ordinalClust
An R package to analyse ordinal data

Margot Selosse¹, Julien Jacques¹, Christophe Biernacki²

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

July 2019
Summary

1 Introduction

2 BOS distribution [1]

3 Co-clustering

4 Application in Oncology

5 Conclusion
Introduction
**Definition**: An ordinal variable $x$ takes values among $m$ full ordered levels.

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

**Examples**:
- Marketing: customer satisfaction surveys
- Sociology: education levels
R package available on CRAN (version 1.3.3) to:

- classify,
- cluster,
- co-cluster

ordinal data.
BOS distribution [1]
**Figure** – BOS distribution $p(x; \mu, \pi)$: shape for $m = 5$ and for different values of $\mu$ and $\pi$
Co-clustering
Classical Latent Block Model

**Figure** — Latent Block Model: each block \((gh)\) follows a BOS distribution of parameters \((\mu_{gh}, \pi_{gh})\)
Model hypothesis

- \( x \) matrix with \( N \) lines, \( J \) columns
- \( G \) clusters in line, \( H \) clusters in column
- We have the one-hot matrix \( v \) which indicates the row-cluster belonging
- We have the one-hot matrix \( w \) 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 \( v \) and in column \( w \) are independent: \( p(v, w) = p(v) \times p(w) \)
- Element \( x_{ij} \) are i.i.d, conditionally to partitions: \( p(x|v, w) = \prod_{ij} p(x_{ij}|v, w) \)
Model inference

Aim

- Find $\theta = (\mu_{gh}, \pi_{gh}, \gamma_{g}, \rho_{h}) \quad \forall (g, h)$
- partitions $v$ (rows) and $w$ (columns) are missing

Using EM algorithm?

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

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

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 $w$: we have to estimate $v$ and $\theta$
- Classification does not have $v$ nor $w$, we just have to estimate the parameters $\theta$
Application in Oncology
Getting started with ordinalClust

```r
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
- \( \text{nbSEM} \) : number of iterations
- \( \text{nbSEMburn} \) : number of iterations for burn-in period
- \( \text{init} \) : 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

\[ \text{coclust} \leftarrow \text{boscoclust}(x = x, \ kr = 3, \ kc = 3, \ m = 4, \ nbSEM = nbSEM, \ nbSEMburn = nbSEMburn, \ init = init) \]

\textbf{Figure} – Co-clustering obtained when following the given example.
**Classification**

```r
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.

Selosse, Margot and Jacques, Julien and Biernacki, Christophe, ordinalClust : an R package for analyzing ordinal data, https://hal.inria.fr/hal-01678800

