ordinalClust

An R package to analyse ordinal data

Margot Selosse [1], Julien Jacques[1], Christophe Biernacki[2]

[1] Laboratoire ERIC, Université Lumière Lyon 2
[2] INRIA Lille, Université de Lille

July 2019

Summary

Introduction

Ordinal Data ?

Definition : An ordinal variable *x* takes values among *m* full ordered levels.

$$\mu \in \{1, ..., m\} \text{ with } 1 < ... < m$$

Examples :

- Marketing : customer satisfaction surveys
- Sociology : education levels

ordinalClust ?

R package available on CRAN (version 1.3.3) to :

- classify,
- cluster,
- co-cluster

ordinal data.

BOS distribution [1]

Distribution

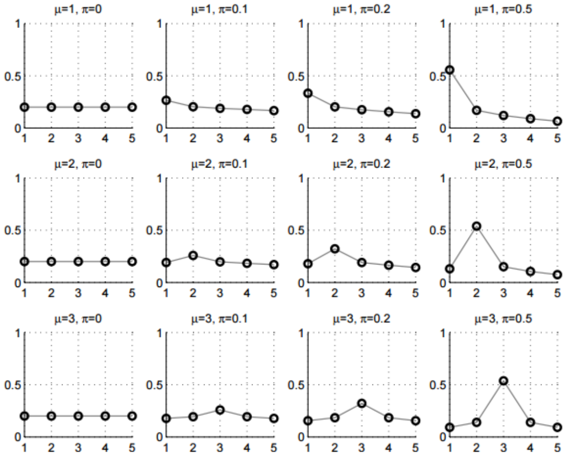

FIGURE – BOS distribution $p(x; \mu, \pi)$: shape for $m = 5$ and for different values of μ and π

Co-clustering

Classical Latent Block Model

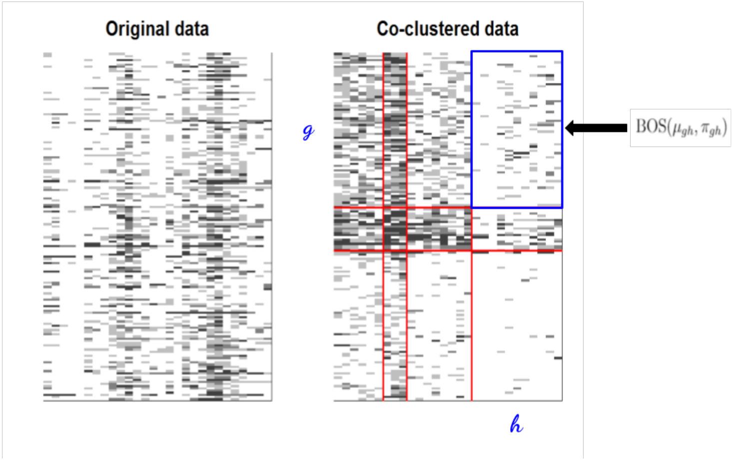

FIGURE – Latent Block Model : each block (gh) follows a BOS distribution of parameters (μ_{gh}, π_{gh})

Model hypothesis

- $\boldsymbol{x}$ matrix with N lines, J columns
- G clusters in line, H clusters in column
- We have the one-hot matrix $\boldsymbol{v}$ which indicates the row-cluster belonging
- We have the one-hot matrix $\boldsymbol{w}$ which indicates the column-cluster belonging
- The crossing between the g^{th} row-cluster and the h^{th} column cluster is called **a block**

- partitions in line $\boldsymbol{v}$ and in column $\boldsymbol{w}$ are independent : $p(\boldsymbol{v}, \boldsymbol{w}) = p(\boldsymbol{v}) \times p(\boldsymbol{w})$
- Element x_{ij} are i.i.d, conditionally to partitions : $p(\boldsymbol{x}|\boldsymbol{v}, \boldsymbol{w}) = \prod_{ij} p(x_{ij}|\boldsymbol{v}, \boldsymbol{w})$

Model inference

Aim

- Find $\theta = (\mu_{gh}, \pi_{gh}, \gamma_g, \rho_h) \;\; \forall \, (g, h)$
- partitions $\boldsymbol{v}$ (rows) and $\boldsymbol{w}$ (columns) are missing

Using EM algorithm ?

E step requires the computation of the joint conditional distributions of the missing labels :

$$p(v_{ig} w_{jh} = 1 | \boldsymbol{x}; \theta) \; \forall i, j, g, h.$$

It implies to compute $G^N \times H^J$ terms at each iteration.
$\Rightarrow$ The SEM-Gibbs algorithm [5] is used.

What about clustering and classification ?

They are the same models but :

- Clustering does not have column-partitions $\boldsymbol{w}$: we have to estimate $\boldsymbol{v}$ and θ
- Classification does not have $\boldsymbol{v}$ nor $\boldsymbol{w}$, we just have to estimate the parameters θ

Application in Oncology

Getting started with ordinalClust

```
library(ordinalClust)
data("dataqol")
data("dataqol.classif")
```

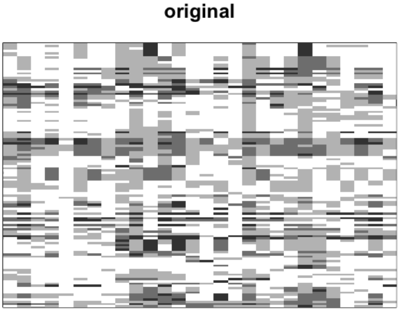

FIGURE – Original data.

Main arguments for ordinalClust

- x : ordinal data set
- m : number of levels of ordinal data
- kr : number of row-clusters
- kc : number of column-clusters
- nbSEM : number of iterations
- nbSEMburn : number of iterations for burn-in period
- init : type of initialization (random, kmeans...)

Clustering

```
clust <- bosclust(x = x, kr = 3, m = 4, nbSEM = nbSEM,
                  nbSEMburn = nbSEMburn, init = init)
```

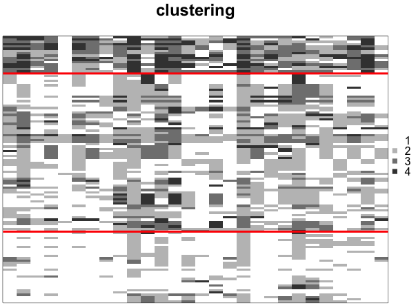

FIGURE – Clustering obtained when following the given example.

Co-clustering

```
coclust <- boscoclust(x = x, kr = 3, kc = 3, m = 4,
                      nbSEM = nbSEM, nbSEMburn = nbSEMburn,
                      init = init)
```

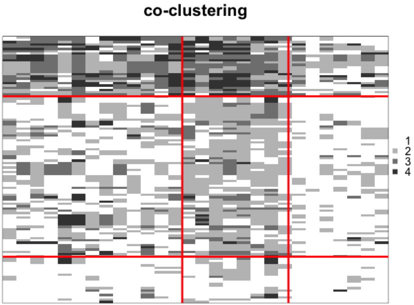

FIGURE – Co-clustering obtained when following the given example.

Classification

```
classif <- bosclassif(x = x.train, y = y.train,
             kr = 2, kc = 3, m = m, nbSEM = nbSEM,
             nbSEMburn = nbSEMburn, init = init)
new.prediction <- predict(classif, x.val)
```

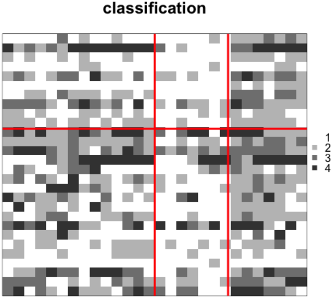

FIGURE – Classification plot obtained when following the given example.

Conclusion

Conclusion

- A documentation is available on HAL [2].
- the package is able to take into account variables that do not have the same number of levels m [3]
- Package needs better summary function and visualization as well.
- Models are applicable to mixed-type data. [4] Another package (mixedClust) will be available soon on CRAN.

References

- C. Biernacki, J. Jacques. Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm. *J. Stat Comput* 26 : 929, 2016.
- Selosse, Margot and Jacques, Julien and Biernacki, Christophe, ordinalClust : an R package for analyzing ordinal data, https ://hal.inria.fr/hal-01678800
- M. Selosse, and J. Jacques and C. Biernacki and F. Cousson-Gélie. Analyzing quality of life survey using constrained co-clustering model for ordinal data and some dynamic implication.
- M. Selosse, J. Jacques, Julien and C. Biernacki. Model-based co-clustering for mixed type data.
- Keribin, C. and Govaert, G. and Celeux, G., "Estimation d'un modèle à blocs latents par l'algorithme SEM". *2èmes Journéees de Statistique*, 2010.