Received: 21 July 2018 | Revised: 20 September 2018 | Accepted: 22 September 2018

DOI: 10.1002/wics.1456

WILEY WIREs COMPUTATIONAL STATISTICS

ADVANCED REVIEW

Distance-based clustering of mixed data

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

[1]Department of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

[2]Department of Economics and Law, University of Cassino and Southern Lazio, Cassino FR, Italy

[3]Department of Primary Education, Democritus University of Thrace, Xanthi, Greece

Correspondence
Michel van de Velden, Department of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands.
Email: vandevelden@ese.eur.nl

Cluster analysis comprises of several unsupervised techniques aiming to identify a subgroup (cluster) structure underlying the observations of a data set. The desired cluster allocation is such that it assigns similar observations to the same subgroup. Depending on the field of application and on domain-specific requirements, different approaches exist that tackle the clustering problem. In distance-based clustering, a distance metric is used to determine the similarity between data objects. The distance metric can be used to cluster observations by considering the distances between objects directly or by considering distances between objects and cluster centroids (or some other cluster representative points). Most distance metrics, and hence the distance-based clustering methods, work either with continuous-only or categorical-only data. In applications, however, observations are often described by a combination of both continuous and categorical variables. Such data sets can be referred to as *mixed* or *mixed-type* data. In this review, we consider different methods for distance-based cluster analysis of mixed data. In particular, we distinguish three different streams that range from basic data preprocessing (where all variables are converted to the same scale), to the use of specific distance measures for mixed data, and finally to so-called joint data reduction (a combination of dimension reduction and clustering) methods specifically designed for mixed data.

This article is categorized under:
Statistical Learning and Exploratory Methods of the Data Sciences > Clustering and Classification
Statistical Learning and Exploratory Methods of the Data Sciences > Exploratory Data Analysis
Statistical and Graphical Methods of Data Analysis > Dimension Reduction

KEYWORDS

cluster analysis, dimension reduction, distance based methods, joint dimension reduction and clustering, mixed data

1 | INTRODUCTION

Cluster analysis aims at defining meaningful groups of observations (i.e., the clusters) in such a way that observations assigned to the same cluster are similar to each other. Many clustering procedures are so-called distance based, where the clusters are obtained by first defining an appropriate distance measure and then applying an algorithm that assigns observations being close to each other to the same cluster. Popular distance-based clustering methods are hierarchical clustering, where points are clustered sequentially by considering intercluster distances, and K-means clustering, where points are assigned in such a way that the squared distances of points to the cluster means (i.e., the within cluster variance) is minimized. An alternative approach to cluster analysis is to assume that parametric probability distibutions define "true" and homogeneous clusters, and that the observed data can be modeled as a mixture of such distributions. Such methods can be referred to as model-based clustering methods (Fraley & Raftery, 2002). In this review, however, we focus exclusively on distance-based clustering methods.

The selection of an appropriate distance or dissimilarity measure crucially affects the clustering solution and depends on the nature of the considered variables. Most distance measures (and hence corresponding clustering methods) concern the analysis of either continuous only or categorical only data. In practice, however, data sets are often comprised of both continuous and categorical variables. In this review, we concern ourselves with distance-based clustering methods that deal with this situation, which we refer to as mixed data.

The distance-based clustering methods of mixed data reviewed here are presented in a sequence of progressively enhanced approaches. We start by briefly summarizing the most straightforward approach where all the variables are converted to the same scale. That is, either continuous to categorical or vice versa. Next, we review various different distance measures for mixed data together with corresponding clustering methods. We then consider several, more advanced approaches that consists of a combination of dimension reduction and clustering, applied both sequentially and jointly. A selection of the reviewed methods is illustrated using three small examples.

2 | CONVERT ALL VARIABLES TO THE SAME SCALE

Since most clustering procedures are designed to deal with variables measured on the same or at least on commensurable scales, a simple strategy is to homogenize variables in a preprocessing phase. For mixed data this would mean either re-coding the continuous variables as categorical ones or vice versa.

3 | RECODE THE CONTINUOUS VARIABLES

Recoding of continuous variables can be achieved via discretization. For example, the range of a continuous variable is split into intervals and values are labeled according to the interval to which the value belongs. Of course, such kind of discretization implies a loss of information (Hennig, Meila, Murtagh, & Rocci, 2015: 187). Moreover, it is not trivial to choose a specific discretization. For some variables, it may be possible to use "meaningful" (general or application specific) intervals. For example, it may be customary in certain fields to use a specific division in age groups following generally accepted conventions (e.g., using age groups for age; nonadult, young adults, elderly etc.). In many cases such intervals are not present and the researchers may choose the intervals. However, the choice of discretization directly influences the calculation of distances and consequently the allocation to clusters. This ambiguity together with the complicated appraisal of cluster analysis solutions (Hennig, 2015) makes it difficult to motivate a certain choice of discretization.

An alternative transformation through discretization is to code the original values of a continuous variable into a prespecified number of *fuzzy* categories, that is, to a set of k nonnegative values that sum to 1, quantifying the "possibility" of the variable to be in each category (van Rijckevorsel (1988); Greenacre (2014); Aşan and Greenacre (2011)). These "pseudo-categorical" values represent each value of a continuous variable uniquely and exactly, that is, the numerical information of the original variable is preserved (Aşan & Greenacre, 2011). Rather than cut-points, h membership functions are used, for example, triangular membership functions or second-order B-splines (van Rijckevorsel, 1988). To define triangular membership functions, h "hinge points" or pivots are needed, for example, for $h = 3$ the hinge points could be the minimum value, the median and the maximum value. Alternative membership functions can be also used, like trapezoidal, Gaussian, and generalized bell-shaped (or Cauchy) membership functions, which have various other theoretical advantages (Aşan & Greenacre, 2011).

4 | RECODE THE CATEGORICAL VARIABLES

A general preprocessing and standardization approach for categorical variables is presented in (Mirkin, 2005: 85–91). Categorical variables are first dummy-coded, that is, for each category a separate column is created and observed categories are coded using ones, whereas all other objects are zeros. Then, each dummy variable is standardized by shifting the origin to the mean and dividing it by the standard deviation, the range or another quantity reflecting the variable's spread. In particular, the mean of a dummy variable corresponds to the proportion, $\hat{\pi}$, of objects falling in the corresponding category. The standard deviation can be either $\sqrt{\hat{\pi}(1-\hat{\pi})}$ (Bernoulli distribution) or $\sqrt{\hat{\pi}}$ (Poisson distribution). The range of a dummy variable is always 1. Distance-based clustering methods, such as K-means, are not invariant against affine transformations (Hennig et al., 2015: 716) and the choice of scaling option has an effect on clustering perfomance; an optimal choice that will apply across different data sets or clustering techniques is impossible (Foss, Markatou, & Ray, 2018).

Alternatively, there exists several methods that allow for a recoding of categorical data as continuous data (Gifi, 1990: 65–240). Typically, these methods use dimension reduction techniques to quantify the data. In multiple correspondence analysis, for example, quantifications for the categories are obtained that can be used to calculate coordinates (or scores) for the observations in a, user-specified, low-dimensional space. These coordinates best capture the variation, that is, the deviation from the independence condition, in the complete categorical data set.

Replacing the original categorical data by the low-dimensional coordinates yields a numerical data set that can be analyzed using clustering methods for continuous data. However, the quantified version of the categorical data is a low-dimensional approximation of the full dimensional data. Using dimension reduction to quantify the categorical variables results in a hybrid data set consisting of the original numerical variables supplemented by a set of "new" variables. We explore more options that combine dimension reduction and cluster analysis in more detail later in this review.

5 | DEFINING DISSIMILARITY MEASURES FOR MIXED DATA

Instead of recoding either the categorical or numerical variables, one may alternatively construct a dissimilarity measure that can be applied directly to mixed data. Typically, such a dissimilarity measure can be constructed by defining and combining dissimilarity measures for each type of variable. If we have a distance measure for the mixed data, we can apply most common distance-based clustering algorithms directly.

Let $\mathbf{X}_{n \times p}$ a data set with n objects and p variables of mixed type. Gower's similarity coefficient is one of the most popular measures of proximity for mixed data types (Gower, 1971):

$$g(x_i,x_{i'}) = \frac{\sum_{v=1}^{p} w_v(x_i,x_{i'})s_v(x_i,x_{i'})}{\sum_{v=1}^{p} w_v(x_i,x_{i'})} \tag{1}$$

where $w_v(x_i, x_i^{'})$ is the weight of variable v for the pair of objects $(x_i, x_i^{'})$, with $i, i^{'} \in \{1, 2, \ldots, n\}$ and $i \neq i^{'}$; $s_v(x_i, x_i^{'})$ is the similarity between x_i and $x_i^{'}$ on v. Variable weights are usually equal to 1 unless a variable's value for one or both objects are missing, when the corresponding weight is 0. Differential weighting can be used to express domain-specific knowledge on variable importance (e.g., Hennig & Liao, 2013). However, there is no general way to choose the variable weights.

Concerning the definition of the similarity coefficient in Equation (1), we consider the following cases:

For continuous variables, $s_v(x_i,x_{i'}) = \frac{|x_{iv}-x_{i'v}|}{R_v}$, where R_v the sample range of variable v. This is a range-normalized Manhattan distance converted into similarity.

For binary variables, $s_v(x_i, x_i^{'}) = 0$ if $x_{iv} \neq x_{i\,v}^{'}$ and $s_v(x_i, x_i^{'}) = 1$ if $x_{iv} = x_{i\,v}^{'} = 1$ or $x_{iv} = x_{i\,v}^{'} = 0$. This reduces to the simple matching coefficient. If v is asymmetric binary, where only the value "1" carries information, then same as for binary except that $w_v = 0$ if $x_{iv} = x_{i\,v}^{'} = 0$.

For nominal variables, $s_v(x_i, x_i^{'}) = 0$, if $x_{iv} \neq x_{i\,v}^{'}$ and $s_v(x_i, x_i^{'}) = 1$ if $x_{iv} = x_{i\,v}^{'}$, which corresponds to an extension of the simple matching coefficient to nominal data.

For ordinal variables, a straightforward option is to replace the original values by their associated rank values; more options are available, see for example, Hennig and Liao (2013). Then, the range-normalized Manhattan distance can be used, as in the continuous case. An alternative approach is to use Podani's extension to ordinal variables (Podani, 1999).

By taking $1 - g(x_i, x_i^{'})$ we obtain a dissimilarity measure. Gower's dissimilarity is implemented in the functions **`daisy()`** of the **`R`** package **`cluster`** (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2018) and **`gowdis()`** of the **`R`** package **`FD`** (Laliberté, Legendre, & Shipley, 2014), in the latter with Podani's extension. Once a dissimilarity matrix is derived, distance-based clustering algorithms can be applied. For instance, Partitioning Around Medoids (PAM), a more robust and flexible version of K-means (Kaufman & Rousseeuw, 1990: 68–125), can be used in conjunction with Gower's dissimilarity.

A similar line of research has focused on partitional clustering algorithms that extend K-means to mixed data by calculating distances between objects and cluster centroids for categorical variables and continuous variables, and combine them in a single objective function. Among the most representative methods are K-prototypes (Huang, 1998), K-means for mixed data (Ahmad & Dey, 2007) and Modha–Spangler convex K-means clustering (Modha & Spangler, 2003). The three methods are briefly introduced below using a unified notation.

K-prototypes (Huang, 1998) are a variant of K-means that is based on the weighted combination of the squared Euclidean distance for continuous variables and the matching distance for categorical variables. The objective of K-prototypes is to group the n objects of a data set $\mathbf{X}_{n \times p}$ into K clusters by minimizing the cost function

$$\sum_{k=1}^{K}\sum_{i=1}^{n} z\{ik\}_K d(x_i, Q_k), \tag{2}$$

where Q_k is the centroid or prototype of cluster k and $z\{ik\}_K$ is an element of the $(n \times K)$ partition membership matrix $\mathbf{Z}_K$. Given m continuous variables and $p - m$ categorical variables, $d(x_i, Q_k)$ is a dissimilarity measure given by:

$$d(x_i,Q_k)=d_{con}(x_i,Q_k)+w_l d_{cat}(x_i,Q_k)=\sum_{c=1}^{m}(x_{ic}-q_{kc})^2+w_k\sum_{t=m+1}^{p}\delta(x_{it},q_{kt}), \tag{3}$$

where x_{ic} is the value of the cth continuous variable for object x_i, x_{it} is the value of the tth categorical variable for object x_i, q_{kc} is the mean of the cth continuous variable in cluster k, q_{kt} is the mode of the tth categorical variable in cluster k, $\delta(x_{it}, q_{kt}) = 0$ for $x_{it} = q_{kt}$, and $\delta(x_{it}, q_{kt}) = 1$ for $x_{it} \neq q_{kt}$, and w_k is a user-defined weight of the significance of the entire group of categorical variables in cluster k.

K-prototypes is implemented in the **R** package **clustMixType** (Szepannek, 2017). For extensions of K-prototypes see (Bushel, Wolfinger, & Gibson, 2007; Ji, Bai, Zhou, Ma, & Wang, 2013).

K-means for mixed data (Ahmad & Dey, 2007) combines the squared Euclidean distance for continuous with a special distance for categorical variables, where the distance between two categories is computed as a function of their co-occurrence with other categories. Similar to K-prototypes the aim is to minimize Equation (2), where the dissimilarity between an object and a cluster centroid is given by:

$$d(x_i,Q_k)=\sum_{c=1}^{m}w_c(x_{ic}-q_{kc})^2+\sum_{t=m+1}^{p}\delta(x_{it},q_{kt}),$$

where the first term denotes the distance of object x_i from its closest cluster centroid for continuous variables and the second term denotes the distance between object x_i and its closest cluster centroid for categorical variables. The cluster centroid for continuous variables is calculated as the mean of all values for objects assigned to that cluster, whereas for categorical variables it represents the proportional distribution of each categorical value in the cluster. The weight or significance of the cth continuous variable, w_c, is automatically determined within the algorithm, based on the probability of an object to be pulled toward a cluster depending on the distribution of the different values present in the cluster, see Ahmad and Dey (2007) for a more detailed description.

The computation of the special distance $\delta(\cdot, \cdot)$ used for categorical variables can be described as follows. Let v_1 denote a categorical variable, two of whose values are a and b. In order to find the distance between a and b, the algorithm considers the overall distribution of a and b in the data set along with their co-occurrence with values of other variables. Let v_2 denote another categorical variable. Let ω denote a subset of values of v_2 and $\overline{\omega}$ the complementary set of values occurring for variable v_2. Let $P(\omega|a)$ denote the conditional probability that an object having value a for v_1, has a value belonging to ω for v_2. Using the same notation, $P(\omega|b)$ denotes the conditional probability that an object having value b for v_1, has a value belonging to ω for v_2. The distance between values a and b for v_1 with respect to v_2 is given by $\delta(a,b)=P(\omega|a)+P(\overline{\omega}|b)-1$, where ω is the subset ω of values of v_2 that maximizes the quantity $P(\omega|a) + P(\omega|b)$. The distance between a and b is computed with respect to every other variable and the average value of distances will be the overall distance between a and b in the data set. The distance between a and b with respect to a continuous variable is computed by first discretizing the continuous variable (see Ahmad & Dey, 2007 for more details). The method is implemented in the **R** package **DisimForMixed** (Pathberiya, 2016).

Modha and Spangler (2003) described a convex K-means algorithm which considers a weighted combination of the squared Euclidean distance and the cosine distance for continuous and dummy-coded categorical variables, respectively. Note that the cosine distance between two binary variables equals one minus the cosine of the angle between the corresponding vectors and is widely used in document clustering (Modha & Spangler, 2003). The overall dissimilarity between an object and a cluster centroid is defined similar to K-prototypes, see Equation (3). The weight, w_k, that corresponds to the significance of the entire group of categorical variables in the cluster k is automatically determined within the algorithm. In order to select the optimal weight, Modha and Spangler (2003) define the average within-cluster dispersion separately for continuous and categorical variables, as follows:

$$\Gamma_{con}=\sum_{l=1}^{k}\sum_{i:x_i\in l}d_{con}(x_i,Q_k)$$

$$\Gamma_{cat}=\sum_{l=1}^{k}\sum_{i:x_i\in l}d_{cat}(x_i,Q_k)$$

where Q_k denotes the centroid of cluster k. The cluster centroid for continuous variables is calculated as the mean of all values for objects assigned to that cluster, whereas for categorical variables it is given by $\sum_{i:x_i\in k}\sum_{t=m+1}^{p}x_{it}/\left\|\sum_{i:x_i\in k}\sum_{t=m+1}^{p}x_{it}\right\|$, where $\|\cdot\|$ denotes the vector L_2 norm. The average between-cluster dispersion is defined as

$$\Lambda_{con} = \sum_{i=1}^{n} d_{con}\left(x_i, \overline{Q}_{con}\right) - \Gamma_{con}$$

$$\Lambda_{cat} = \sum_{i=1}^{n} d_{cat}\left(x_i, \overline{Q}_{cat}\right) - \Gamma_{cat}$$

where $\overline{Q}_{con}$ ($\overline{Q}_{cat}$) denote the centroids taken across all continuous (categorical) variables. Finally, the aim is to minimize the product of the continuous and categorical dispersion ratio:

$$\frac{\Gamma_{con}}{\Lambda_{con}} \times \frac{\Gamma_{cat}}{\Lambda_{cat}}.$$

The weight, w_k, is identified through a brute-force search over a range of user-specified values, and the within- to between-cluster dispersion ratio is calculated separately for continuous and categorical variables for each value tested; the value of w_k that minimizes the aforementioned product is selected. Although Modha–Spangler clustering accounts for variable significance within the algorithm, it is vulnerable to individual noninformative variables, due to the fact that the single weight does not allow individual variables to be up- or downweighted (Foss, Markatou, Ray, & Heching, 2016). The Modha–Spangler algorithm is implemented in `R` package **`kamila`** (Foss & Markatou, 2018).

6 | DIMENSION REDUCTION AND CLUSTERING

There exist several dimension reduction techniques for numerical data (e.g., Principal Component Analysis; PCA, [Jolliffe, 2002]), categorical data (e.g., Multiple Correspondence Analysis; MCA [Greenacre, 2017: 137–144]) and mixed data (e.g., Principal Component Analysis or Factor Analysis of Mixed Data; PCAMIX or FAMD [Pagès, 2004]). In addition, there are several methods that combine the use of dimension reduction and cluster analysis. We consider methods that apply distance-based clustering to coordinates in the reduced space, to be distance-based methods as well. In this section, we review some of these methods in a unified framework.

7 | TANDEM APPROACH

Extant dimension reduction techniques all result in new numerical scores (coordinates) for the observations. Hence, an obvious approach is to perform a two-step analysis where cluster analysis is applied to the results of dimension reduction. Such an approach is often referred to as a "tandem analysis" (Hubert & Arabie, 1985).

For data reduction of mixed data, FAMD/PCAMIX, was originally proposed independently by several authors (de Leeuw & van Rijckevorsel, 1980; Hill & Smith, 1976; Kiers, 1991; Pagès, 2004) and can be seen as a compromise between PCA and MCA. In this method, categorical variables are transformed into dummy variables and concatenated with the continuous variables. Each continuous variable is standardized; that is, centered and divided by its standard deviation, and each dummy variable is divided by the squared root of the proportion of objects taking the associated category. PCA is then performed on the resulting matrix. It is not difficult to see that this procedure is equivalent to PCA when there are only continuous variables and to MCA when there are only categorical variables (Vichi, Vicari, & Kiers, 2009).

This type of scaling has a set of desirable properties. The first principal component, denoted f_1, maximizes the link between the continuous and categorical variables in the following sense (Audigier, Husson, & Josse, 2016):

$$\sum_{c=1}^{m} r^2(v_c, f_1) + \sum_{t=m+1}^{p} \eta^2(v_t, f_1)$$

with v_c being the cth continuous variable, v_t the tth categorical variable, m the number of continuous variables, $p - m$ the number of categorical variables, r^2 the square of the correlation coefficient and η^2 the square of the correlation ratio. The second principal component maximizes the aforementioned criterion among orthogonal variables to the first principal component, etc. The first d principal components are preserved to be used in the subsequent clustering step.

FAMD can be described as a PCA of the weighted matrix $\mathbf{XD}_{\Sigma}^{1/2}$, where $\mathbf{D}_{\Sigma}$ is the matrix with diagonal elements equal to $s_1^2, \ldots, s_m^2, \hat{\pi}_{m+1}, \ldots, \hat{\pi}_p$, with s_c the standard deviation of the cth continuous variable and $\hat{\pi}_t$ the proportion of objects taking the associated category of the tth dummy variable. It is trivial to show that this is equivalent to the Singular Value Decomposition of $\mathbf{XD}_{\Sigma}^{1/2} - \mathbf{M}$, where $\mathbf{M}$ the matrix where each row equals to the vector of the means of each column of $\mathbf{XD}_{\Sigma}^{1/2}$.

This specific weighting also implies that the distance between two objects x_i and $x_i^{'}$ in the initial space is a combination of the Euclidean distance used in PCA, for the first m continuous variables and a weighted distance in the spirit of the chi-square distance used in MCA, for the next $p - m$ categorical variables:

$$d^2(x_i, x_{i'}) = \sum_{c=1}^{m} \frac{(x_{ic} - x_{i'c})^2}{s_c^2} + \sum_{t=m+1}^{p} \frac{(x_{it} - x_{i't})^2}{\hat{\pi}_t}$$

FAMD is implemented in **R** packages **FactoMineR** (Lê, Josse, & Husson, 2008), **PCAmixdata** (Chavent, Kuentz-Simonet, Labenne, & Saracco, 2017) and **ade4** (Hill & Smith, 1976).

In the second step of tandem analysis, a clustering method is applied to the object scores on the selected dimensions, *d*. However, the choice of *d* in the first step, can be critical for the final clustering solution. Moreover, notice that in this two-step approach, two different criteria are optimized. While dimension reduction aims at defining a reduced set of combinations of the original variables that maximize the original variability, cluster analysis aims to minimize within-group variability while maximizing between-group variability. This discrepancy of objectives may lead to the so-called cluster masking problem. This problem occurs when the dimension reduction step hides away the underlying cluster structure. This could, for example, occur when variables that are not related to the cluster structure are strongly correlated with each other. An illustrative example of the cluster masking problem is provided in Vichi and Kiers (2001), and further discussed in Van Buuren and Heiser (1989) and De Soete and Carroll (1994). A solution to this problem has been provided by jointly optimizing the two criteria.

8 | JOINT DATA REDUCTION OF MIXED DATA

Several methods for joint dimension reduction and cluster analysis have been proposed that deal with either continuous or categorical data. In particular, for continuous data we distinguish Reduced K-means (RKM) clustering (De Soete & Carroll, 1994), or, equivalently, projection pursuit (Bock, 1987), Factorial K-means (FKM; Vichi & Kiers, 2001) as well as a compromise version of these two methods (Vichi et al., 2009). For categorical data, we have cluster correspondence analysis (van de Velden, Iodice D'Enza, & Palumbo, 2017), which is closely related to GROUPALS (Van Buuren & Heiser, 1989), multiple correspondence analysis and K-means (MCA K-means) (Hwang, Dillon, & Takane, 2006), and iterative factorial clustering of binary variables (i-FCB) (Iodice D'Enza & Palumbo, 2013). As dimension reduction methods effectively transform the complete data to (numerical) data measured on commensurable scales, the joint dimension reduction and cluster analysis methods may also be particularly suited for the analysis of mixed data.

In order to describe the joint dimension reduction and clustering method(s) in a unified framework that allows us to relate the different variants/options, it is practical to introduce some matrix notation. Let X denote a centered and standardized $n \times p$ data matrix, B is a $p \times d$ columnwise orthonormal loadings matrix, that is, $\mathbf{B}^{'}\mathbf{B} = \mathbf{I}_d$, where d is the user-supplied dimensionality of the reduced space. Finally, we use $\mathbf{G}$ to denote the $K \times d$ matrix of cluster centroids in the d-dimensional space.

9 | RKM CLUSTERING OF MIXED DATA

In RKM (De Soete & Carroll, 1994), the simultaneous dimension reduction and cluster analysis problem is tackled in such a way that the cluster allocation and dimension reduction maximizes the *between* variance of the clusters in the reduced space. The RKM objective function is

$$\min \phi_{\text{RKM}}(\mathbf{B}, \mathbf{Z}_K, \mathbf{G}) = \|\mathbf{X} - \mathbf{Z}_K \mathbf{G} \mathbf{B}'\|^2, \tag{4}$$

where $\|\cdot\|$ denotes the Frobenius norm. It is not difficult to see that as solution for the cluster means we have $\mathbf{G} = \left(\mathbf{Z}_K'\mathbf{Z}_K\right)^{-1}\mathbf{Z}_K'\mathbf{X}\mathbf{B}$. Inserting this we obtain

$$\min \phi_{\text{RKM}}(\mathbf{B}, \mathbf{Z}_K) = \|\mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{B}\mathbf{B}'\|^2, \tag{5}$$

where $\mathbf{P} = \mathbf{Z}_K\left(\mathbf{Z}_K'\mathbf{Z}_K\right)^{-1}\mathbf{Z}_K'$ is a projection matrix. Using the projector matrix $\mathbf{P}$ and the trace operator for the sum of diagonal elements of a matrix, we see that

$$\|\mathbf{X} - \mathbf{P}\mathbf{X}\mathbf{B}\mathbf{B}'\|^2 = trace(\mathbf{X}'\mathbf{X}) - trace(\mathbf{B}'\mathbf{X}'\mathbf{P}\mathbf{X}\mathbf{B}). \tag{6}$$

Hence, as $\mathbf{X}$ is constant, minimizing ϕ_{RKM} is equivalent to

$$\max\phi'_{\text{RKM}}(\mathbf{B},\mathbf{Z}_K) = trace\,\mathbf{B}'\mathbf{X}'\mathbf{PXB} \tag{7}$$

For categorical variables, van de Velden et al. (2017) introduced cluster CA. As objective they formulate their method as the maximization of between cluster variation in reduced space. Let $\mathbf{Z}_j$ denote an $n \times p_j$ indicator matrix. That is, each row corresponds to an observation, and the columns represent the p_j categories of the jth categorical variable. Observed categories are coded by ones and all other elements are zero. Consequently, $\mathbf{Z}_j\mathbf{1}_{p_j} = \mathbf{1}_n$. Data on several categorical variables can be collected in a so-called superindicator matrix $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p]$. Furthermore, let $\mathbf{M} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}'_n/n$ denote a centering matrix. The objective of cluster CA can be expressed as

$$\max\phi_{\text{clusca}}(\mathbf{B},\mathbf{Z}_K) = trace\,\mathbf{B}'\mathbf{Z}'\mathbf{MPMZB} \tag{8}$$

subject to

$$\frac{1}{np}\mathbf{B}'\mathbf{D}_z\mathbf{B} = \mathbf{I}_K.$$

Comparing this equation to Equation (7), we see that the methods are closely related. In fact, upon defining $\mathbf{B}^* = \frac{1}{\sqrt{np}}\mathbf{D}_z^{1/2}\mathbf{B}$, the cluster CA objective may also be formulated as

$$\min\phi_{\text{CCA}}\left(\mathbf{B}^*,\mathbf{Z}_K,\mathbf{G}\right) = \left\|\mathbf{D}_z^{-1/2}\mathbf{MZ} - \mathbf{Z}_K\mathbf{G}\mathbf{B}^{*\prime}\right\|^2, \tag{9}$$

subject to

$$\mathbf{B}^{*\prime}\mathbf{B}^* = \mathbf{I}_K$$

Comparing Equations (4) and (9) we see that cluster CA can be seen as RKM of a centered and standardized indicator matrix. In contrast to standardization of numerical values, that is, division by sample standard deviations, for the categorical variables standardization is achieved by dividing through the squared roots of the marginals. Typically, in cluster CA, the loadings are standardized accordingly.

Defining $\mathbf{X}^* = \left(\mathbf{X}\mathbf{D}_z^{-1/2}\mathbf{MZ}\right)$ and inserting this in Equation (4) yields the objective for a joint analysis of mixed data:

$$\min\phi_{\text{mixed RKM}}(\mathbf{B},\mathbf{Z}_K,\mathbf{G}) = \left\|\mathbf{X}^* - \mathbf{Z}_K\mathbf{G}\mathbf{B}'\right\|^2, \tag{10}$$

where $\mathbf{B}'\mathbf{B} = \mathbf{I}_d$.

Hence, as solution for the cluster means we have $\mathbf{G} = \left(\mathbf{Z}'_K\mathbf{Z}_K\right)^{-1}\mathbf{Z}'_K\mathbf{X}^*$ and the loadings/optimal scaling values can be obtained from the eigenequation:

$$\mathbf{X}^{*\prime}\mathbf{PXB} = \mathbf{B}\boldsymbol{\Lambda}. \tag{11}$$

To obtain the cluster allocations $\mathbf{Z}_K$, we need to apply K-means to $\mathbf{X}^*\mathbf{B}$. By iterating between these steps (i.e., K-means and optimal scaling) the objective value of Equation (10) either decreases or does not change resulting in a (local) optimum.

10 | FKM FOR NUMERICAL AND CATEGORICAL DATA

The objective function for FKM, an alternative method for combining cluster analysis and dimension reduction first proposed by Vichi and Kiers (2001), can be expressed as:

$$\min\phi_{\text{FKM}}(\mathbf{B},\mathbf{Z}_K,\mathbf{G}) = \|\mathbf{XB} - \mathbf{Z}_K\mathbf{G}\|^2, \tag{12}$$

Alternatively, using the projector matrix $\mathbf{P}$, we can write

$$\min\phi_{\text{FKM}}(\mathbf{B},\mathbf{Z}_K) = \|\mathbf{XB} - \mathbf{PXB}\|^2. \tag{13}$$

In line with the previous derivations of the relationship between RKM and cluster correspondence analysis, we can formulate an objective function corresponding to the categorical alternative to FKM. That is,

$$\min\phi_{\text{FCCA}}\left(\mathbf{B}^*,\mathbf{Z}_K,\mathbf{G}\right) = \left\|\mathbf{D}_z^{-1/2}\mathbf{MZB}^* - \mathbf{Z}_K\mathbf{G}\right\|^2, \tag{14}$$

subject to

$$\mathbf{B}^{*\prime}\mathbf{B}^{*}=\mathbf{I}_K$$

The loadings/optimal scaling values can be obtained from the eigenequation:

$$\mathbf{Z}^{*\prime}(\mathbf{P}-\mathbf{I})\mathbf{Z}^{*}\mathbf{B}=\mathbf{B}\boldsymbol{\Lambda}. \tag{15}$$

Similar as before, by replacing $\mathbf{X}$ by $\mathbf{X}^{*}=\left(\mathbf{X}\mathbf{D}_z^{-1/2}\mathbf{M}\mathbf{Z}\right)$ we obtain a mixed version of FKM.

11 | GENERAL METHOD FOR DIMENSION REDUCTION AND CLUSTER ANALYSIS

It can be verified (see also, Yamamoto & Hwang, 2014) that the RKM objective function (5), can be decomposed as:

$$\|\mathbf{X}-\mathbf{P}\mathbf{X}\mathbf{B}\mathbf{B}'\|^2=\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{B}'\|^2+\|\mathbf{X}\mathbf{B}-\mathbf{P}\mathbf{X}\mathbf{B}\|^2. \tag{16}$$

Hence, RKM can be seen as a compromise of PCA (the first part of the decomposition) and FKM. Rather than assigning equal weights to the two parts, Vichi et al. (2009) propose to minimize a convex combination of them. They name the resulting method clustering and dimensional reduction (CDR). The objective of CDR thus becomes:

$$\min\phi_{\text{CDR}}(\mathbf{B},\mathbf{Z}_K)=\alpha\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{B}'\|^2+(1-\alpha)\|\mathbf{X}\mathbf{B}-\mathbf{P}\mathbf{X}\mathbf{B}\|^2. \tag{17}$$

Using the trace operator and collecting terms, we see that minimizing ϕ_{CDR} amounts to maximizing:

$$trace\mathbf{B}'\mathbf{X}'((1-\alpha)\mathbf{P}-(1-2\alpha)\mathbf{I})\mathbf{X}\mathbf{B}. \tag{18}$$

Hence, for known $\mathbf{Z}_K$, the loadings $\mathbf{B}$ can be obtained by taking the eigendecomposition of $\mathbf{X}'((1-\alpha)\mathbf{P}-(1-2\alpha)\mathbf{I})\mathbf{X}$, and by selecting orthonormal eigenvectors corresponding to the d largest eigenvalues. On the other hand, for known $\mathbf{B}$, only $trace\mathbf{B}'\mathbf{X}'((1-\alpha)\mathbf{P})\mathbf{X}\mathbf{B}$ from Equation (18) needs to be maximized. This maximization problem is equivalent to a standard K-means clustering objective function applied to $\mathbf{X}\mathbf{B}$.

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, a mixed method is obtained by applying the CDR algorithm to $\mathbf{X}^{*}$.

Combining these two parts, we can formulate, for given α, the following alternating least-squares algorithm:

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 $\mathbf{X}^{*\prime}((1-\alpha)\mathbf{P}-(1-2\alpha)\mathbf{I})\mathbf{X}^{*}$.
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.

12 | APPLICATIONS

In this section, we apply five of the reviewed distance-based methods to real mixed data sets: Gower's dissimilarity measure followed by PAM, K-prototypes, Ahmad's mixed K-means, FAMD followed by K-means (tandem analysis) and mixed RKM (joint analysis). For these five methods, software implementations are readily available through packages in `R`. In particular, we used the packages `cluster` for Gower/PAM (Maechler et al., 2018), `clustMixType` for K-prototypes (Szepannek, 2017), `DisimForMixed` (Pathberiya, 2016) for mixed K-means (Pathberiya, 2016), `FactoMineR` for FAMD/K-means (Lê et al., 2008) and `clustrd` for mixed RKM (Markos, Iodice D'Enza, & van de Velden, 2018). Three real-world data sets, characterized by a different proportion of continuous and categorical variables, were considered: a data set with more categorical than continuous variables, one with more continuous than categorical variables and a balanced scenario.

Appraising the results of different clustering algorithms is a complex task as it is not possible to define a unique criterion. A suitable measure for cluster quality typically depends on the context-specific clustering goal, may it be pure exploration, data reduction, or comparison of the clusters with some further information (Hennig, 2015). For this brief demonstration, the adjusted Rand index (ARI) (Hubert & Arabie, 1985) was used to assess the level of agreement between clustering partitions obtained with different methods, as well as their agreement with an a priori known clustering structure or an informative external criterion, where available. Higher ARI values indicate higher agreement between clustering partitions. All methods were applied to the three data sets with 500 random starts. Assuming no prior knowledge on variable importance, equal variable weights were used for Gower/PAM and K-prototypes.

12.1 | Cleveland heart disease

The Cleveland Heart Disease data set was obtained from the UCI machine learning repository and consists of five continuous and eight categorical variables measured on 303 patients. Following earlier studies that used this data set (Hunt & Jorgensen, 2011), the heart disease variable was recoded in two categories denoting the presence or absence of heart disease. This variable was used as an external partition to evaluate the obtained clustering solutions. Observations that had missing values in any of the 13 variables analyzed were omitted from further analysis, leaving 297 observations.

For this application, we set the number of clusters to 2. For tandem analysis and mixed RKM the number of dimensions was set to 1. Table 1 shows agreement between methods in terms of ARI. As expected, Tandem and mixed RKM resulted in the most similar partitions (0.92), whereas mixed K-means produced the least similar cluster structure with other methods, with ARI values ranging from 0.32 to 0.34. In addition, higher agreement with the "true" clustering partition was observed for mixed RKM (0.41) and tandem analysis (0.39), followed by Gower/PAM (0.38), K-prototypes (0.31), and mixed K-means (0.17).

12.2 | FIFA 18 player data

The Kaggle FIFA 18 Complete Player data set, is publicly available (https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset) and contains data on about 17,000 players of 42 soccer leagues, extracted from the 2017 edition of the EA Sports FIFA gaming series. In this study, we considered only a subset of 1,039 players of three different leagues, the Dutch Eredivisie, the Italian League Serie A and the Greek Super League. For our analysis, we selected 43 continuous variables (player age, height and weight, 37 player performance indicators on the 0–100 scale, player wage, value and release clause in €), and four categorical variables (player's position, attacking work rate, defending work rate and preferred foot). Player's position, that is forward (FW), midfielder (Mid), defender (Def) and goalkeeper (Gk), was used as an external variable to validate the clustering solutions. The number of clusters was set to four making it possible to directly link the clusters to the positions. The number of dimensions for tandem and mixed RKM was set to three.

Mixed RKM and Gower/PAM showed the strongest association with player's position (ARI = 0.51 and 0.50, respectively), followed by K-prototypes (0.49) and Tandem (0.44). Mixed K-means associated the least with the external variable. All methods except mixed K-means placed goalkeepers in a single cluster. The most similar partitions were those of mixed RKM and K-prototypes (.89).

12.3 | TopGear

The data set is included in the **`R`** package **`robustHD`** (Alfons, 2016) and contains information on 242 car models featured on the website of the popular BBC television show Top Gear (http://www.topgear.com). There are 16 categorical and 13 continuous variables. Continuous variables refer to features such as car size (e.g., height, weight), performance (e.g., top speed and acceleration), price, and a review score ranging from 0 to 10. Five of these variables (price, displacement, BHP, torque, top speed) are highly skewed, and were logarithmically transformed. Categorical variables refer to car features and accessories (e.g., alarm system, cruise control), indicating whether each accessory comes with the car as a standard equipment, is optional, or is not present. Other categorical variables indicate manufacturer's origin and type of engine (e.g., petrol, diesel).

A solution with four clusters was selected as this solution most clearly identified distinct and meaningful patterns. The number of dimensions for tandem and mixed RKM was set to 3. Two of the clusters, containing approximately one-fourth to one-third of the observations each, were clearly present in all clustering partitions: a cluster of high performing, expensive and highly regarded cars, with a rear-wheel or four-wheel drive system, which come with many standard features, and another

TABLE 1 Agreement between methods (ARI values) for each of the considered data sets; from left to right, Cleveland heart data, FIFA and TopGear

Data set	Heart ($K = 2$)					FIFA ($K = 4$)					TopGear ($K = 4$)				
Method	#1	#2	#3	#4	#5	#1	#2	#3	#4	#5	#1	#2	#3	#4	#5
#1 PAM	1					1					1				
#2 K-prototypes	0.63	1				0.80	1				0.42	1			
#3 Mixed KM	0.18	0.33	1			0.36	0.45	1			0.32	0.56	1		
#4 Tandem	0.60	0.76	.34	1		0.72	0.86	.41	1		0.58	0.49	0.43	1	
#5 Mixed RKM	0.56	0.69	.32	.92	1	0.80	0.89	.47	.79	1	0.49	0.69	0.66	0.62	1

cluster of compact, cheap, fuel efficient cars, with slow acceleration, a front-wheel drive and no accessories and standards. A less compact cluster, with a size ranging from 25 to 45% across different methods, contains mostly large, diesel-powered cars, a four-wheel drive system and optional equipment available. Mixed RKM, mixed K-means and K-prototypes resulted in the most similar partitions (see Table 1). Both mixed RKM and mixed K-means yielded a tiny cluster consisting of only three car models: BMW i3, Chevrolet Volt, and Vauxhall Ampera. Note that these three models are hybrid or purely electric cars with an additional petrol-powered engine. They were identified as outliers in studies evaluating robust methods (Alfons, Croux, & Gelper, 2016).

13 | CONCLUSIONS

In this review, we considered distance-based clustering methods for mixed data, ranging from data preprocessing to more advanced joint dimension reduction and clustering methods. Despite the ubiquity of mixed data, it appears that no consensus exists concerning the best distance-based clustering approach for such data. One reason for this lies in the difficult issue of assessing the quality of cluster solutions. The context-specific clustering goal, may it be pure exploration, data reduction, or comparison of the clusters with some further information available, conditions the choice of the most suitable technique (Hennig, 2015).

As our examples show, the reviewed methods typically yield distinct solutions and it is not trivial to decide which solution is better. However, some observations appear to hold more generally and are worthwhile mentioning here. First of all, the mixed K-means approach by Ahmad and Dey (2007) yields solutions that are quite different from the results obtained by the other algorithms. The solutions also appear to perform worse when comparing with an informative external cluster variable. Moreover, Ahmad's algorithm is computationally much less efficient than the other methods. Running times were considerably larger than those for any of the other algorithms. For instance, it took more than 48 hours to return a solution for the FIFA data whereas the second slowest method, mixed RKM, only took 5 minutes on the same computer. Secondly, although differences between the joint dimension reduction methods and the tandem approach were limited, Vichi and Kiers (2001) and van de Velden et al. (2017), showed that for, respectively, continuous and categorical data, the tandem approach indeed suffers from the so-called masking problem, where the data subspace picked in the dimension reduction step is not suitable for the clustering step. Note also that, in the case of categorical data, van de Velden et al. (2017) showed considerably better recovery of true cluster structure for the joint methods than full dimensional clustering using Gower's dissimilarity and PAM.

Concerning the joint methods, it is worth mentioning that these methods can be used to generate visualizations of the clusters (and variables) in the reduced spaces. Such visualizations may be useful with respect to cluster interpretation. Examples can be found in Vichi and Kiers (2001) and Timmerman, Ceulemans, Kiers, and Vichi (2010), for numerical data, Hwang et al. (2006), van Dam and van de Velden (2015) and van de Velden et al. (2017) for categorical data, and Vichi et al. (2009), for mixed data.

Finally, it is important to outline that in this review we restricted ourselves to distance-based clustering methods. We are aware that there are several so-called model-based clustering methods. Model-based approaches to clustering typically assume that the observations follow a finite mixture model. There is a wide literature on mixture models, depending on the type of data, that is, continuous or categorical, and on the component distributions being used. In particular, for continuous data, mixtures of Gaussian distributions (Fraley & Raftery, 2002) have been proposed, as well as for example, Gamma (Mayrose, Friedman, & Pupko, 2005), skew normal (Lin, 2009), and generalized hyperbolic distributions (Browne & McNicholas, 2015). Mixture models have also been successfully applied to categorical data (Cai, Song, Lam, & Ip, 2011; Fong & Yip, 1993; Foss et al., 2016).

A mixture model for mixed-type data is the normal-multinomial mixture model (Fraley & Raftery, 2002; Hunt & Jorgensen, 2011). The component distributions are, in this case, joint normal-multinomial. The formulation of the normal-multinomial model changes with the assumption on the within-cluster dependency structure that characterizes the variables in question. In particular, the general model assumes within-cluster dependence for the continuous variables but not for the categorical ones; furthermore, conditional independence between continuous and categorical variables is assumed. In order to specify the conditional dependence for both continuous and categorical variables, Everitt (1988) proposed to discretize the categorical variables with a flexible covariance structure; such method becomes computationally unfeasible in case of a large number of categorical variables. Similar approaches have been proposed by McParland and Gormley (2016) and Browne and McNicholas (2012).

An alternative approach to account for the conditional dependencies between continuous and categorical variables consists of a mixture of location models Lawrence and Krzanowski (1996). A location model defines a new categorical variable: the levels of such variable correspond to each of the possible combinations of the categories characterizing the original (categorical) variables; then the continuous variables are modeled via a multivariate normal density function with parameters depending on the levels of the new-built categorical variable. The mixture of location models and extensions, however, suffer of both computational issues and lack of identifiability (Willse & Boik, 1999).

For a thorough review of model-based clustering methods for mixed-type data we refer the reader to Foss et al. (2018).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLE

Clustering mixed data

ORCID

Michel van de Velden https://orcid.org/0000-0002-9807-9057

Alfonso Iodice D'Enza https://orcid.org/0000-0002-6147-0052

Angelos Markos https://orcid.org/0000-0002-4204-3573

REFERENCES

Ahmad, A., & Dey, L. (2007). A *k*-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, *63*(2), 503–527. https://doi.org/10.1016/j.datak.2007.03.016

Alfons, A. (2016). *robustHD: Robust Methods for High-Dimensional Data [Computer Software manual]*. R package version 0.5.1. Retrieved from https://CRAN.R-project.org/package=robustHD

Alfons, A., Croux, C., & Gelper, S. (2016). Robust groupwise least angle regression. *Computational Statistics & Data Analysis*, *93*, 421–435. https://doi.org/10.1016/j.csda.2015.02.007

Aşan, Z., & Greenacre, M. (2011). Biplots of fuzzy coded data. *Fuzzy Sets and Systems*, *183*(1), 57–71. https://doi.org/10.1016/j.fss.2011.03.007

Audigier, V., Husson, F., & Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, *10*(1), 5–26. https://doi.org/10.1007/s11634-014-0195-1

Bock, H. (1987). On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In H. Bozdogan & A. Gupta (Eds.), *Multivariate statistical modeling and data analysis* (pp. 17–34). Dordrecht: Springer. https://doi.org/10.1007/978-94-009-3977-6_2

Browne, R. P., & McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference*, *142*(11), 2976–2984. https://doi.org/10.1016/j.jspi.2012.05.001

Browne, R. P., & McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *The Canadian Journal of Statistics*, *43*(2), 176–198. https://doi.org/10.1002/cjs.11246

Bushel, P. R., Wolfinger, R. D., & Gibson, G. (2007). Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Systems Biology*, *1*(1), 15. https://doi.org/10.1186/1752-0509-1-15

Cai, J. H., Song, X. Y., Lam, K. H., & Ip, E. H. S. (2011). A mixture of generalized latent variable models for mixed mode and heterogeneous data. *Computational Statistics and Data Analysis*, *55*(11), 2889–2907. https://doi.org/10.1016/j.csda.2011.05.011

Chavent, M., Kuentz-Simonet, V., Labenne, A., & Saracco, J. (2017). Multivariate analysis of mixed data: The PCAmixdata R package. *arXiv preprint arXiv: 1411.4911.*

de Leeuw, J., & van Rijckevorsel, J. (1980). HOMALS and PRINCALS-some generalizations of principal components analysis. *Data Analysis and Informatics*, *2*, 231–242.

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, & B. Burtschy (Eds.), *New approaches in classification and data analysis* (pp. 212–219). Berlin, Germany: Springer-Verlag. https://doi.org/10.1007/978-3-642-51175-2_24

Everitt, B. S. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, *6*(5), 305–309. https://doi.org/10.1016/0167-7152(88)90004-1

Fong, D. Y., & Yip, P. (1993). An EM algorithm for a mixture model of count data. *Statistics & Probability Letters*, *17*(1), 53–60. https://doi.org/10.1016/0167-7152(93)90195-O

Foss, A., & Markatou, M. (2018). kamila: Clustering mixed-type data in R and Hadoop. *Journal of Statistical Software*, *83*(1), 1–44. https://doi.org/10.18637/jss.v083.i13

Foss, A., Markatou, M., & Ray, B. (2018). Distance metrics and clustering methods for mixed-type data. *International Statistical Review*. https://doi.org/10.1111/insr.12274

Foss, A., Markatou, M., Ray, B., & Heching, A. (2016). A semiparametric method for clustering mixed data. *Machine Learning*, *105*(3), 419–458. https://doi.org/10.1007/s10994-016-5575-7

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631. https://doi.org/10.1198/016214502760047131

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, England: Wiley.

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871. https://doi.org/10.2307/2528823

Greenacre, M. (2014). Data doubling and fuzzy coding. In M. Greenacre & J. Blasius (Eds.), *Visualization and verbalization of data* (pp. 239–253). Boca Raton, FL: CRC Press.

Greenacre, M. (2017). *Correspondence analysis in practice*. Boca Raton, FL: Chapman and Hall/CRC.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62.

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, *62*(3), 309–369. https://doi.org/10.1111/j.1467-9876.2012.01066.x

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2015). *Handbook of cluster analysis*. Boca Raton, FL: CRC Press.

Hill, M., & Smith, A. (1976). Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon*, *25*, 249–255. https://doi.org/10.2307/1219449

Huang, Z. (1998). Extensions to the *k*-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304. https://doi.org/10.1023/A:1009769707641

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Hunt, L., & Jorgensen, M. (2011). Clustering mixed data. *WIREs Data Mining and Knowledge Discovery*, *1*(4), 352–361. https://doi.org/10.1002/widm.33

Hwang, H., Dillon, W. R., & Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. *Psychometrika*, *71*, 161–171. https://doi.org/10.1007/s11336-004-1173-x

Iodice D'Enza, A., & Palumbo, F. (2013). Iterative factor clustering of binary data. *Computational Statistics*, *28*(2), 789–807. https://doi.org/10.1007/s00180-012-0329-x

Ji, J., Bai, T., Zhou, C., Ma, C., & Wang, Z. (2013). An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing*, *120*, 590–596. https://doi.org/10.1016/j.neucom.2013.04.011

Jolliffe, J. (2002). *Principal component analysis*. New York: Springer-Verlag.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An Introduction to cluster analysis*. Hoboken, NJ: John Wiley & Sons.

Kiers, H. A. (1991). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, *56*(2), 197–212. https://doi.org/10.1007/BF02294458

Laliberté, E., Legendre, P., & Shipley, B. (2014). FD: measuring functional diversity from multiple traits, and other tools for functional Ecology [Computer software manual]. R package version 1.0-12. Retrieved from: https://CRAN.R-project.org/package=FD.

Lawrence, C. J., & Krzanowski, W. J. (1996). Mixture separation for mixed-mode data. *Statistics and Computing*, *6*(1), 85–92. https://doi.org/10.1007/BF00161577

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Lin, T. I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, *100*(2), 257–265. https://doi.org/10.1016/j.jmva.2008.04.010

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2018). cluster: Cluster analysis basics and extensions [Computer Software manual]. R package version 2.0.7-1. Retrieved from https://CRAN.R-project.org/package=cluster.

Markos, A., Iodice D'Enza, A., & van de Velden, M. (2018). clustrd: Methods for joint dimension reduction and clustering [Computer Software manual]. R package version 1.2.3. Retrieved from https://CRAN.R-project.org/package=clustrd.

Mayrose, I., Friedman, N., & Pupko, T. (2005). A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, *21*(Suppl 2, 151–158. https://doi.org/10.1093/bioinformatics/bti1125

McParland, D., & Gormley, I. C. (2016). Model based clustering for mixed data: ClustMD. *Advances in Data Analysis and Classification*, *10*(2), 155–169. https://doi.org/10.1007/s11634-016-0238-x

Mirkin, B. (2005). *Clustering: A data recovery approach*. London: CRC Press.

Modha, D. S., & Spangler, W. S. (2003). Feature weighting in *k*-means clustering. *Machine Learning*, *52*(3), 217–237. https://doi.org/10.1023/A:1024016609528

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée*, *52*(4), 93–111.

Pathberiya, H. A. (2016). DisimForMixed: Calculate dissimilarity matrix for dataset with mixed attributes [Computer software manual]. R package version 0.2. Retrieved from https://CRAN.R-project.org/package=DisimForMixed.

Podani, J. (1999). Extending Gower's general coefficient of similarity to ordinal characters. *Taxon*, *48*, 331–340. https://doi.org/10.2307/1224438

Szepannek, G. (2017). clustMixType: *k*-prototypes clustering for mixed variable-type data [Computer software manual]. R package version 0.1-29. Retrieved from https://CRAN.R-project.org/package=clustMixType

Timmerman, M. E., Ceulemans, E., Kiers, H. A., & Vichi, M. (2010). Factorial and reduced *k*-means reconsidered. *Computational Statistics & Data Analysis*, *54*(7), 1858–1871.

Van Buuren, S., & Heiser, W. J. (1989). Clustering *n* objects into *k* groups under optimal scaling of variables. *Psychometrika*, *54*(4), 699–706. https://doi.org/10.1007/BF02296404

van Dam, J. W., & van de Velden, M. (2015). Online profiling and clustering of facebook users. *Decision Support Systems*, *70*, 60–72. https://doi.org/10.1016/j.dss.2014.12.001

van de Velden, M., Iodice D'Enza, A., & Palumbo, F. (2017). Cluster correspondence analysis. *Psychometrika*, *82*(1), 158–185. https://doi.org/10.1007/s11336-016-9514-0

van Rijckevorsel, J. (1988). Fuzzy coding and B-splines. In J. van Rijckevorsel & J. de Leeuw (Eds.), *Component and correspondence analysis. Dimension reduction by functional approximation* (pp. 33–54). Chichester, England: Wiley.

Vichi, M., & Kiers, H. A. (2001). Factorial *k*-means analysis for two-way data. *Computational Statistics & Data Analysis*, *37*(1), 49–64. https://doi.org/10.1016/S0167-9473(00)00064-5

Vichi, M., Vicari, D., & Kiers, H. (2009). Clustering and dimensional reduction for mixed variables. *Behaviormetrika* 2018. Unpublished manuscript

Willse, A., & Boik, R. J. (1999). Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing*, *9*(2), 111–121. https://doi.org/10.1023/A:1008842432747

Yamamoto, M., & Hwang, H. (2014). A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika*, *41*(1), 115–129.

How to cite this article: van de Velden M, Iodice D'Enza A, Markos A. Distance-based clustering of mixed data. *WIREs Comput Stat*. 2018;e1456. https://doi.org/10.1002/wics.1456