A full factorial benchmarking study of non-parametric partitioning methods for mixed-type data

Imperial College London

Efthymios Costa

Supervised by: Dr. Ioanna Papatsouma & Prof. Alastair Young

Motivation & Aims

- Clustering: the task of assigning data points into a number of groups/clusters such that data points within each cluster are more similar to each other than to points in other groups.
- Mixed data sets are often encountered and performing meaningful cluster analysis is crucial for practitioners.
- Benchmarking studies could serve as a guide to help with the choice of clustering technique but these need to disentangle possible interactions between the various data set characteristics. [1]

Non-Parametric Methods

Dissimilarities between data objects are defined by distance functions:

- **K-Prototypes** [2]:

$$d(X_i, Q_l) = \sum_{j=1}^{p_r}(x_{ij} - q_{lj})^2 + \gamma_l \sum_{j=p_r+1}^{p} \delta(x_{ij}, q_{lj}).$$

- **Gower's dissimilarity** [3]:

$$d_G(X_i, X_j) = 1 - \frac{\sum_{k=1}^{p} w_k(X_i,X_j) s_k(X_i,X_j)}{\sum_{k=1}^{p} w_k(X_i,X_j)}$$

- **Mixed K-Means** [4]: $d_M(X_i, Q_l) = \sum_{j=1}^{p_r}(w_j(x_{ij} - q_{lj}))^2 + \sum_{j=p_r+1}^{p} \Omega(x_{ij}, q_{lj})^2$
- **Modha-Spangler K-Means** [5]: $d_{MS}(X_i, Q_l) =$

$$\sum_{j=1}^{p_r}(x_{ij} - q_{lj})^2 + \gamma_l \left(1 - \frac{\sum_{j=p_r+1}^{P^*} x_{ij} q_{lj}}{\sqrt{\sum_{j=p_r+1}^{P^*} x_{ij}^2}\sqrt{\sum_{j=p_r+1}^{P^*} q_{lj}^2}}\right)$$

$$E = \sum_{l=1}^{k}\sum_{i=1}^{n} y_{il}\, d(X_i, Q_l) \qquad (1)$$

(1) is the 'trace of the within cluster dispersion matrix' cost function that we want to minimise.

Factor Analysis Techniques

Motivation

- Data sets can consist of a very large number of columns (variables), some of which may be irrelevant to the existing cluster structure.
- Dimensionality reduction techniques can be particularly helpful in such cases.
- How can they be achieved for both continuous & categorical data?

Methods Considered:

- **Factor Analysis for Mixed Data** [6]:
  - Sequential dimensionality reduction and clustering method.
  - The i^{th} principal component is given by:

$$\boldsymbol{F}_i^* = \underset{\boldsymbol{F}_i \perp\!\!\!\perp \boldsymbol{F}_{i-1},\dots,\boldsymbol{F}_1}{\arg\max} \sum_{j=1}^{p_r} R^2\left(\boldsymbol{F}_i, \boldsymbol{X}_{con_j}\right) + \sum_{j=p_r+1}^{p} \eta^2\left(\boldsymbol{F}_i, \boldsymbol{X}_{cat_j}\right).$$

  - K-Means is applied on the lower dimensional representation.
- **Mixed Reduced K-Means** [7]:
  - Joint dimensionality reduction and clustering method.
  - The 'optimal' cluster allocation is given by:

$$\boldsymbol{Z}_k^* = \underset{\boldsymbol{Z}_k}{\arg\min}\, \phi_{RKM}\left(\boldsymbol{B}, \boldsymbol{Z}_k, \boldsymbol{G}\right) = \underset{\boldsymbol{Z}_k}{\arg\min}\, \|\boldsymbol{X} - \boldsymbol{Z}_k \boldsymbol{G} \boldsymbol{B}^\intercal\|_F^2$$

  - Minimisation via an alternating least squares algorithm.

Experimental Design & Results

Experimental Design

- Aspects Investigated:
  - Number of observations (300, 600, 1200)
  - Number of variables (6, 10, 12)
  - Number of clusters (3, 4, 5)
  - Cluster sphericity (Spherical/Non-Spherical)
  - Average cluster overlap: $\omega_{ij} = \omega_{i|j} + \omega_{j|i}$, where $\omega_{i|j} = \mathbb{P}_{\boldsymbol{X}}\left(\pi_j \phi\left(\boldsymbol{X}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) < \pi_i \phi\left(\boldsymbol{X}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) | \boldsymbol{X} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)\right)$ [8] (0.01, 0.05, 0.10, 0.15, 0.20)
  - Cluster density, i.e. whether clusters are balanced (Balanced/Highly Unbalanced)
- Data sets simulated from Gaussian mixtures, half of the variables discretised by quantile discretisation.
- Cluster recovery performance evaluated using the Adjusted Rand Index (ARI) [9].

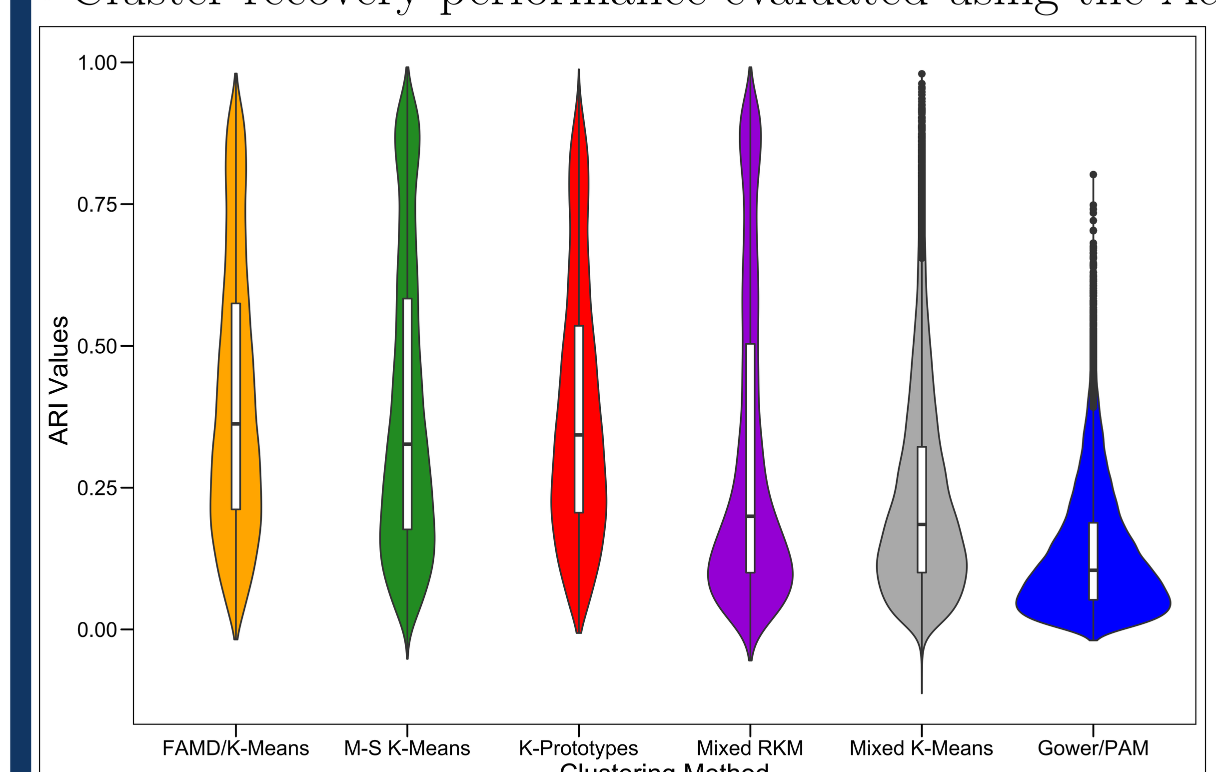
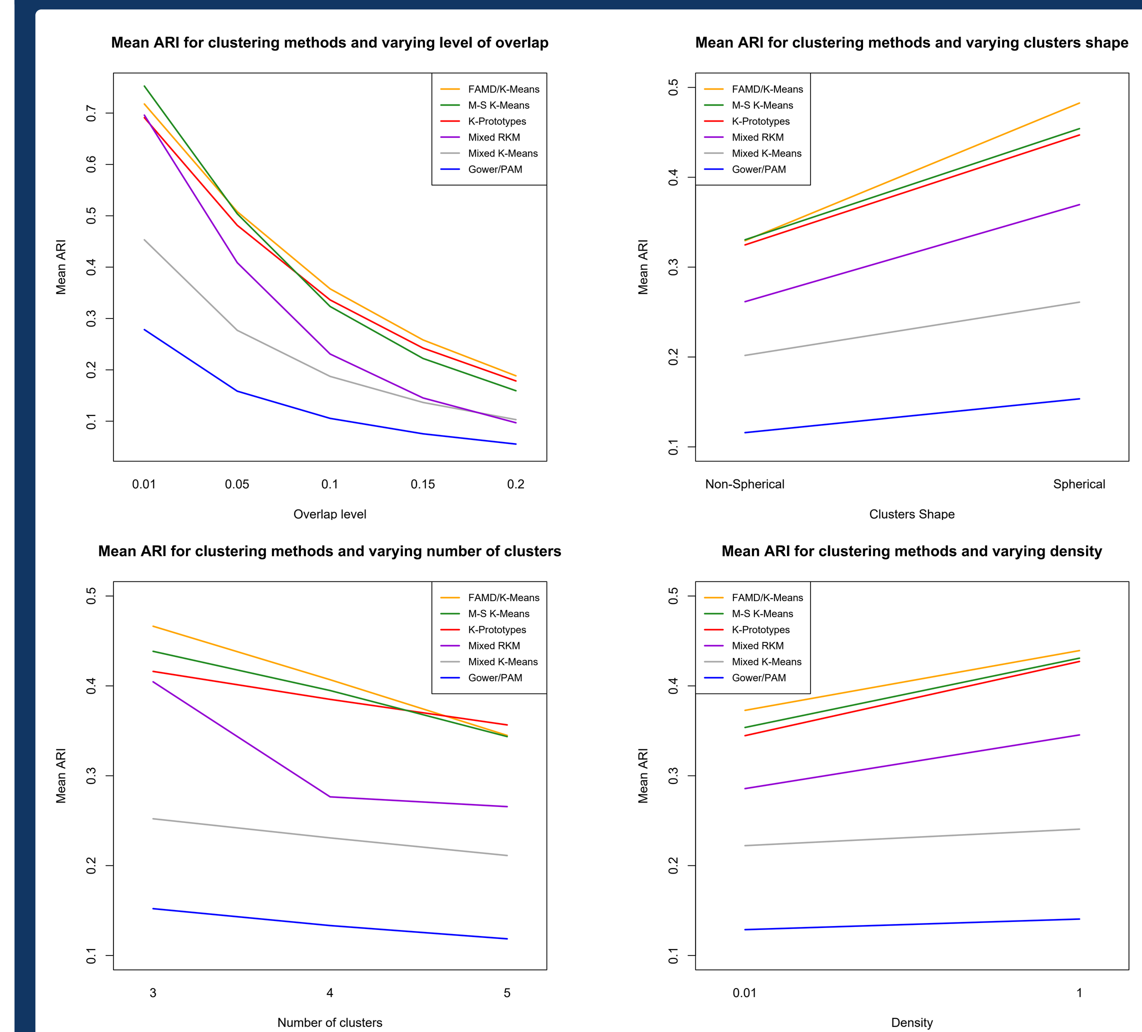


Figure: Violin/box plots of Adjusted Rand Index values by method

Effect	Source	partial η^2
Between data sets effects	overlap	.804
	shape	.268
	# clusters	.140
	density	.091
	# vars	.012
	# obs	.006
Within data sets effects (univariate tests)	Method (M)	.666
	M*overlap	.501
	M*vars	.206
	M*clusters	.153
	M*density	.093
	M*shape	.024
	M*obs	.002

Table: Repeated measures ANOVA for six clustering methods on ARI (factors ordered by decreasing effect size, partial η^2)

Additional Results Plots



Future Work Plans

- Investigate the effect of the ratio of categorical to continuous variables in clustering performance.
- Generate purely mixed-type data, i.e. purely categorical variables and purely continuous variables with a cluster structure.
- Look at high-dimensional data ($n \ll p$) and conduct a similar study.

References

[1] I. Van Mechelen, A.-L. Boulesteix, R. Dangl, N. Dean, I. Guyon, C. Hennig, F. Leisch, and D. Steinley, "Benchmarking in cluster analysis: A white paper," *arXiv preprint arXiv:1809.10496*, 2018.

[2] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proceedings of the 1st Pacific-Asia conference on knowledge discovery and data mining (PAKDD)*, pp. 21–34, Citeseer, 1997.

[3] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, pp. 857–871, 1971.

[4] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2007.

[5] D. S. Modha and W. S. Spangler, "Feature weighting in k-means clustering," *Machine learning*, vol. 52, no. 3, pp. 217–237, 2003.

[6] J. Pagès, *Multiple Factor Analysis By Example Using R*, ch. 3, pp. 67–78. Chapman and Hall/CRC, 2014.

[7] M. Vichi, D. Vicari, and H. A. Kiers, "Clustering and dimension reduction for mixed variables," *Behaviormetrika*, vol. 46, no. 2, pp. 243–269, 2019.

[8] R. Maitra and V. Melnykov, "Simulating data to study performance of finite mixture modeling and clustering algorithms," *Journal Of Computational And Graphical Statistics*, vol. 19, no. 2, pp. 354–376, 2010.

[9] L. Hubert and P. Arabie, "Comparing partitions," *Journal Of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

Notation

k: number of clusters, n: number of data points, p: number of variables, p_r: number of continuous variables, P^*: number of continuous & dummy-coded categorical variables, X_i: i^{th} data point, Q_l: prototype/centroid/medoid for l^{th} cluster, $\|\cdot\|_F$: Frobenius norm, y_{il}: $= 1 \iff X_i$ is in l^{th} cluster (else 0), $\boldsymbol{B}$: columnwise orthonormal loadings matrix, $\boldsymbol{G}$: cluster centroids in reduced dimensions, $\boldsymbol{Z}_k$: cluster allocations matrix