k prototypes Clustering
for Mixed-Type Data

Gero Szepannek

fachhochschule stralsund
university of applied sciences
fachbereich wirtschaft
school of business studies

Scope

- Gold standard for clustering in industry:
  - K-means
  - Hierarchical Clustering (x-Linkage & Ward)
  - SOMs
- …originally designed for numeric data.

- For factors only…
  - … latent class analysis can be used (e.g. using {poLCA}; Linzer and Lewis, 2014)
  - The k modes algorithm extends k means to factors (cf. {klaR}; Roever et al., 2014)

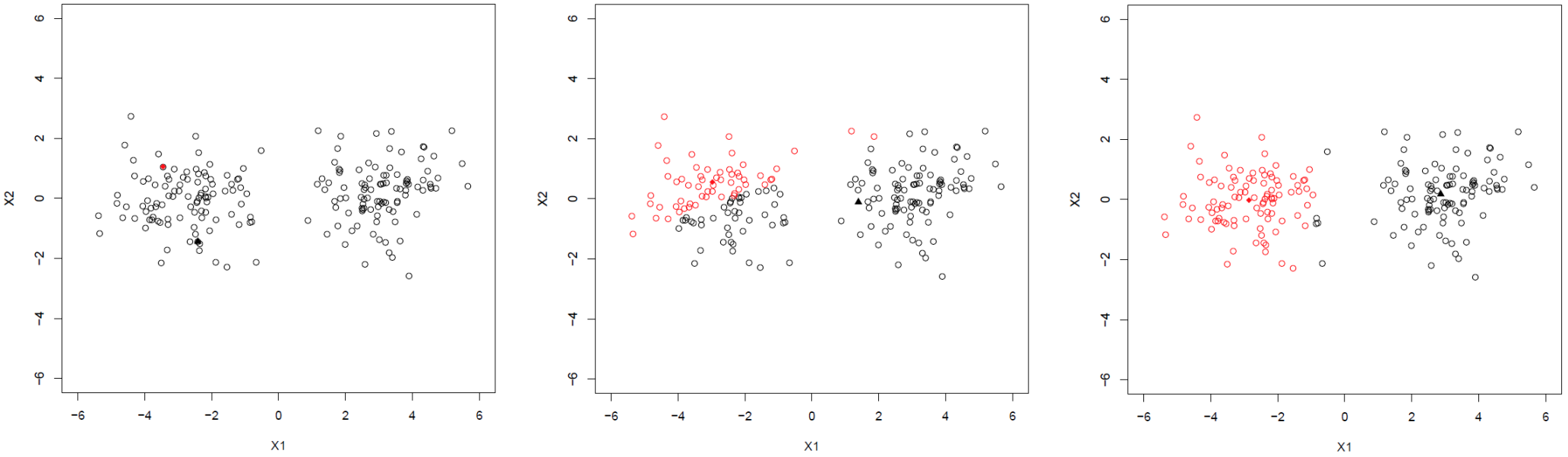
- But for most BI applications both are of interest: numerics and factors.
- Common practice:
  - Hierarchical Clustering using Gower distance (e.g. using {cluster}; Maechler et al., 2016)
  - Pre-Transform of factors into dummies (cf. Weihs and Szepannek, 2010, e.g. using {caret}; Kuhn, 2016)

- An alternative approach: k prototypes (Huang, 1998) generalizes k modes for mixed type data ({clustMixType}; Szepannek, 2016).

Gero Szepannek
Stralsund University of
Applied Sciences

Basic k Means Algorithm

1. Initialize k centers.
2. Assignment of each observation to its closest center.
3. Cluster update of center as means according to the new assigmnent.
4. Repeat 2. and 3. until convergence or maximum number of iterations.



Gero Szepannek
Stralsund University of
Applied Sciences

From k Means to k Prototypes

- For **numeric** variables the cluster mean is the prototype with minimal euclidean distance to all objects (cf. e.g. Hastie et al., 2009).
- Likewise for **factor** variables the mode minimizes simple matching distances, it is:

$$\sum_{i=1}^{n_k} I(x_i \neq x_{mode}) = min$$

(number of objects different from x_{mode})

- Both means and modes are calculated for each variable separately!

This allows do define a prototype update step in the mixed type setting.

Gero Szepannek
Stralsund University of
Applied Sciences

Cluster Assignment

- According to the former considerations a mixed-type distance is defined by:

$$d_{mix}(x,y) = \sum_{j=1}^{p} \left(x^j - y^j\right)^2 + \lambda \sum_{j=p+1}^{q} I(x^j \neq y^j)$$

- …where the upper index *j* denotes the variable (*p* numerics and *(q-p)* factors)
- The parameter **λ** allows to control the trade-off between numeric and factor variables in distance computation.

According to d_{mix} in each step clusters will be re-assigned.

Gero Szepannek
Stralsund University of
Applied Sciences

The R Package `clustMixType`

- The R package clustMixType (Szepannek, 2016) offers access to:
  - …a function `kproto()` to perform k prototype clustering
  - …a `predict()` method for application to new data,
  - …a function `clprofiles()` to visualize cluster profiles.

- Basic call: `kproto(x, k, lambda = NULL, iter.max = 100, nstart = 1)`

  creates an object of class kproto where
  - `x` is a data frame
  - `k` the number of clusters and (note that alternatively a vector of prototype indices can be passed)
  - `lambda` the control parameter (cf. last slide)
  - `iter.max` and `nstart` control termination of the algorithm and multiple calls (as the algorithm only locally converges it depends on initialization. The best solution of repetitve calls will be returned.)

Gero Szepannek
Stralsund University of
Applied Sciences

Example

Design:

- 4 clusters
- 4 variables (2 numeric, 2 factor)
- 100 observations | cluster

Cluster	Num1	Num2	Factor1	Factor2
1	+	+	+	+
2	-	-	+	+
3	+	+	-	-
4	-	-	-	-

Result: `kmeans` like object of class `kproto`

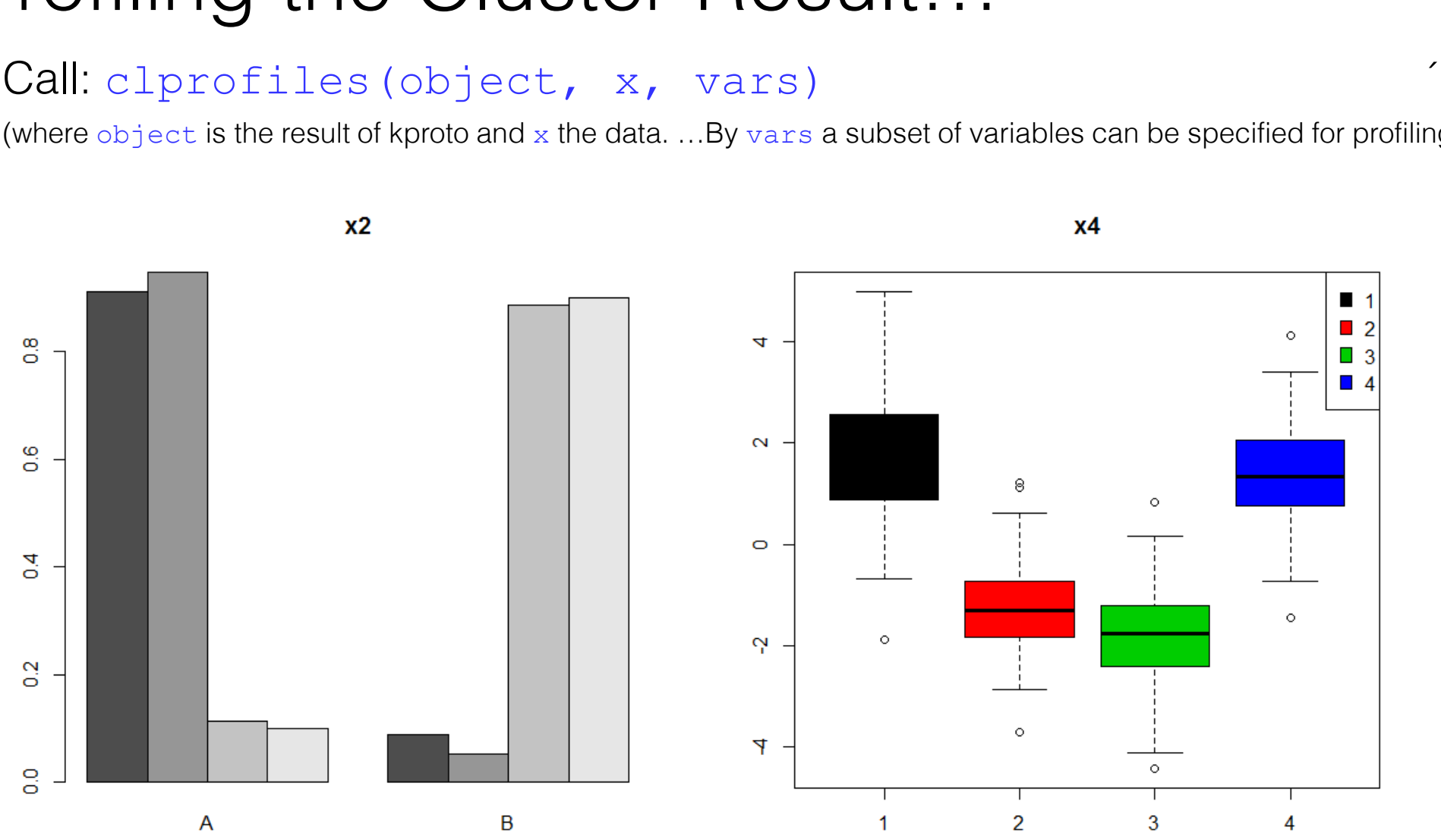
```
List of 9
 $ cluster     : int [1:400] 2 2 2 2 2 2 3 2 2 2 ...
 $ centers     :'data.frame':   4 obs. of  4 variables:
  ..$ x1: Factor w/ 2 levels "A","B": 1 1 2 2
  ..$ x2: Factor w/ 2 levels "A","B": 1 1 2 2
  ..$ x3: num [1:4] 1.65 -1.57 -1.67 1.19
  ..$ x4: num [1:4] 1.76 -1.27 -1.8 1.37
 $ lambda      : num 6.77
 $ size        : 'table' int [1:4(1d)] 101 93 106 100
  ..- attr(*, "dimnames")=List of 1
  .. ..$ clusters: chr [1:4] "1" "2" "3" "4"
 $ withinss    : num [1:4] 374 266 332 336
 $ tot.withinss: num 1309
 $ dists       : num [1:400, 1:4] 16.7 29.3 18.8 21.6 35.3 ...
 $ iter        : num 7
 $ trace       :List of 2
  ..$ tot.dists: num [1:7] 3720 1612 1350 1313 1310 ...
  ..$ moved    : int [1:7] 400 77 33 17 7 2 0
 - attr(*, "class")= chr "kproto"
```

```
Numeric predictors: 2
Categorical predictors: 2
Lambda: 6.768405

Number of Clusters: 4
Cluster sizes: 101 93 106 100
Within cluster error: 374.4996 266.4945 332.0333 336.3587

Cluster prototypes:
    x1 x2        x3        x4
92   A  A  1.646250  1.758957
54   A  A -1.567456 -1.269453
205  B  B -1.666238 -1.796813
272  B  B  1.189727  1.365156
```

Gero Szepannek
Stralsund University of
Applied Sciences



Profiling the Cluster Result…

- Call: `clprofiles(object, x, vars)`

  (where `object` is the result of kproto and `x` the data. …By `vars` a subset of variables can be specified for profiling.)

Gero Szepannek
Stralsund University of
Applied Sciences

References

- Huang, Z. (1998): Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and KDD 2*, 283–304.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009): The Elements of Statistical Learning. Springer.
- Kuhn, M. (2016): caret: Classification and Regression Training, R package version 6.0-71, https://CRAN.R-project.org/package=caret.
- Linzer, D. and Lewis, J.. (2014): poLCA: Polytomous variable Latent Class Analysis, R package version 1.4-1, https://CRAN.R-project.org/package=poLCA.
- Maechler, M., Rousseeuw, P., Struyf, A. Hubert, M. (2016): cluster: "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al, R package version 2.1-4, https://CRAN.R-project.org/package=cluster.
- Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G. and Zentgraf, M. (2014): klaR: Classification and visualization, R package version 0.6-12, https://CRAN.R-project.org/package=klaR.
- Szepannek, G. (2016): clustMixType: k-Prototypes Clustering for Mixed Variable-Type Data, R package version 0.1-16, https://CRAN.R-project.org/package= clustMixType.
- Weihs, C. and Szepannek, G. (2009); Distances in Classification, *Transactions in Case-Based Reasoning 2*, 3-14.

Gero Szepannek
Stralsund University of
Applied Sciences

Thank you!

Gero Szepannek
Stralsund University of
Applied Sciences