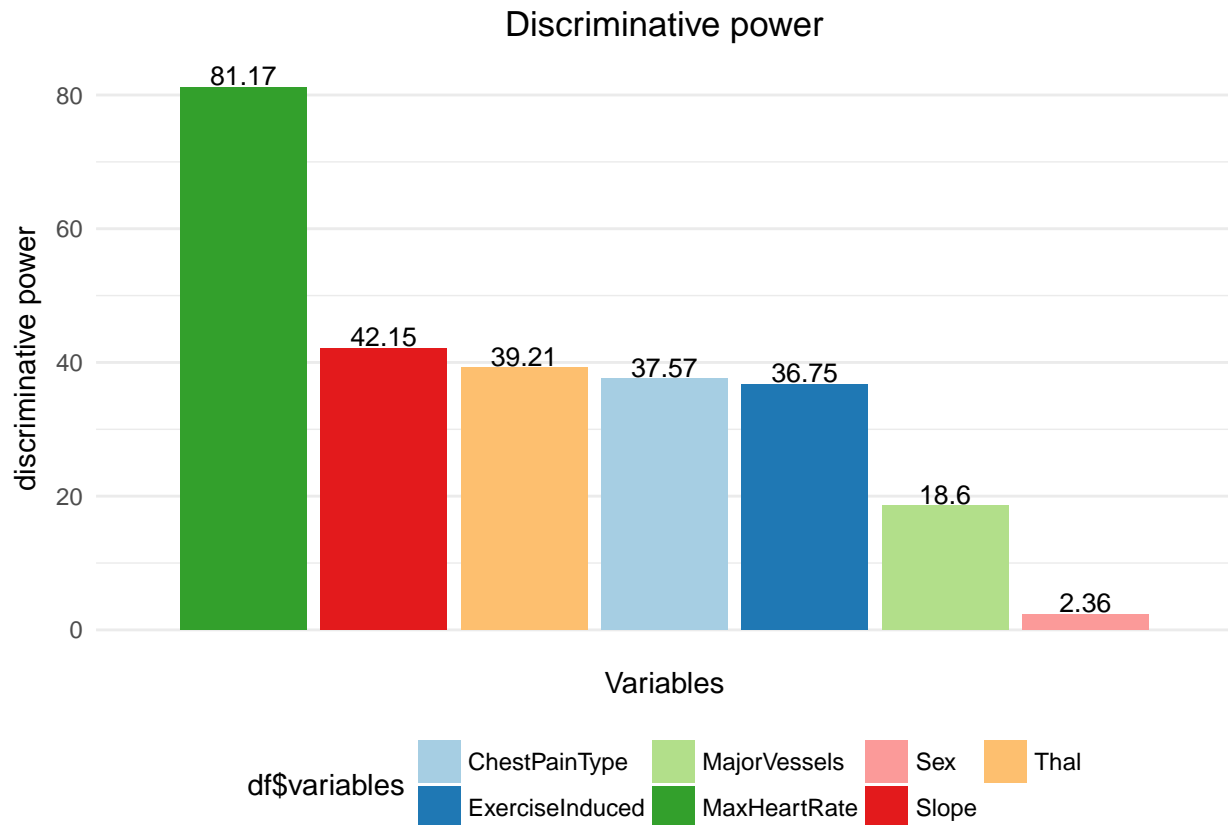
Information Criteria:

```
loglike: -6403.136
```

```
AIC:      -6441.136
BIC:      -6509.506
ICL:      -6638.116
```

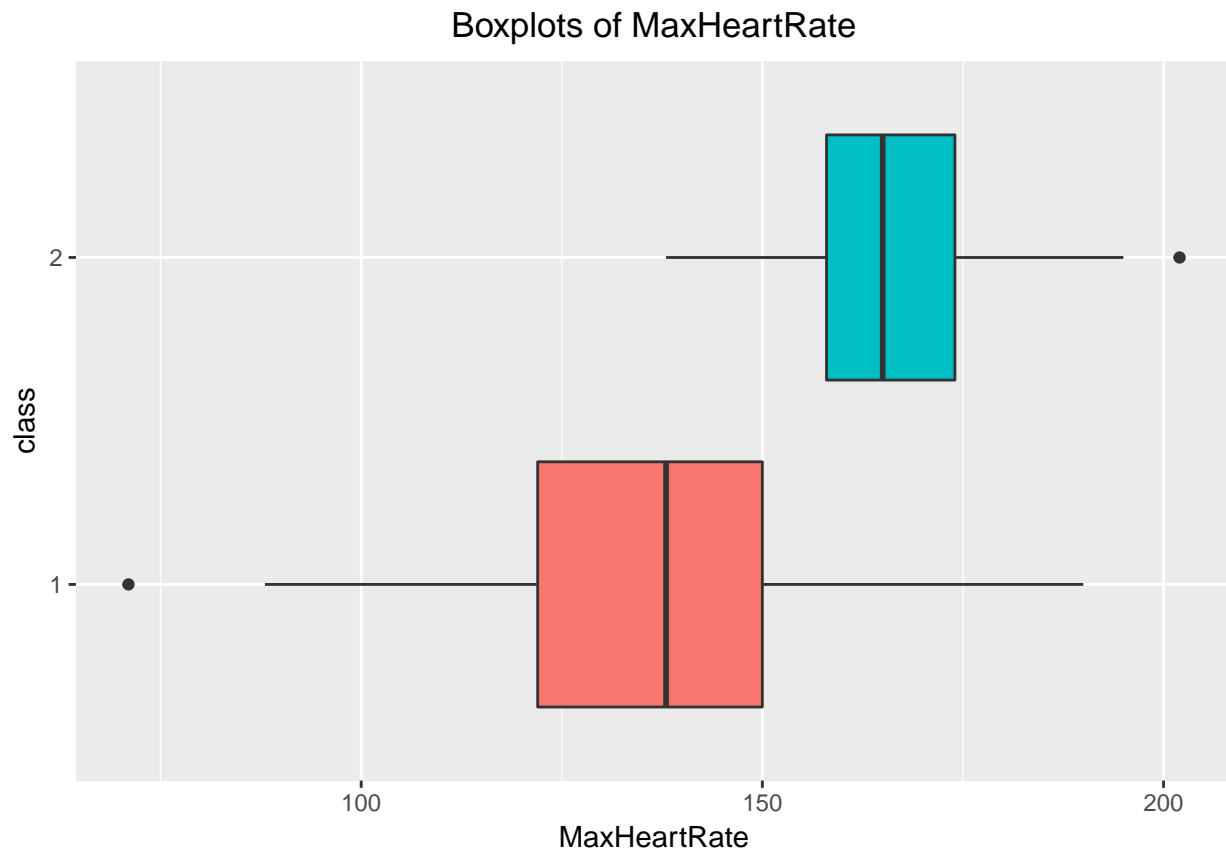
Model interpretation should focus on the most discriminative variables. These variables can be found with the following plot.

```
# Discriminative power of the variables (here, the most discriminative variable is MaxHeartRate)
plot(out, type="bar")
```



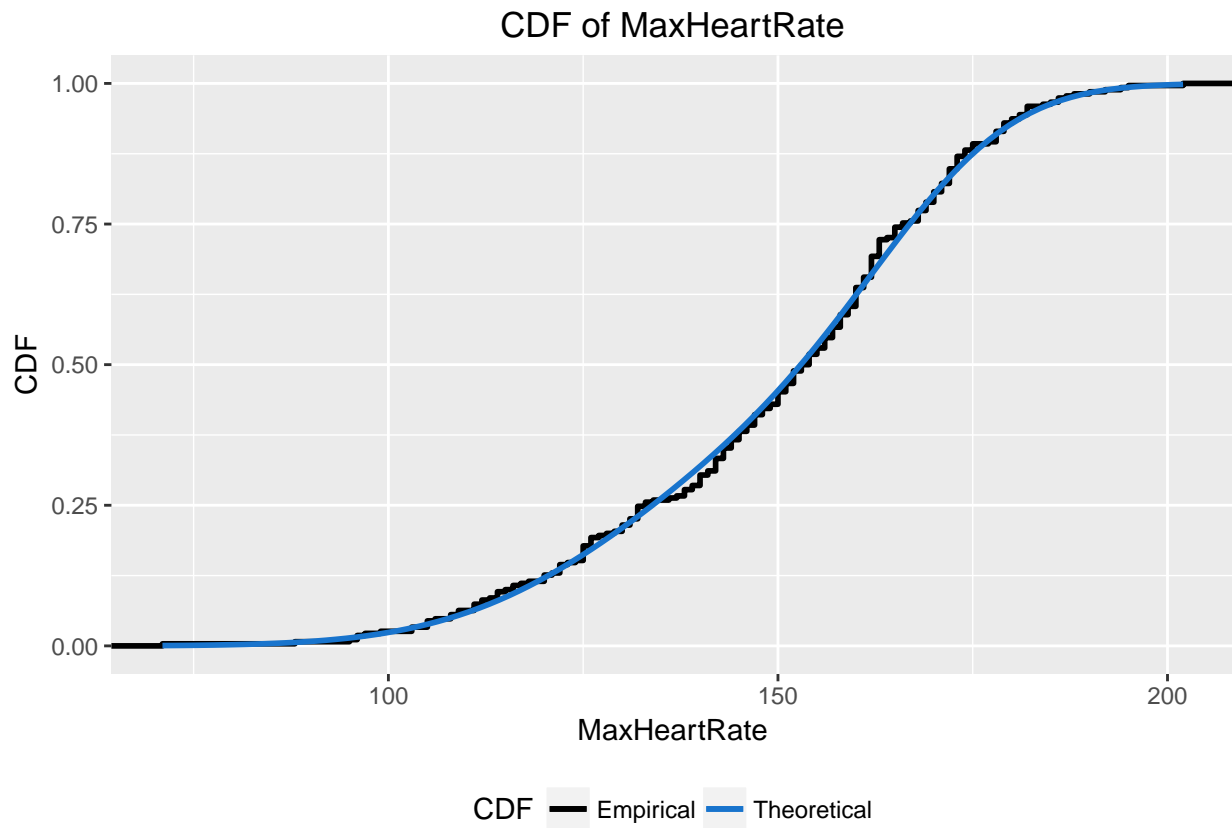
Interpretation of the most discriminative variable is based on its distribution per cluster.

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



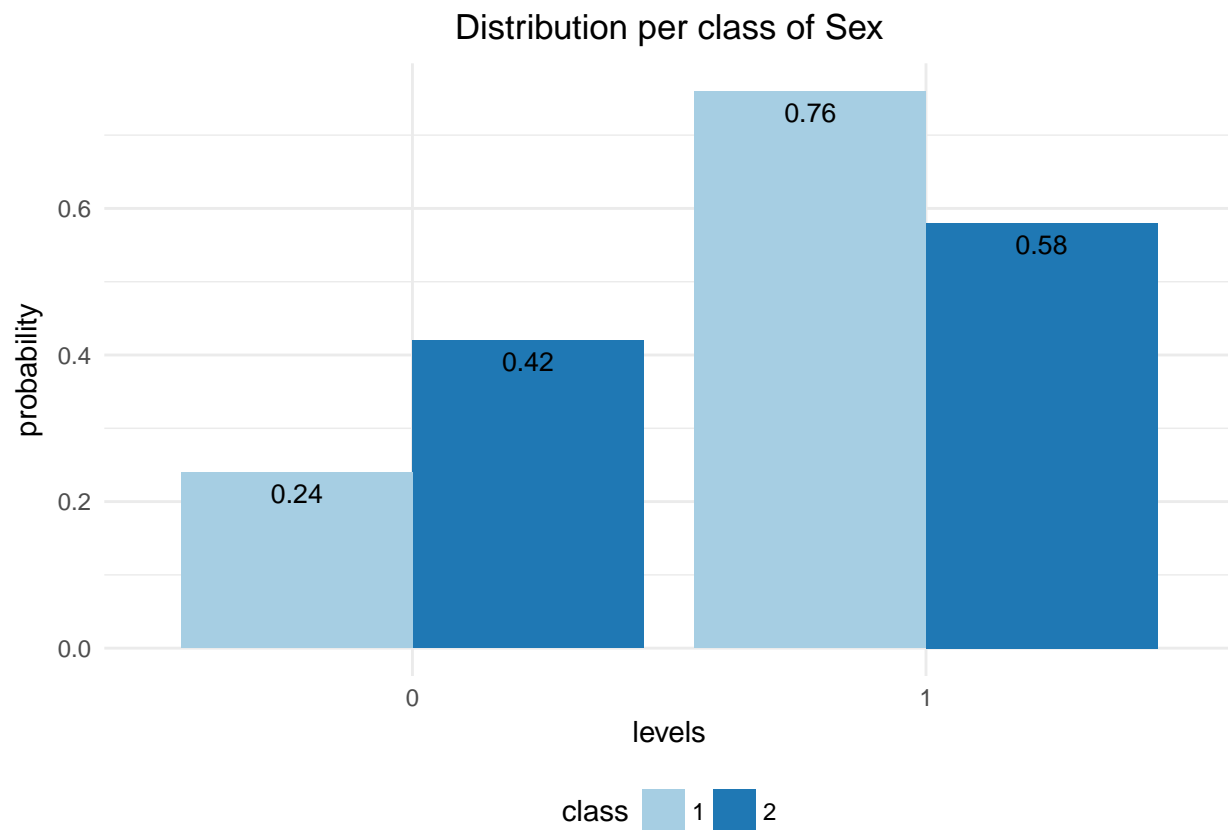
We can check that the distribution used to cluster is relevant.

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



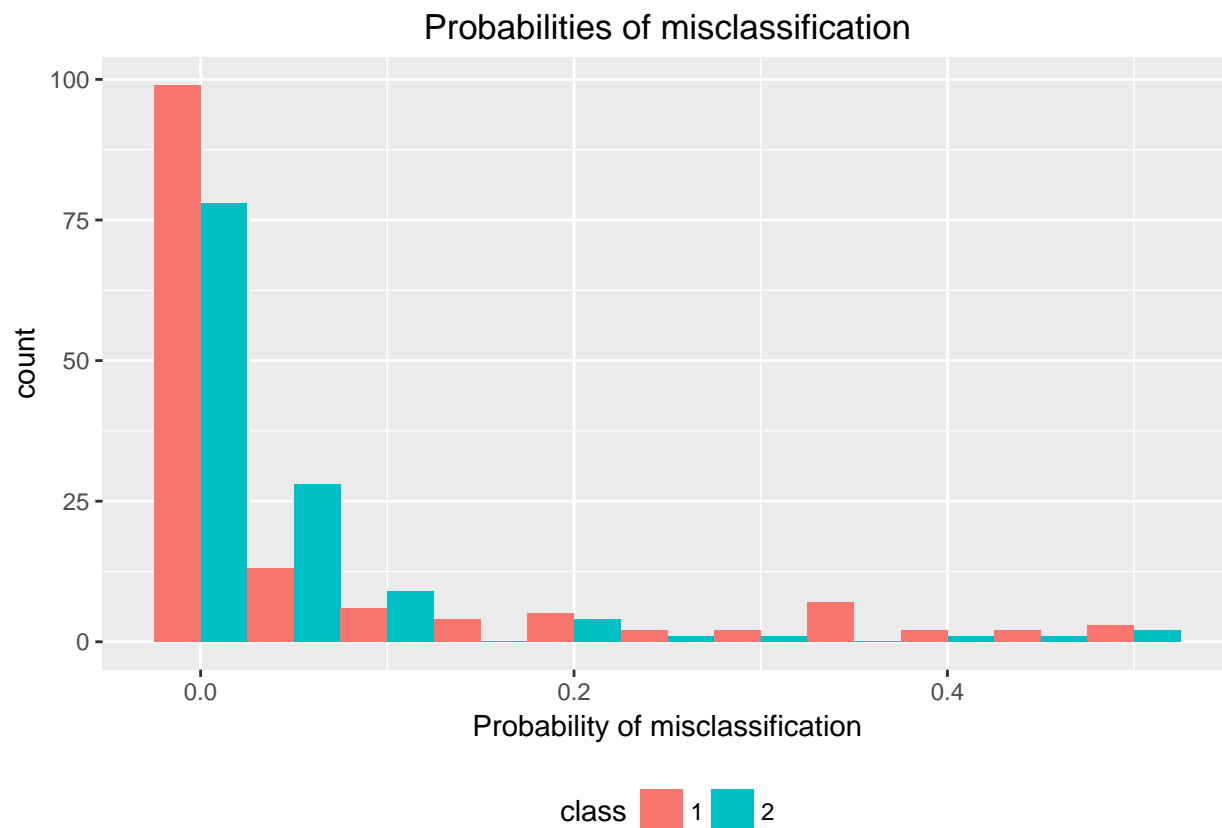
Interpretation of a categorical variable is based on its distribution per cluster.

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



Interpretation of the cluster overlaps by using the probabilities of missclassification.

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



Missing values can be imputed.

Imputation by posterior mean for the first observation

```
not.imputed <- heart[1,-13]
```

```
imputed <- VarSelImputation(out)[1,]
```

```
rbind(not.imputed, imputed)
```

	Age	Sex	ChestPainType	RestBloodPressure	SerumCholestoral
1	NA	1	4	130	322
2	58.11331	1	4	130	322

	FastingBloodSugar	ResElectrocardiographic	MaxHeartRate	ExerciseInduced
1	0	2	109	0
2	0	2	109	0

	Slope	MajorVessels	Thal
1	2	3	3
2	2	3	3