STABILITY OF JOINT DIMENSION REDUCTION CLUSTERING

Angelos Markos[1], Michel van de Velden[2] and Alfonso Iodice D'Enza[3]

[1] Department of Primary Education, Democritus University of Thrace, (e-mail: amarkos@eled.duth.gr)

[2] Department of Econometrics, Erasmus University Rotterdam, (e-mail: vandevelden@ese.eur.nl)

[3] Department of Political Sciences, University of Naples, Federico II, (e-mail: iodicede@unina.it)

ABSTRACT: Several methods for joint dimension reduction and cluster analysis of categorical, continuous or mixed-type data have been proposed over time. These methods combine dimension reduction (PCA/MCA/PCAmix) with partitioning clustering (K-means) by optimizing a single objective function. Cluster stability assessment is a critical and inadequately discussed topic in the context of joint dimension reduction and clustering. We introduce a resampling scheme that combines bootstrapping and a measure of cluster agreement to assess global cluster stability of joint dimension reduction and clustering solutions and a Jaccard similarity approach for empirical evaluation of the stability of individual clusters. Both approaches are implemented in the R package `clustrd`.

KEYWORDS: dimension reduction, k-means, cluster stability, cluster validity.

1 Joint dimension reduction and clustering

Joint dimension reduction refers to a set of algorithmic or non-model based techniques aimining at simultaneously finding an optimal reduction of the variables and an optimal partitioning of the objects of a rectangular data set. Reduced K-means (De Soete & Carroll, 1994) and Factorial K-means (Vichi & Kiers, 2001) combine Principal Component Analysis (PCA) for dimension reduction with K-means for clustering and are suitable for data sets with continuous variables. In the case of categorical variables, MCA K-means (Hwang, Dillon & Takane, 2006), IFC-B (Iodice D'Enza & Palumbo, 2013) and Cluster Correspondence Analysis (van de Velden, Iodice D'Enza & Palumbo, 2017) a variant of Correspondence Analysis is used in the dimension reduction step and K-means is used for clustering. In the case of mixed-type data, that is when

the data set contains both continuous and categorical variables, one can resort to GROUPALS (Van Buuren & Heiser, 1989) and Mixed Reduced/Factorial K-means (Vichi, Vicari & Kiers, 2009). These methods combine PCA for mixed data with K-means.

The general objective can be formulated as follows:

$$\min \phi_{\text{CDR}}(\mathbf{B}, \mathbf{Z}_K) = \alpha \left\| \mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}' \right\|^2 + (1-\alpha) \left\| \mathbf{X}\mathbf{B} - \mathbf{P}\mathbf{X}\mathbf{B} \right\|^2 \qquad (1)$$

where $\mathbf{X}$ is a $n \times Q$ data matrix, $\mathbf{B}$ a $Q \times d$ columnwise orthonormal loadings matrix, d is the user supplied dimensionality of the reduced space, $\mathbf{Z}_K$ a $n \times K$ binary matrix indicating cluster memberships of the n observations into the K clusters, $\mathbf{P} = \mathbf{Z}_K \left(\mathbf{Z}_K' \mathbf{Z}_K\right)^{-1} \mathbf{Z}_K'$ is a projection matrix, and $\mathbf{G} = \mathbf{P}\mathbf{X}\mathbf{B}$ a $K \times d$ cluster centroid matrix.

For categorical variables, the CDR objective can easily be adjusted by substituting $\mathbf{D}_z^{-1/2}\mathbf{M}\mathbf{Z}$ for $\mathbf{X}$ in all equations. Similarly, for mixed-type data, $\mathbf{X}$ is set to $\begin{pmatrix} \mathbf{X} & \mathbf{D}_z^{-1/2}\mathbf{M}\mathbf{Z} \end{pmatrix}$.

For given α, the following alternating least-squares algorithm is used to minimize the loss function in Eq.1:

1. Generate an initial cluster allocation $\mathbf{Z}_K$ (e.g., by randomly assigning subjects to clusters).
2. Find loadings $\mathbf{B}$ by taking the eigendecomposition of $\mathbf{X}^{*\prime}\left((1-\alpha)\mathbf{P} - (1-2\alpha)\mathbf{I}\right)\mathbf{X}$.
3. Update the cluster allocation $\mathbf{Z}_K$ by applying K-means to the reduced space subject coordinates $\mathbf{X}\mathbf{B}$.
4. Repeat the procedure (i.e., go back to step 2) using $\mathbf{Z}_K$ for the cluster allocation matrix, until convergence. That is, until $\mathbf{Z}_K$ remains constant.

Note that, for $\alpha = 1$ CDR reduces to PCAMIX, for $\alpha = 1/2$ we get mixed RKM method and for $\alpha = 0$ we have mixed FKM.

2 Global and local cluster stability via resampling

Cluster validation is important because cluster analysis presents clusters in almost any case. Here we focus on the stability of a partition in the case of joint dimension and clustering, that is, given a new sample from the same population, how likely is it to obtain a similar clustering? Stability can also be used to inform the selection of the number of clusters because if true clusters exist, the corresponding partition should have a high stability.

Resampling approaches (that is, bootstrapping, subsetting, replacement of points by noise) provide an elegant framework to computationally derive the distribution of interesting quantities describing the quality of a partition (Hennig 2007, Dolnicar & Leisch 2010). Simulations so far seem to suggest that resampling makes a lot of difference; the exact scheme used is not that important. Leisch (2015) provides a generic scheme for assessing cluster stability via resampling. Based on this scheme, we provide below two algorithms, one for assessing global stability, or the overall stability of a clustering partition, and one for assessing local or cluster-wise stability, or the stability of each one of the clusters in a given partition.

Algorithm GLOBAL STABILITY
Resampling: Draw bootstrap samples $\mathcal{S}^i$ and $\mathcal{T}^i$ of size n from the data and use the original data as evaluation set $\mathcal{E}^i = \mathbf{X}$. Apply a joint dimension reduction and clustering method to $\mathcal{S}^i$ and $\mathcal{T}^i$ and obtain $C^{\mathcal{S},i}$ and $C^{\mathcal{T},i}$.
Mapping: Assign each observation x_i to the closest centers of $C^{\mathcal{S},i}$ and $C^{\mathcal{T},i}$ using Euclidean distance, resulting in partitions $C^{X\mathcal{S},i}$ and $C^{X\mathcal{T},i}$, where $C^{X\mathcal{S},i}$ is the partition of the original data $\mathbf{X}$ predicted from clustering bootstrap sample $\mathcal{S}^i$ (same for $\mathcal{T}^i$ and $C^{X\mathcal{T},i}$).
Evaluation: Use the Adjusted Rand Index (ARI, Hubert & Arabie, 1985) or the Measure of Concordance (MOC, Pfitzner 2008) as measure of agreement and stability.
Inspect the distributions of ARI/MOC to assess the *global reproducibility* of the clustering solutions.

Algorithm LOCAL STABILITY
Resampling: Draw bootstrap samples $\mathcal{S}^i$ and $\mathcal{T}^i$ of size n from the data and use the original data as evaluation set $\mathcal{E}^i = \mathbf{X}$. Apply a joint dimension reduction and clustering method to $\mathcal{S}^i$ and $\mathcal{T}^i$ and obtain $C^{\mathcal{S},i}$ and $C^{\mathcal{T},i}$.
Mapping: Assign each observation x_i to the closest centers of $C^{\mathcal{S},i}$ and $C^{\mathcal{T},i}$ using Euclidean distance, resulting in partitions $C^{X\mathcal{S},i}$ and $C^{X\mathcal{T},i}$.
Evaluation: Obtain the maximum Jaccard agreement between each original cluster C_k and each one of the two bootstrap clusters, $C_{k'}^{X\mathcal{S},i}$ and $C_{k'}^{X\mathcal{T},i}$ as measure of agreement and stability, and take the average of each pair:

$$s_k^i = \Big(\max_{i \le k' \le K} \frac{C_k \cap C_{k'}^{X\mathcal{S},i}}{C_k \cup C_{k'}^{X\mathcal{S},i}} + \max_{i \le k' \le K} \frac{C_k \cap C_{k'}^{X\mathcal{T},i}}{C_k \cup C_{k'}^{X\mathcal{T},i}} \Big) / 2$$

Inspect the distributions of s_k^i to assess the cluster level (local) stability of the solution.

The two algorithms are implemented in the `R` package `clustrd` via functions `global_bootclus()` and `local_bootclus()`, respectively.

3 Conclusions

Stability is an important aspect of clustering quality. Resampling approaches provide an elegant framework to assess global stability of Joint Dimension Reduction and Clustering solutions, as well as local quality of a cluster. However, maximizing stability for estimating the number of clusters amounts to implicitly defining the "true clustering" as the one with highest stability, which may not be appropriate. A comprehensive simulation study trying different combinations could offer guidance what works best in which situations.

References

DE SOETE, G., & CARROLL, J. D. 1994. K-means clustering in a low-dimensional Euclidean space. *New Approaches in Classification and Data Analysis*, 212–219.

DOLNICAR, S, & LEISCH, F. 2010. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Marketing Letters*, **21**(1), 83–101.

HENNIG, C. 2007. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, **52**(1), 258–271.

HWANG, H, DILLON W, & TAKANE, Y. 2006. An Extension of Multiple Correspondence Analysis for Identifying Heterogenous Subgroups of Respondents. *Psychometrika*, **71**, 161–171.

IODICE D'ENZA, A., & PALUMBO, F. 2013. Iterative factor clustering of binary data. *Computational Statistics*, **28**(2), 789–807.

LEISCH, F. 2015. Resampling methods for exploring cluster stability. *Pages 658–673 of: Handbook of cluster analysis*. Chapman and Hall/CRC.

VAN BUUREN, S, & HEISER, WJ. 1989. Clustering *n* objects into *k* groups under optimal scaling of variables. *Psychometrika*, **54**(4), 699–706.

VAN DE VELDEN, M, IODICE D'ENZA A, & PALUMBO, F. 2017. Cluster Correspondence Analysis. *Psychometrika*, **82**(1), 158–185.

VICHI, M, & KIERS, H. 2001. Factorial *k*-means analysis for two-way data. *Computational Statistics & Data Analysis*, **37**(1), 49–64.

VICHI, M, VICARI D, & KIERS, H. 2019. Clustering and dimensional reduction for mixed variables. *Unpublished manuscript, to appear in Behaviormetrika 2018*.