LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Quest: A Generalized Motif Bicluster Algorithm

Sebastian Kaiser and Friedrich Leisch

Institut für Statistik
Ludwig-Maximilians-Universität München

UseR 2009, 09.07.2009, Rennes, France

Overview

Outline:

I. Introduce Biclustering

II. New Bicluster Algorithm

III. New Developments in the `biclust` Package

IV. Example

V. Summary and Future Work

I. Biclustering

Why Biclustering?

- Simultaneous clustering of 2 dimensions
- Large datasets where traditional clustering of columns **or** rows leads to diffuse results
- Only parts of the data influence each other

I. Biclustering

Initial Situation:

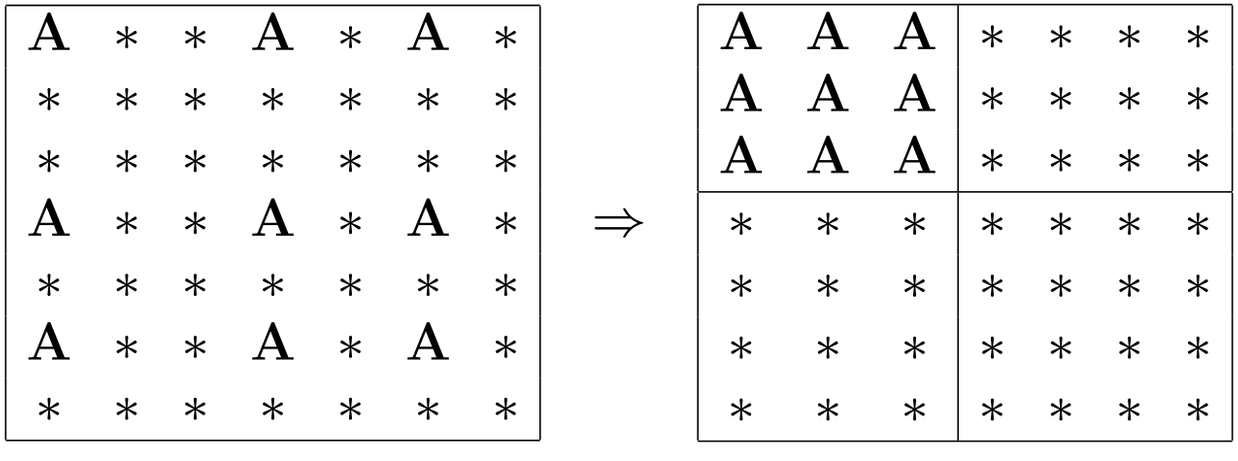
Two-Way Dataset

	c_1	$\dots$	c_i	$\dots$	c_m
r_1	a_{11}	$\dots$	a_{i1}	$\dots$	a_{m1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
r_j	a_{1j}	$\dots$	a_{ij}	$\dots$	a_{mj}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
r_n	a_{1n}	$\dots$	a_{in}	$\dots$	a_{mn}

I. Biclustering

Goal:

Finding subgroups of rows and columns which are as similar as possible to each other and as different as possible to the rest.



I. Biclustering

More than one bicluster? Most Bicluster Algorithms are iterative. To find the next bicluster given n-1 found biclusters you have to either

- ignore the n-1 already found biclusters,
- delete rows and/or columns of the found biclusters or
- mask the found biclusters with random values.

II. Bicluster Algorithms: In the Package

Chosen sample of algorithms in order to cover most bicluster outcomes.

Bimax(Barkow et al., 2006): Groups with ones in binary matrix

CC (Cheng and Church, 2000): Constant values

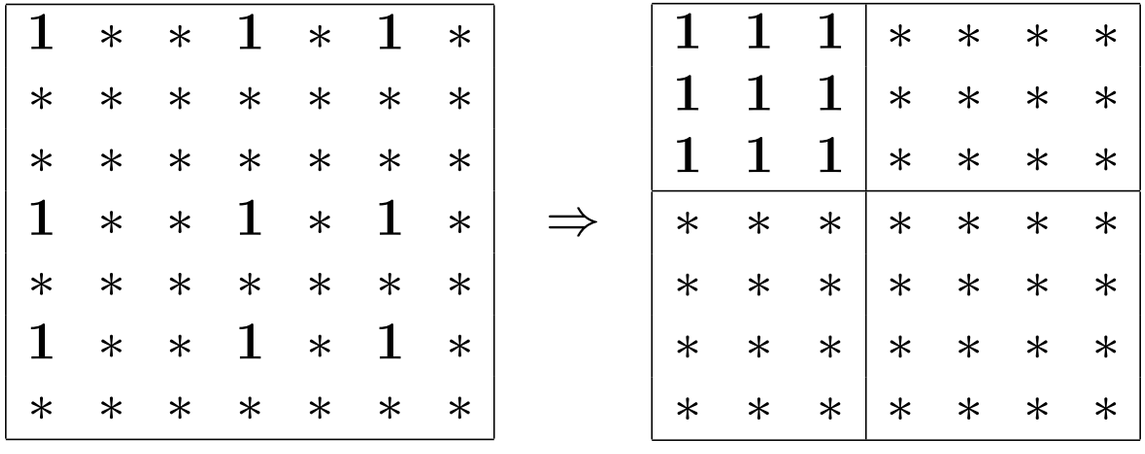
Plaid (Turner et al., 2005): Constant values over rows or columns

Spectral (Kluger et al., 2003): Coherent values over rows and columns

Xmotif (Murali and Kasif, 2003): Coherent correlation over rows and columns

II. Bicluster Algorithms

Bimax

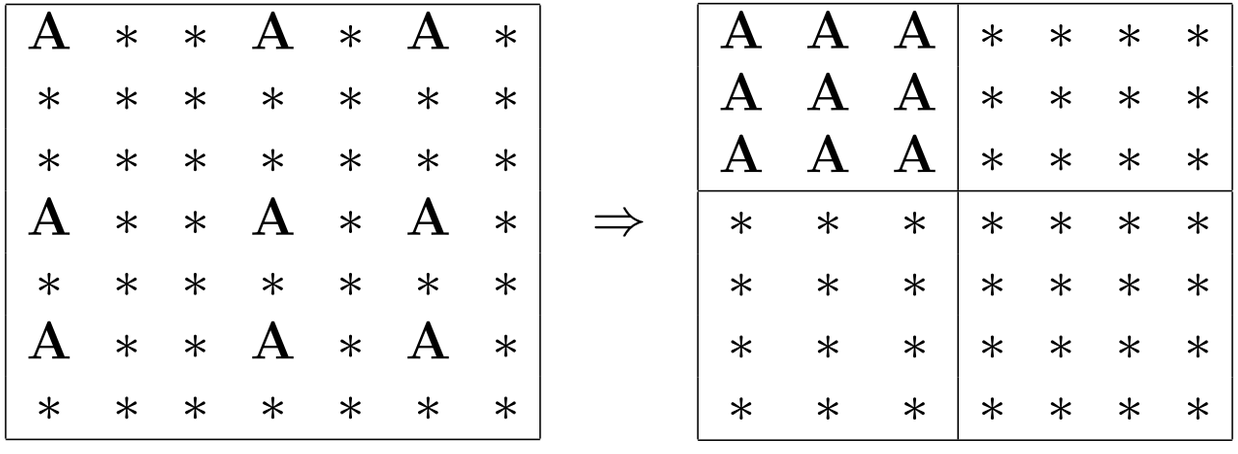


- Finds subgroups of ones in a binary data matrix.

- Suitable if only one kind of outcome is interesting.

II. Bicluster Algorithms

Xmotif

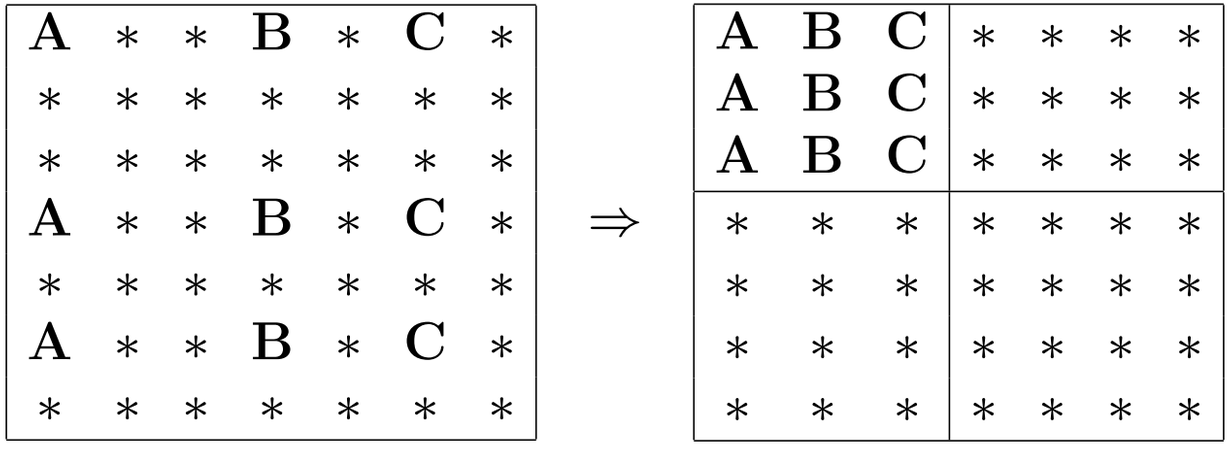


- Finds subgroups of equal outcomes.

- Suiteable if equal nominal or ordinal values are wanted.

II. Bicluster Algorithms

Quest (nominal)

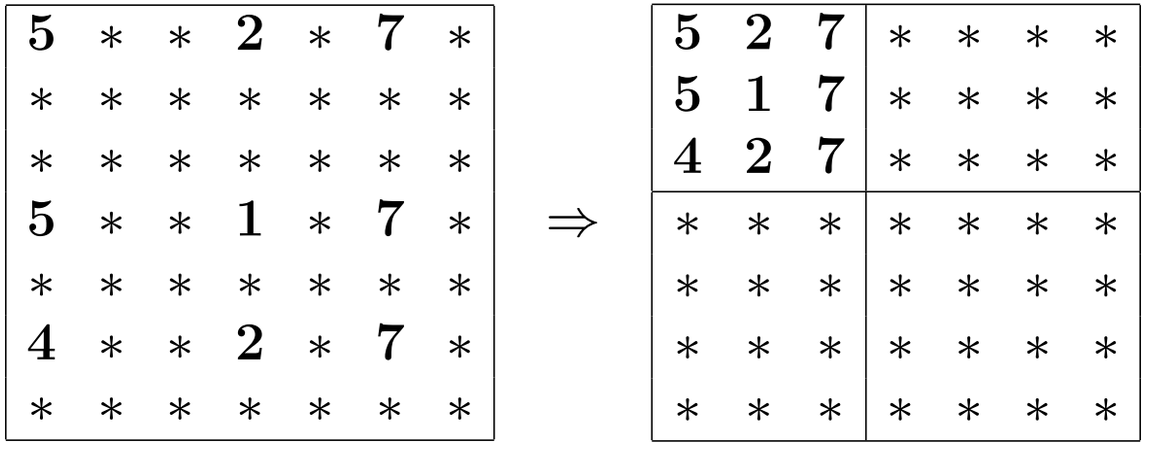


- Finds subgroups of equal outcomes over the variables.

- Suiteable if equal patterns of nominal or ordinal values are wanted.

II. Bicluster Algorithms

Quest (ordinal)



- Finds subgroups of outcomes inside a given intervall or a given size of intervall over the variables.

- Suiteable if similar patterns of ordinal or continuous values are wanted.

II. Bicluster Algorithms

Quest (continuous)

74	$*$	$*$	0.23	$*$	-13	$*$
$*$	$*$	$*$	$*$	$*$	$*$	$*$
$*$	$*$	$*$	$*$	$*$	$*$	$*$
80.5	$*$	$*$	0.35	$*$	-12.75	$*$
$*$	$*$	$*$	$*$	$*$	$*$	$*$
77	$*$	$*$	0.27	$*$	-11.99	$*$
$*$	$*$	$*$	$*$	$*$	$*$	$*$

$\Rightarrow$

74	0.23	-13	$*$	$*$	$*$	$*$
80.5	0.35	-12.75	$*$	$*$	$*$	$*$
77	0.27	-11.99	$*$	$*$	$*$	$*$
$*$	$*$	$*$	$*$	$*$	$*$	$*$
$*$	$*$	$*$	$*$	$*$	$*$	$*$
$*$	$*$	$*$	$*$	$*$	$*$	$*$
$*$	$*$	$*$	$*$	$*$	$*$	$*$

- Finds subgroups of outcomes having a high likelihood for a joint normal distribution over the variables.
- Suiteable if similar patterns of continuous values are wanted.
- Expandable on other distributions.

III. The `biclust` - Package

Function: biclust

The main function of the package is

```
biclust(data,method=BCxxx(),number,...)
```

with:

data: The preprocessed data matrix
method: The algorithm used (E. g. BCCC() for CC)
number: The maximum number of bicluster to search for
... : Additional parameters of the algorithms

Returns an object of class Biclust for uniform treatment.

III. The `biclust` - Package

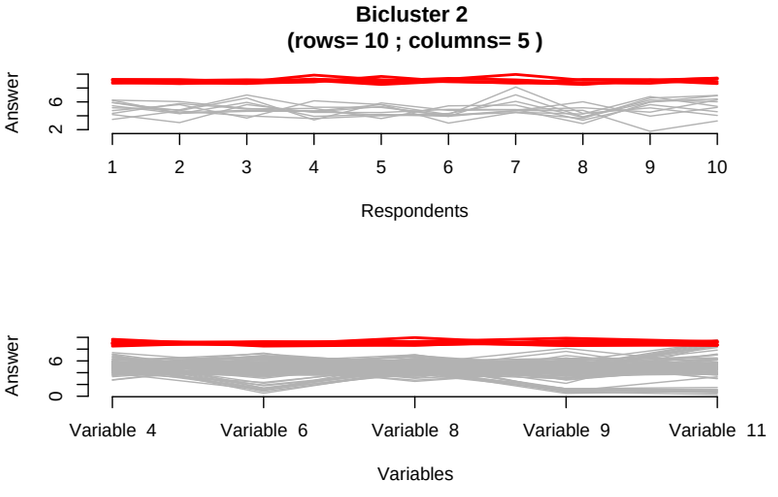
Additional methods

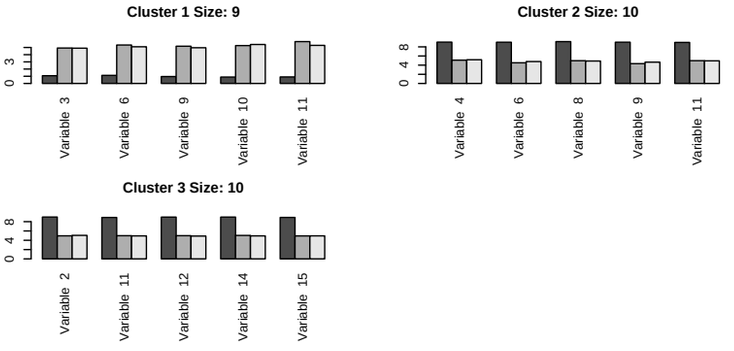
Preprocessing: `discretize()`, `binarize()`, ...

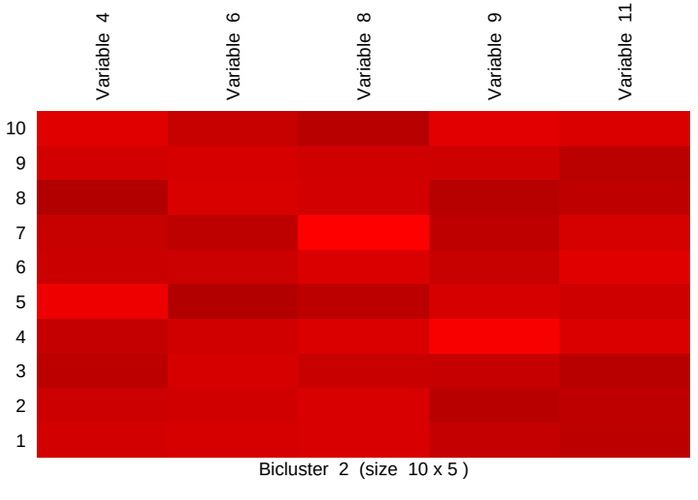
Visualization: `parallelCoordinates()`, `drawHeatmap()`, `plotclust()`, ...

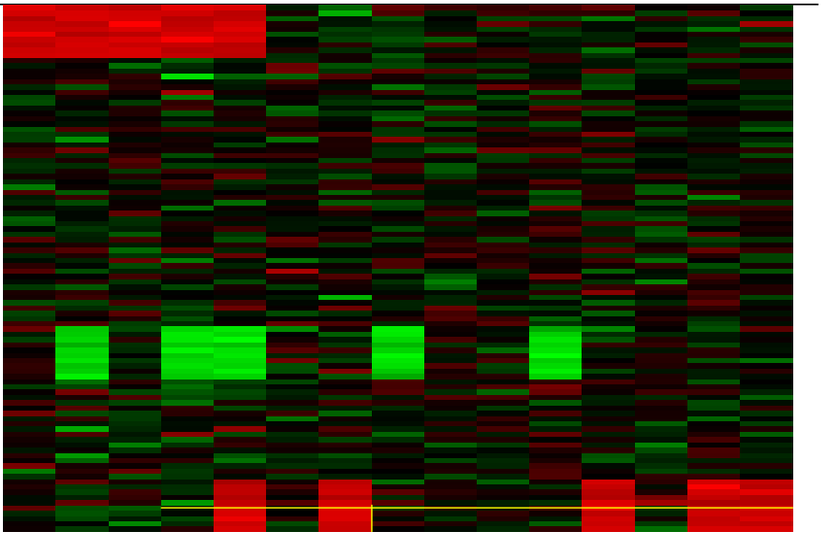
Validation: `jaccardind()`, `clusterVariance()`, ...

III. The biclust - Package: Visualizations



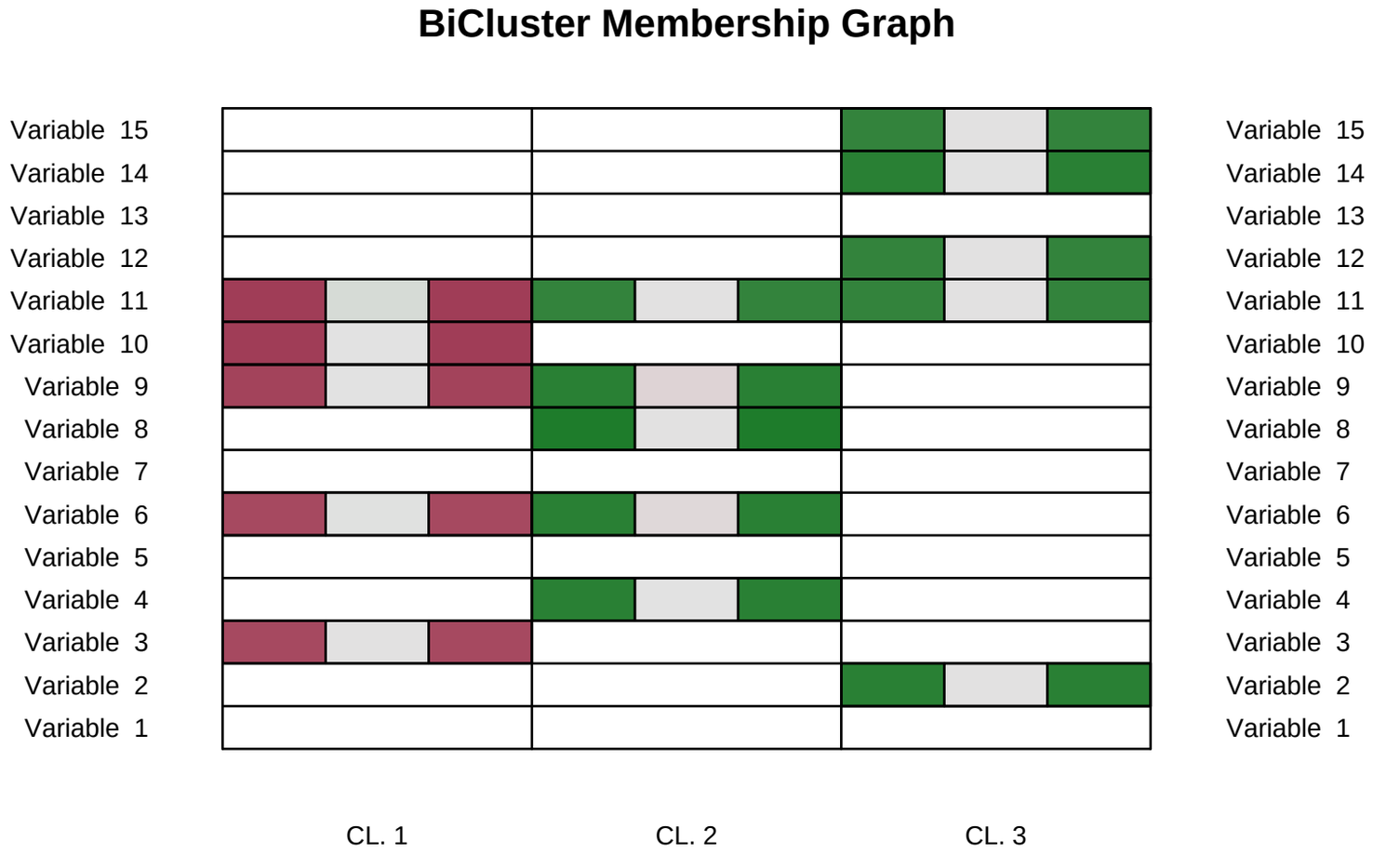






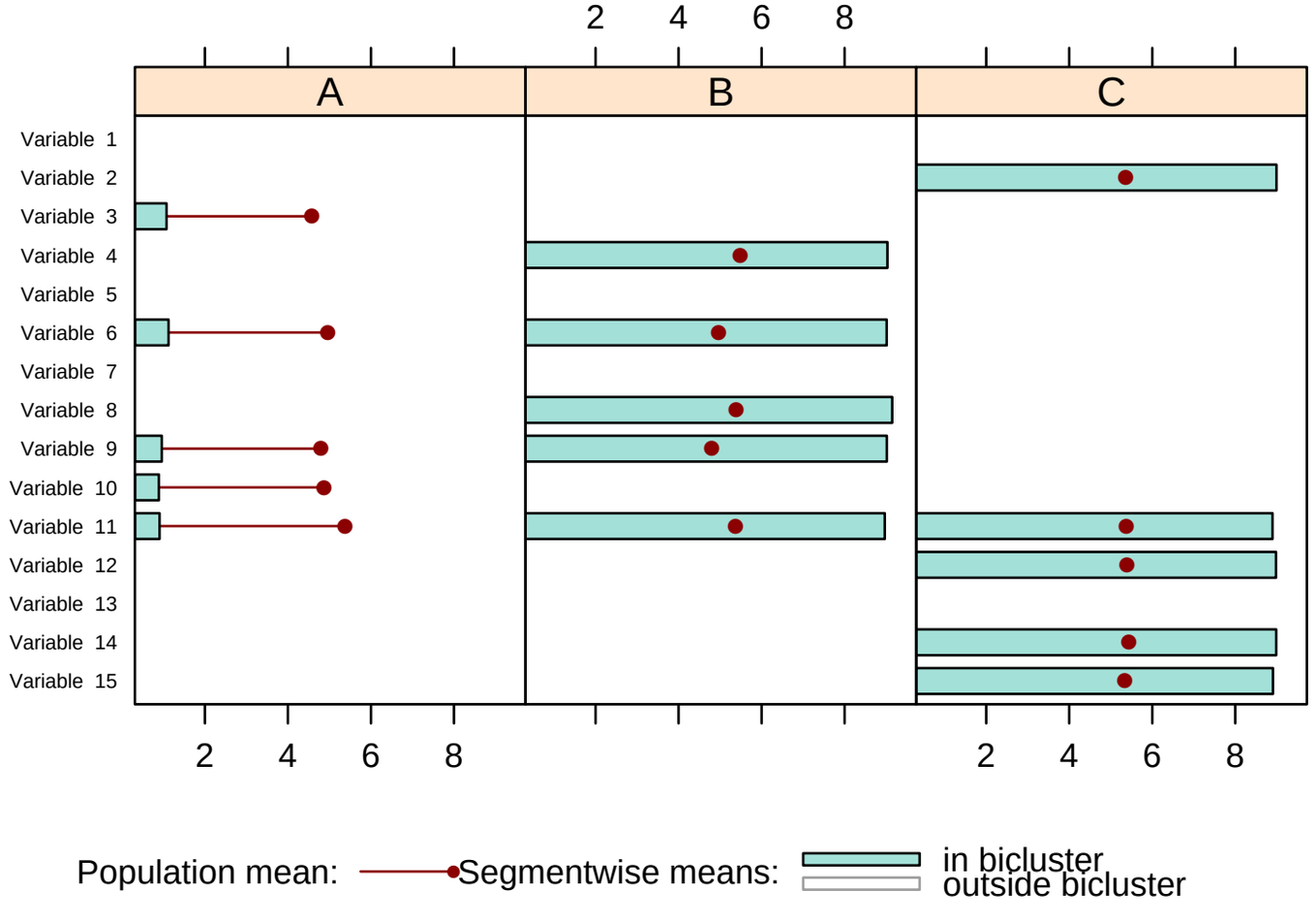
III. The biclust - Package: biclustmember()

```
biclustmember(Biclust,data,number,...)
```



III. The biclust - Package: biclustbarchart()

```
barchart(Biclust,data,number,...)
```



IV. Example: Tourism Survey

Australian Tourism Survey

- Survey conducted by researchers from the Faculty of Commerce, University of Wollongong
- Data collected from a nationally representative online Internet panel
- Questions about travel and unpaid help behavior
- 1003 people, 56 blocks of question à about 5 to 51 questions (around 600 questions)

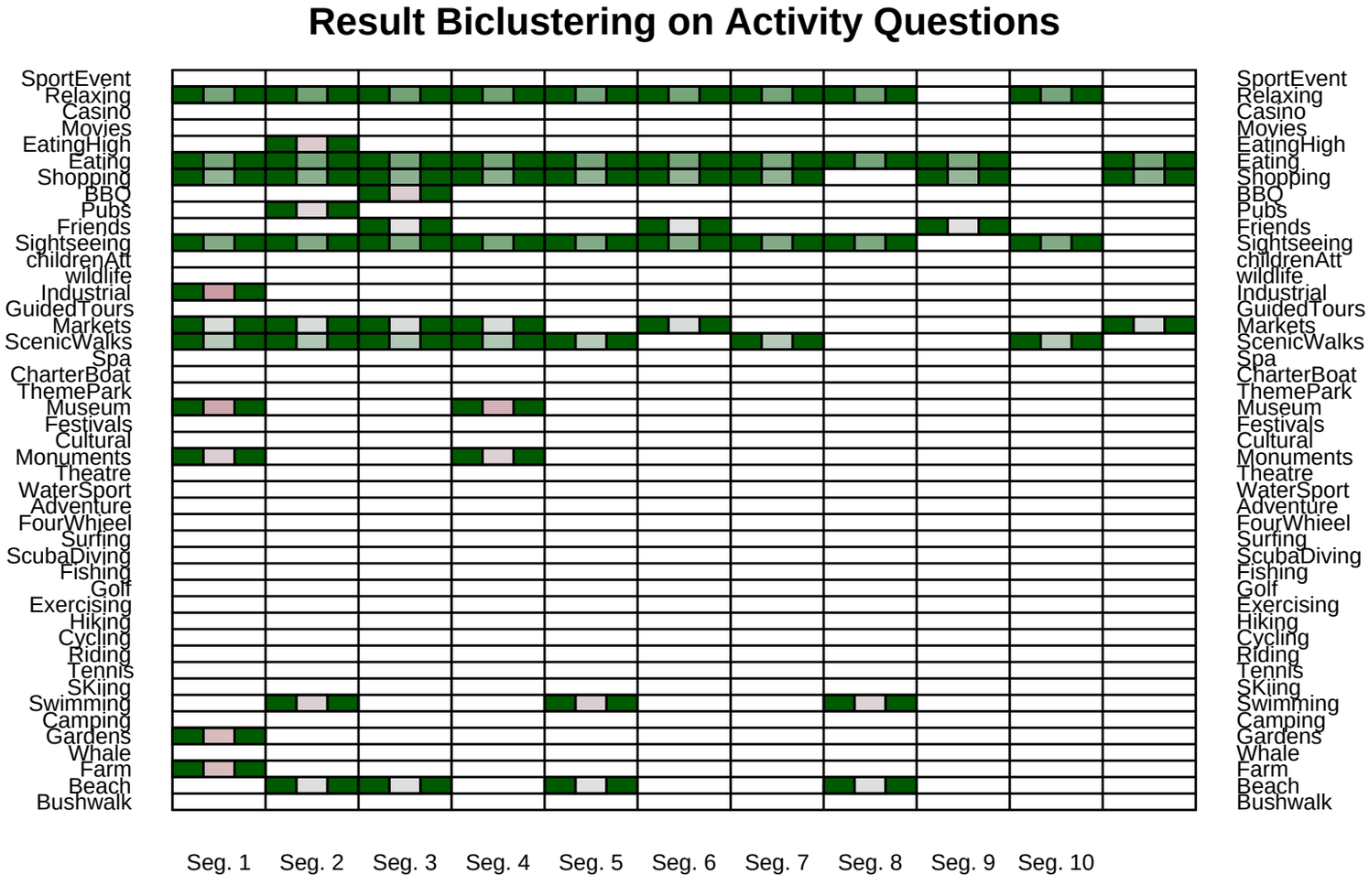
IV. Example: Tourism Survey I

Activity questions: Questions on activities participants did during their vacation.

```
> bimaxres<-biclust(x=activity, method=BCBimax(), number=50,
+ mrow=50, mcol=4)
> bimaxres

        An object of class Biclust

        call:
        biclust(x=activity, method=BCBimax(), number=50, mrow=50, mcol=4)

        Number of Clusters found:  11

        First  5  Cluster sizes:
                            BC 1 BC 2 BC 3 BC 4 BC 5
        Number of Rows:     "74" "59" "55" "50" "75"
        Number of Columns: "11" "10" " 9" " 8" " 7"
```

IV. Example: Tourism Survey I

```
biclustmember(res=bimaxres,data=activity,number=1,...)
```



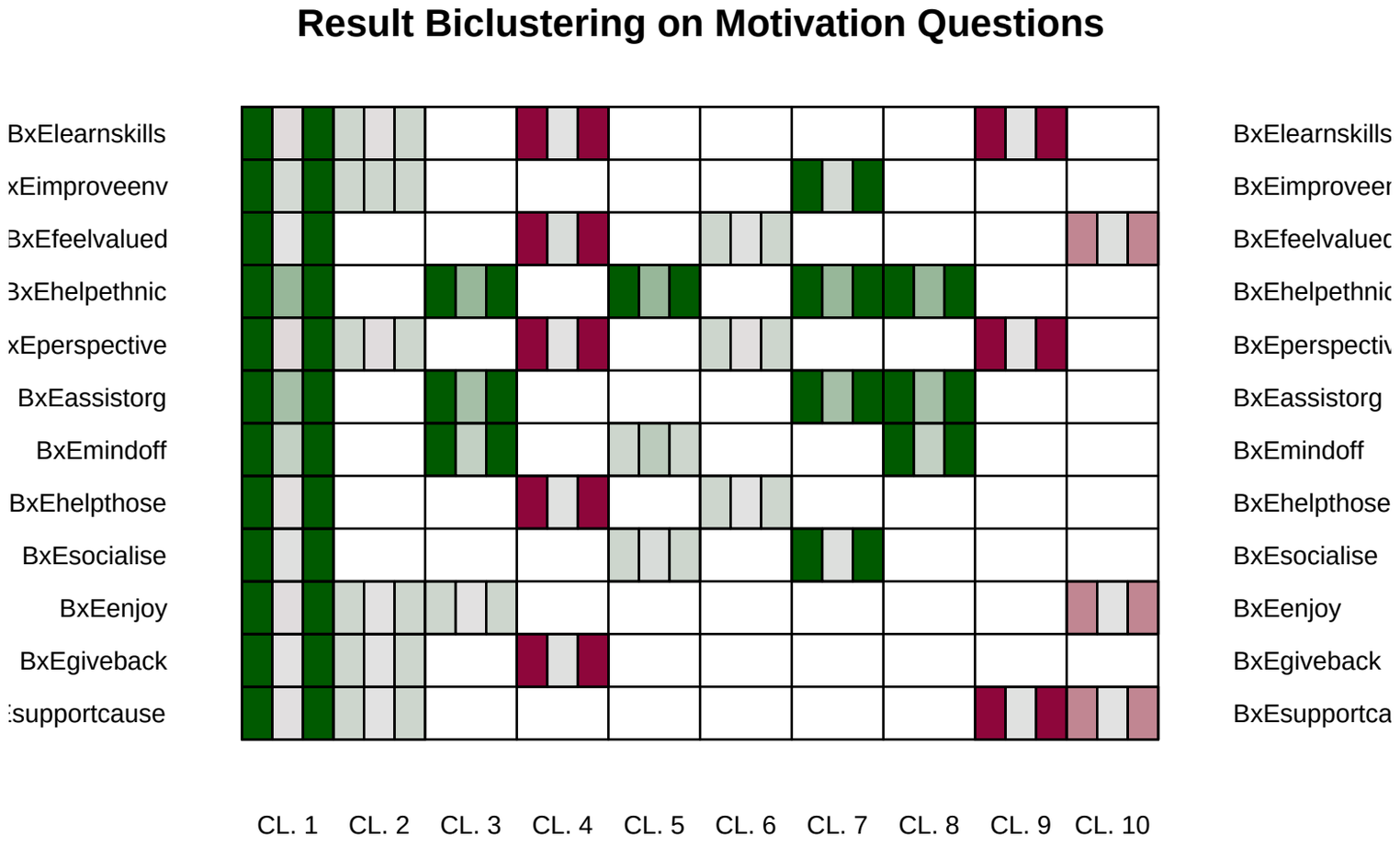
IV. Example: Tourism Survey I

Motivation questions: Questions on motivations for unpaid help weighted with importance.

```
> questres<-biclust(x=motivation, method=BCQuestord(), d=2, ns = 500,
+ nd = 500, sd = 1, alpha = 0.05, number = 10)
> questres
    An object of class Biclust

    call:
            biclust(x = motivation, method = BCQuestord(), ns = 500,
            nd = 500, sd = 1, alpha = 0.05, number = 10)

    Number of Clusters found:  10

    First  5  Cluster sizes:
                    BC 1 BC 2 BC 3 BC 4 BC 5
    Number of Rows:    "76" "69" "77" "59" "57"
    Number of Columns: "12" " 6" " 4" " 5" " 3"
```

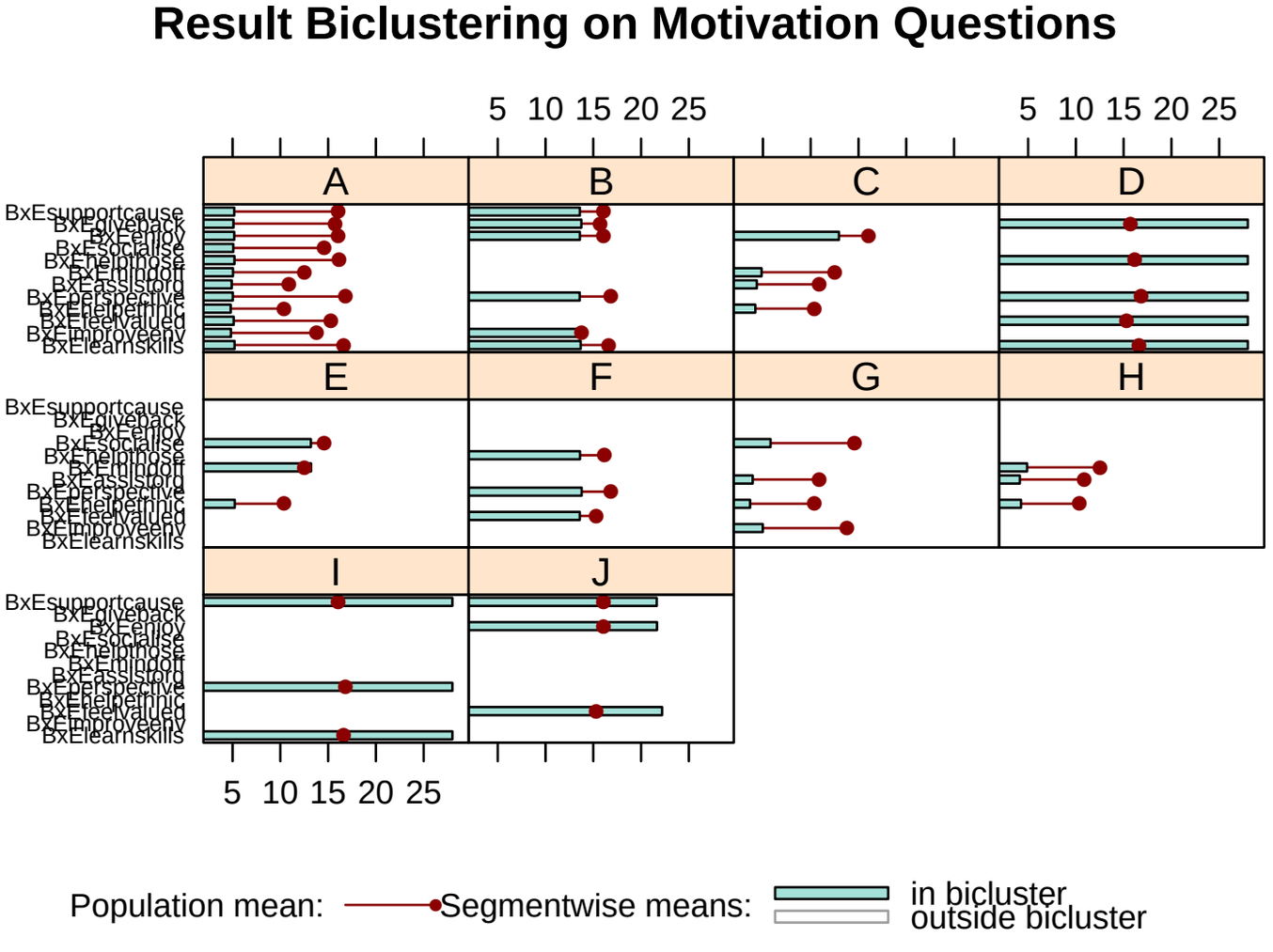
IV. Example: Tourism Survey II

```
biclustmember(res=questres,data=motivation,number=1,...)
```



IV. Example: Tourism Survey II

```
barchart(res=questres,data=motivation,number=1,...)
```



V. Summary and Future Work

Summary

- New bicluster algorithm to deal with nominal, ordinal and continuous data
- New developments in the `biclust` package
- Example on tourism data

Future Work

- Simultaneous clustering of nominal, ordinal and continuous data (Questionaire)
- Fully model based biclustering

Acknowledgments

Market segmentation is a joint work with Sara Dolnicar from the School of Management and Marketing of the University of Wollongong in Australia.

The package `biclust` is a joint work with Microarray Analysis and Visualization Effort, University of Salamanca, Spain, especially Rodrigo Santamaria.

References

biclust - A Toolbox for Bicluster Analysis in R,

Kaiser S. and Leisch F., In Paula Brito, editor, *Compstat 2008–Proceedings in Computational Statistics*, pages 201-208. Physica Verlag, Heidelberg, Germany.

BICLUSTERING: Overcoming data dimensionality problems in market segmentation,

Dolnicar S., Kaiser S., Lazarevski K., Leisch F., submitted 2009.

Links:

`http://cran.r-project.org/package=biclust/` official release

`http://r-forge.r-project.org/projects/biclust/` newest developments

`http://www.statistik.lmu.de/~kaiser/bicluster.html` Papers and Links