CluMix: Clustering and Visualization of Mixed-Type Data

Manuela Hummel   Annette Kopp-Schneider

June 2, 2016

Contents

1 Introduction

In real data situations various factors of interest are measured on different scales, e.g. quantitative gene expression values and categorical clinical features like gender, disease stage etc. In many cases (pre-selected) gene expression data are visualized in heatmaps, while further patient characteristics are only added ”informatively” on top. This can be visually quite confusing in case there are more than just a few such additional features. Also, it might be of interest to include clinical information in the process of clustering patients. Further, by standard heatmaps relationships between the quantitative features used for clustering and the information added on top are not explored explicitely. This package offers an integrative heatmap for data of mixed types to overcome those limitations of classical heatmaps.

In order to create a heatmap for variables measured on different scales, special similarity measures are necessary defining i) distances between subjects (e.g. patients) based on features of different types, and ii) distances between the different variables. Similarities between subjects are measured by Gower’s general similarity coefficient [2] with an extension of Podani [4] for ordinal variables. Similarities between variables are assessed by combination of appropriate measures of association for different pairs of data types [3]. Then standard hierarchical clustering with complete linkage is applied. Alternatively, variables can also be clustered by the ’ClustOfVar’ approach [1].

2 Mixed-Data Heatmap

We use a small simulated example dataset with quantitative, ordinal and categorical variables, that is included in the package for illustration.

```
> library(CluMix)
> data(mixdata)
> str(mixdata)
```

```
'data.frame':        40 obs. of  10 variables:
 $ X1.cat    : Factor w/ 3 levels "1","2","3": 3 1 3 2 3 2 2 3 1 3 ...
 $ X2.quant : num  0.465 -1.095 -0.699 4.208 1.394 ...
 $ X3.quant : num  -1.4703 1.3512 -0.0678 4.6036 -0.2339 ...
 $ X4.ord    : Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 3 1 1 4 1 4 4 1 3 2 ...
 $ X5.ord    : Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 3 2 4 5 3 5 5 4 1 1 ...
 $ X6.quant : num  -3.47 -3.16 -3.51 -1.15 5.85 ...
 $ X7.quant : num  -4.2637 -4.5057 -3.4813 -0.0388 4.5707 ...
 $ X8.cat    : Factor w/ 3 levels "1","2","3": 1 1 1 3 2 3 3 2 3 2 ...
 $ X9.ord    : Ord.factor w/ 5 levels "1"<"2"<"3"<"4"<..: 2 2 1 3 4 3 3 5 1 5 ...
 $ X10.quant: num  1.478 0.407 1.716 0.355 -1.555 ...
```

The mixed-data heatmap with subjects in the columns and variables in the rows is created by the mix.heatmap function (see Figure 1). Some options are available to manipulate labels, colors and legend. Note that in the current implementation the heatmap is limited to 200 variables.

```
> mix.heatmap(mixdata, rowmar=7, legend.mat=TRUE)
```

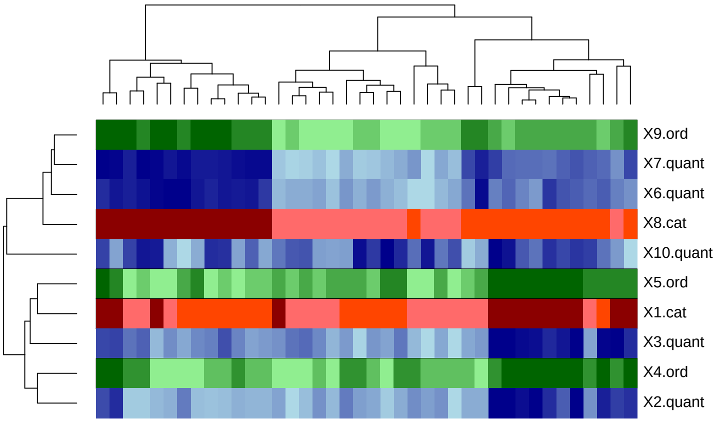

Figure 1: Mixed-data heatmap using Gower's distances for clustering subjects (columns) and combination of association measures (CluMix approach) for clustering variables (rows).

For clustering subjects, variable weights can be provided to give more importance to certain variables in the calculation of Gower's distances (see Figure 2).

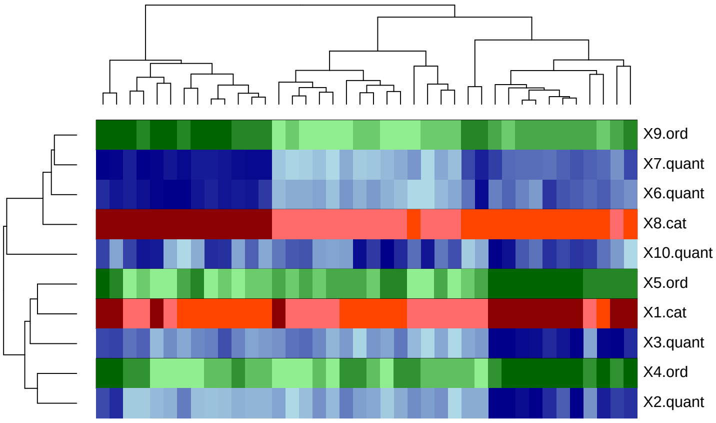

Figure 2: Mixed-data heatmap using weighted Gower's distances for clustering subjects (columns) and combination of association measures (CluMix approach) for clustering variables (rows).

```
> w <- rep(1:2, each=5) > mix.heatmap(mixdata, varweights=w, rowmar=7)
```

To choose the 'ClustOfVar' approach for clustering variables (see Figure 3) instead of the default approach using a combination of different association measures, you can specify *dist.variables.method = "ClustOfVar"*.

```
> mix.heatmap(mixdata, dist.variables.method="ClustOfVar", rowmar=7)
```

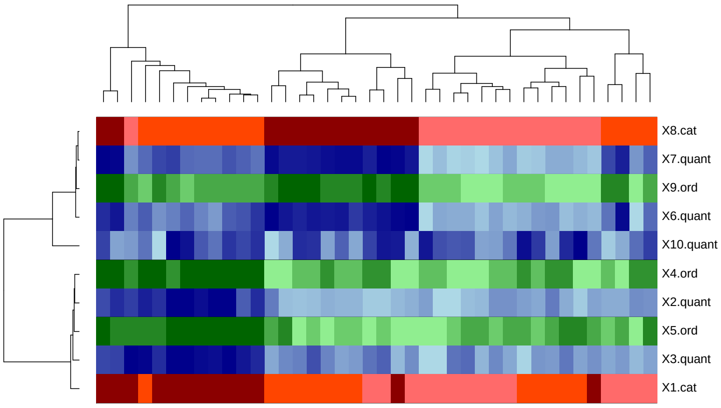

Figure 3: Mixed-data heatmap using the ClustOfVar approach for clustering variables.

The user can also provide previously calculated distance matrices or dendrograms (by functions `dist.subjects`, `dist.variables`, `dendro.subjects`, and `dendro.variables` from this package or anyhow).

```
> D.subjects <- dist.subjects(mixdata)
> dend.variables <- dendro.variables(mixdata)
> mix.heatmap(mixdata, D.subjects=D.subjects, dend.variables=dend.variables)
```

Colored bars can be added on top and to the left of the heatmap in order to provide additional information on subjects and/or variables. We give a random example, see Figure 4.

```
> colbar <- sample(c("purple", "darkgrey"), nrow(mixdata), replace=T)
> mix.heatmap(mixdata, ColSideColors=colbar, legend.colbar=c("aa",
"bb"), rowmar=7)
```

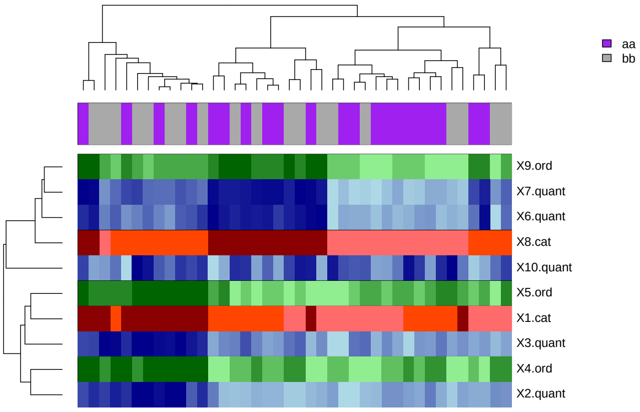

Figure 4: Mixed-data heatmap with added column color bar.

3 Similarity Matrix Heatmap

Instead of drawing a heatmap for both samples and variables simultaneously, one can also visualize a similarity matrix for either samples or variables, see Figure 5 for an example.

```
> distmap(mixdata, what="variables", margins=c(6,6))
```

Similarity matrices can also be derived before hand by `similarity.subjects` or `similarity.variables` (or anyhow), and provided to the `distmap` function as the *data* argument.

```
> S <- similarity.variables(mixdata)
> distmap(S)
```

4 Confounder Plot

We further propose an illustration that might be useful in regression analysis. The similarities of all variables in a dataset with two variables of special interest (i.e. predictor and outcome of a regression model) are simultaneously visualized in a scatter plot, where the x-axis shows similarities to the predictor and

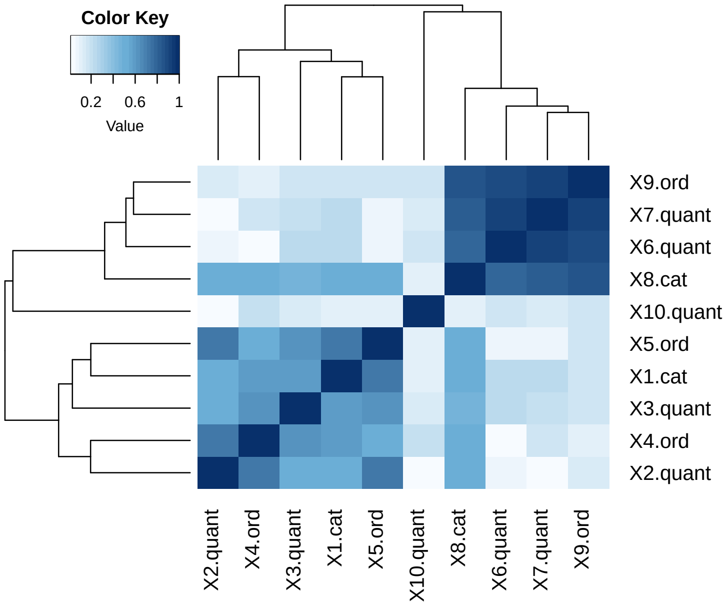

Figure 5: Similarity matrix heatmap for variables.

the y-axis similarities to the outcome, see Figure 6 for an example. The height of the predictor variable's point indicates its association with the outcome and hence its predicting ability. Variables in the upper right part are potential confounders for which prediction model should be adjusted, or collinear variables that should be removed. Variables in the lower right part are strongly related to the predictor, but not associated with the outcome. Variables very close to the outcome variable's point are potential surrogate outcomes. Note that distances between points in the plot do not directly correspond to variable similarities.

```
> confounderPlot(mixdata, x="X4.ord", y="X1.cat")
```

5 Session Information

```
> toLatex(sessionInfo())
```

- R Under development (unstable) (2016-06-01 r70695), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=German_Germany.1252`, `LC_MONETARY=German_Germany.1252`, `LC_NUMERIC=C`, `LC_TIME=German_Germany.1252`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: CluMix 1.1
- Loaded via a namespace (and not attached): Biobase 2.30.0, BiocGenerics 0.16.1, ClustOfVar 0.8, DescTools 0.99.15, FD 1.0-12,

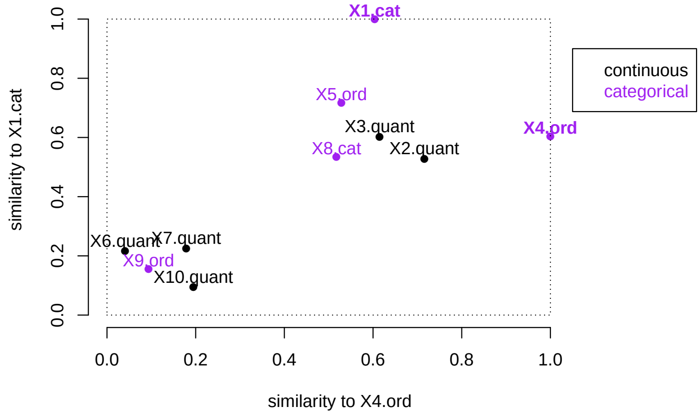

Figure 6: Similarity of each variable with 'X1.cat' (y-axis) plotted against respective similarities with 'X4.ord' (x-axis).

Formula 1.2-1, Hmisc 3.17-1, KernSmooth 2.23-15, MASS 7.3-45, Matrix 1.2-4, RColorBrewer 1.1-2, Rcpp 0.12.3, TSP 1.1-3, acepack 1.3-3.3, ade4 1.7-3, ape 3.4, bitops 1.0-6, boot 1.3-18, caTools 1.17.1, chron 2.3-47, cluster 2.0.3, codetools 0.2-14, colorspace 1.2-6, data.table 1.9.6, extracat 1.7-4, foreach 1.4.3, foreign 0.8-66, gdata 2.17.0, geometry 0.3-6, ggplot2 2.0.0, gplots 2.17.0, grid 3.4.0, gridExtra 2.2.1, gtable 0.1.2, gtools 3.5.0, hexbin 1.27.1, iterators 1.0.8, lattice 0.20-33, latticeExtra 0.6-26, limma 3.26.5, magic 1.5-6, magrittr 1.5, manipulate 0.98.507, marray 1.48.0, mgcv 1.8-11, munsell 0.4.2, mvtnorm 1.0-4, nlme 3.1-127, nnet 7.3-12, parallel 3.4.0, permute 0.9-0, plyr 1.8.3, reshape2 1.4.1, rpart 4.1-10, scales 0.3.0, splines 3.4.0, stringi 1.0-1, stringr 1.0.0, survival 2.38-3, tools 3.4.0, vegan 2.3-3

References

[1] Marie Chavent, Vanessa Kuentz-Simonet, Benoit Liquet, and Jerome Saracco. Clustofvar: An r package for the clustering of variables. *Journal of Statistical Software*, 50(1):1–16, 2012.

[2] J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27:857–871, 1971.

[3] Manuela Hummel and Annette Kopp-Schneider. Clustering of samples and variables with mixed-type data. *work in progress*.

[4] J. Podani. Extending gower's general coefficient of similarity to ordinal characters. *Taxon*, 48:331–340, 1999.