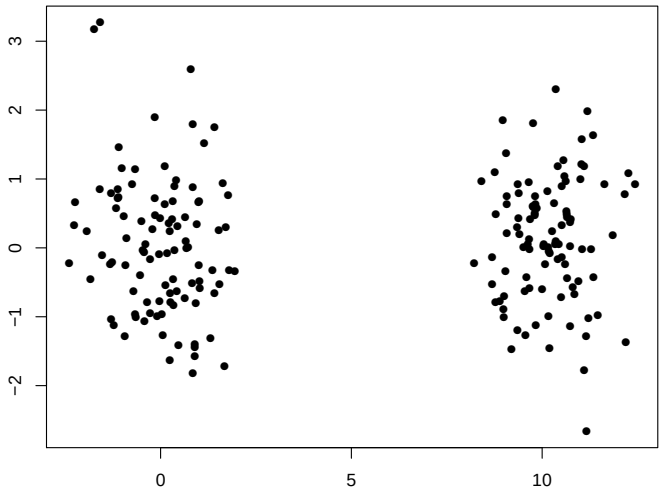
VarSelLCM: an R/C++ package for feature selection in model-based clustering of mixed-data with missing values

Matthieu Marbac[1] Mohammed Sedki[2]

[1]CREST - Ensai

[2]University of Paris-Sud and UMR Inserm-1181

Clustering and feature selection

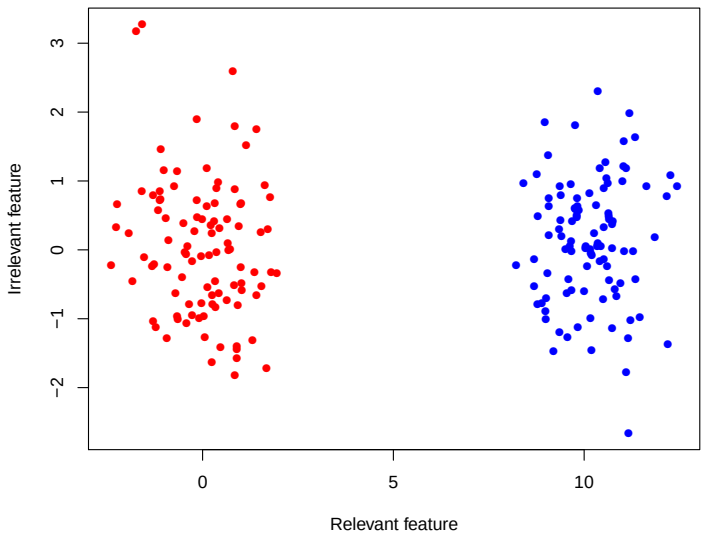


Features can be

- continuous
- categorical
- integer
- mixed-type

Missing values can occurred (MCAR)

Clustering and feature selection



- Estimation of a classification rule
- Evaluation of the risk of misclassification.
- Interpretation of the clusters.
- Estimation of the number of clusters.
- Detection of relevant features.

Model-based clustering

Main idea:

Model the distribution of the observed data $\boldsymbol{X}$.

Mixture model:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x};\boldsymbol{\theta}_k)$$

where $\boldsymbol{\theta}$ groups all the parameters.

The distribution of the components depends on the type of features. To cluster continuous data, VarSelLCM uses Gaussian mixtures:

$$f_k(\boldsymbol{x};\boldsymbol{\theta}_k) = \phi(\boldsymbol{x};\mu_k,\Sigma_k)$$

Mixture model permits:

- computation of probabilities of classification
- model selection with information criteria (BIC, MICL)

Mixture model and feature selection

Main idea

Only a subset of variables explains the unobserved partition.

Feature selection permits

- reduction the variance of the estimators
- an easier interpretation

Mixture model with feature selection:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = f(\boldsymbol{x}_{irrelevant};\boldsymbol{\alpha}) \sum_{k=1}^{g} \pi_k f_k(\boldsymbol{x}_{relevant};\boldsymbol{\theta}_k)$$

VarSelLCM a simultaneous estimation of the partition and the role of the features (with BIC or MICL).

A real example: clustering

The function of clustering with its main arguments

```
VarSelCluster(x,
              gvals,
              vbleSelec = TRUE,
              crit.varsel = "BIC",
              nbcores = 1)
```

- **x** contains the data to cluster. Continuous variables must be "numeric", count variables must be "integer" and categorical variables must be "factor"
- **gvals** defines number of components to consider.
- **vbleSelec** indicates if a feature selection is done
- **crit.varsel** defines the information criterion used for the feature selection
- **nbcores** defines the number of cores used by the algorithm

Information criteria for model selection

BIC

- Classical criterion for model selection
- Many observations are required
- Specific EM algorithm performs simultaneously model selection and maximum likelihood inference

MICL

- Derived from ICL (clustering criterion)
- An algorithm performs model selection before maximum likelihood inference
- Computationally intensive if too many observations ($> 10^4$).

A real example: clustering

```
library(VarSelLCM)
data(heart)
ztrue <- heart[,"Class"]
x <- heart[,-13]
x[1,1] <- NA
```

Standard clustering

```
res_without <-   VarSelCluster(x, gvals = 1:3,
                               vbleSelec = FALSE,
                               crit.varsel = "BIC")
```

Cluster analysis with feature selection

```
res_with <-   VarSelCluster(x, gvals = 1:3,
                            vbleSelec = TRUE,
                            crit.varsel = "BIC")
```

Benefits of feature selection

BIC is improved by feature selection.

```
c(BIC(res_without), BIC(res_with))
```

```
[1] -6516.216 -6509.506
```

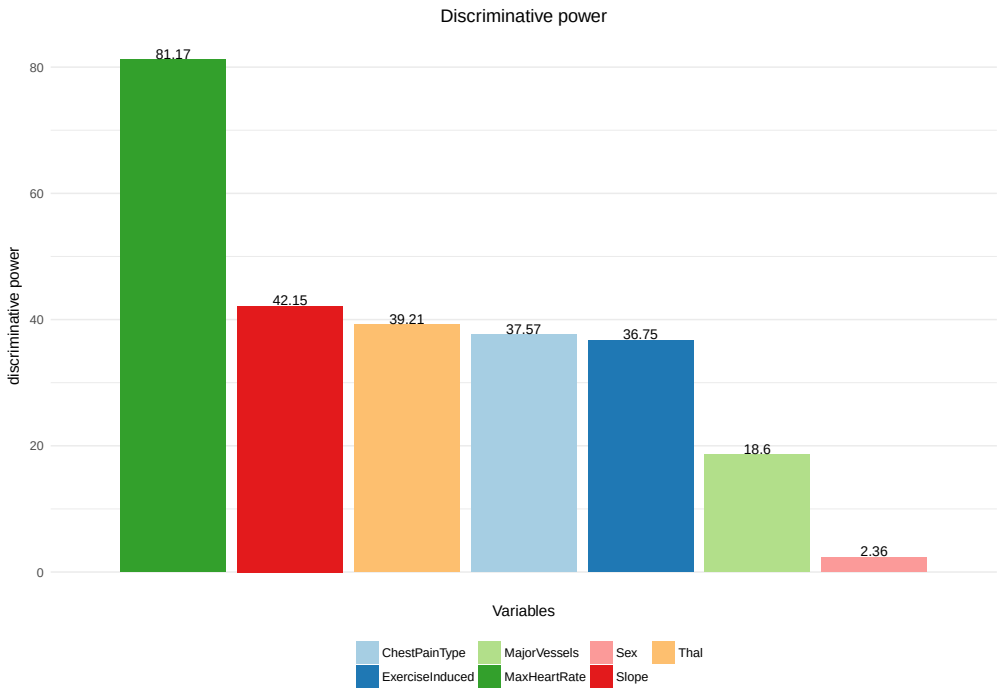
The partition is improved by feature selection.

```
c(ARI(ztrue, fitted(res_without)),
  ARI(ztrue, fitted(res_with)))
```

```
[1] 0.2218655 0.2661321
```

Results interpretation

```
plot(res_with)
```

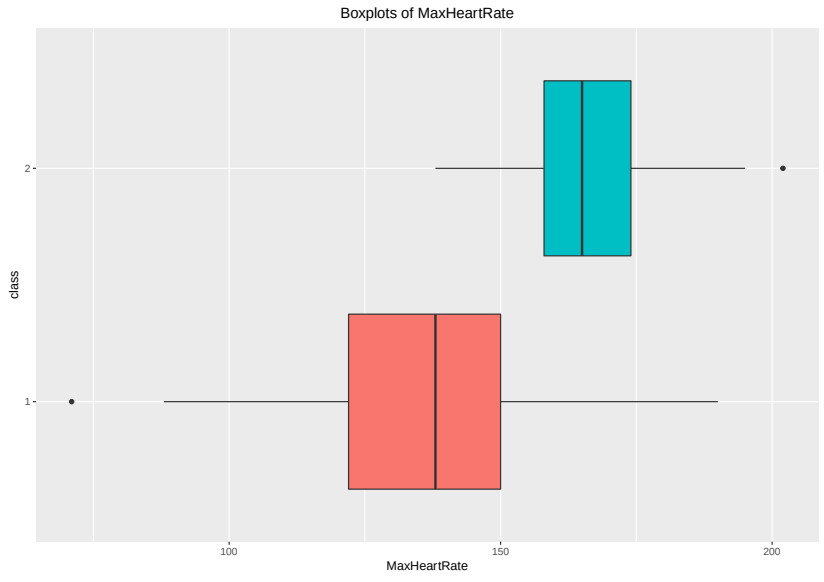


The greater this index, the more the feature distinguishes the clusters.

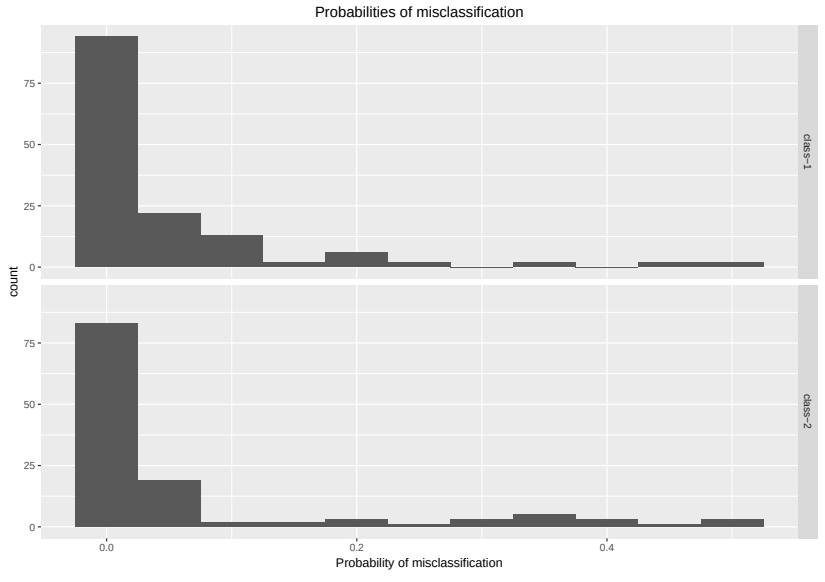
$$\ln \frac{p(\{X_j discrim\}|\hat{\mathbf{z}}, \mathbf{x})}{p(\{X_j notdiscrim\}|\hat{\mathbf{z}}, \mathbf{x})}$$

Results interpretation

```
plot(x=res_with,
     y="MaxHeartRate")
```




```
plot(x=res_with,
     type="probs-class")
```



Partition and probabilities of classification

Classification

```
round(
 predict(
   res_with,
   newdata = x[2,]),
 2)
```

```
     class-1 class-2
[1,]    0.62    0.38
```

```
predict(
  res_with,
  newdata = x[2,],
  type = "partition")
```

```
[1] 1
```

Imputation

```
VarSelImputation(
  res_with,
  newdata = x[1,],
  method = "postmean")[1:3]
```

```
       Age Sex ChestPainType
1 58.11326   1             4
```

```
VarSelImputation(
  res_with,
  newdata = x[1,],
  method = "sampling")[1:3]
```

```
  Age Sex ChestPainType
1  54   1             4
```

Some applications

The EDEN mother-child study

- 2,000 children (many missing values)
- 25 features (continuous and categorical)
- BIC is used for model selection

Application in human population genomics

- 1,318 individuals from 35 populations of western central Africa
- 160K independent markers
- MICL is used for model selection

Conclusion

VarSelLCM.2.1.2:

- model-based clustering
- mixed-type data with missing value
- feature selection
- Shiny interface (for results interpretation)

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.