Comput Stat (2016) 31:771–798
DOI 10.1007/s00180-016-0644-8

ORIGINAL PAPER

Density-based clustering with non-continuous data

Adelchi Azzalini[1] · Giovanna Menardi[1]

Received: 29 September 2014 / Accepted: 21 January 2016 / Published online: 18 February 2016

Abstract Density-based clustering relies on the idea of associating groups with regions of the sample space characterized by high density of the probability distribution underlying the observations. While this approach to cluster analysis exhibits some desirable properties, its use is necessarily limited to continuous data only. The present contribution proposes a simple but working way to circumvent this problem, based on the identification of continuous components underlying the non-continuous variables. The basic idea is explored in a number of variants applied to simulated data, confirming the practical effectiveness of the technique and leading to recommendations for its practical usage. Some illustrations using real data are also presented.

Keywords Density estimation · Mixed variables · Modal clustering · Model-based clustering · Multidimensional scaling

1 Background and motivation

Cluster analysis refers to a widespread class of methods for exploring data with the aim of finding groups of similar objects. This goal has been traditionally achieved by evaluating some measure of distance/dissimilarity between the observations. A

Electronic supplementary material The online version of this article (doi:10.1007/s00180-016-0644-8) contains supplementary material, which is available to authorized users.

✉ Giovanna Menardi
menardi@stat.unipd.it

Adelchi Azzalini
azzalini@stat.unipd.it

[1] Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Padova, Italy

popular account is the book of Kaufman and Rousseeuw (1990). An alternative, more recent strand in cluster analysis has examined the problem via the introduction of some notion of 'density' associated to the data. In the statistical literature, the term density is intended as density of a probability distribution, from which the observations are supposed to be sampled. Even with this specification, there exist at least two distinct approaches within this framework.

The most consolidated approach to density-based clustering is the one denoted as 'model-based'. Although its formulation goes back to Wolfe (1970), model-based clustering has become increasingly popular since the works of Fraley and Raftery (1998, 2002). In this setting, the d-dimensional density $f(x)$ underlying the observed data is assumed to be a mixture of a number, G say, of component densities, $f_1, \ldots, f_G$ which belong to some specified parametric family of distributions, each with different parameters. The problem then translates into the one of estimating the parameters of $f_1, \ldots, f_G$ and the vector of mixing probabilities; this task is typically tackled by maximum likelihood with the aid of the Expectation Maximization algorithm. The basic parametric assumption about $f_1, \ldots, f_G$ is that they are all d-dimensional normal densities. In recent years, more flexible parametric families have been considered (e.g., Lin 2010). In model-based clustering, each component density f_g corresponds to a cluster. The number of clusters, G, is either pre-assigned or is estimated using some additional criterion, typically information-based.

Within the density-based approach, an alternative to model-based clustering arises when the underlying density $f(x)$ is estimated non-parametrically and its modes are regarded as identifiers of the clusters; hence the term 'modal' or 'non-parametric clustering' will sometimes be used for this methodology. One formulation within this approach aims at estimating the modes of f and associates each cluster to the set of points along the steepest ascent path towards a mode. Most of contributions which follow this direction can be considered as a refinement of the *mean-shift* clustering, early proposed by Fukunaga and Hostetler (1975). Another class of methods associates the clusters to disconnected density level sets of the sample space so that the modes correspond to the innermost points of these sets. More in detail, the intersection $f(x) = k$, with a given level value k, singles out high-density sets; moving k along its feasible range gives rise to a tree structure of the high-density sets, hence of the clusters. The basic idea of this formulation was put forward a long time ago (Wishart 1969; Hartigan 1975, Sect. 11.13) but it is only relatively recently that it has been translated into some fully developed and operational procedures. Among these, we recall the ones of Stuetzle (2003), Azzalini and Torelli (2007), and their subsequent developments reviewed by Stuetzle and Nugent (2010) and Menardi and Azzalini (2014), respectively.

It is worth to underline that both density-based formulations recalled above incorporate a well-defined concept of what constitutes a 'cluster', although with a slight difference between the two approaches, and provide a way to estimate the number of clusters. This fact represents both a conceptual and a practical advantage over traditional distance-based methods, which do not supply a similar outcome, and justifies an extra effort which may be required to apply these methods even in the presence of some limitations.

Whatever specific formulation in density-based clustering is taken, one has to live with the assumption, intrinsic to this approach, that the observations are of continuous type. In some cases, this condition represents a severe limitation, since in a range of applications at least some of the variables are instead of non-continuous type. Social and economic studies are unarguably those where non-continuous observations occur most frequently, but not by any means the only ones. The most common type of non-continuous variables is represented by the categorical ones. However, sometimes even numeric variables can be problematic, when they are highly discretized; for instance, the number of successful pregnancies of a woman in her lifetime spans a small number of values, especially in Western countries.

In principle, the model-based approach is applicable also to non-continuous data. The classical approach for categorical data, often referred to as 'latent class analysis', assumes f to be a mixture of multinomial distributions (Goodman 1974); a recent advancement which considers correlated variables has been proposed by Marbac et al. (2015). In practice, the model specification becomes difficult when the observed variables are of heterogeneous type, because of the requirement to formulate a probability model which combines variables of different nature, potentially involving a joint distribution with continuous, discrete and categorical components. Some attempts in this direction have been pursued by Vermunt and Magidson (2002) and by Hunt and Jorgensen (2003); both formulations rely on some disputable assumption of independence between blocks of variables of different type. In the same context, the work of Browne and McNicholas (2012) explores the use of latent variables mixture models to cluster data of mixed type; in this case the assumption is the observed variables are independent conditional to the latent variables.

On the contrary, within the non-parametric formulation of the density-based clustering, to the best of our knowledge there has been no attempt to overcome the crucial assumption of observing continuous data only.

The aim of the present contribution is to put forward a technique to circumvent the restriction of density-based methods to continuous variables, to examine its working in a range of situations and to provide recommendations for its practical usage. This technique builds on the widespread idea of reconstructing the continuous latent structure underlying the observed data, and to apply density-based clustering subsequently. Also due to its simplicity, the proposed technique can be applied to model-based clustering and to modal clustering of data comprising variables of mixed type, quantitative and qualitative.

2 On the reconstruction of a continuous latent structure

2.1 Formalization

In many cases, although certainly not universally, categorical variables are collected having in mind that they are representatives of some underlying continuous variables. This can occur in two distinct forms. In the first one, a continuous variable is known to exist, but its direct measurement is not feasible or it is at least problematic. The most typical example is represented by personal income, which in many surveys is not asked directly, to avoid a grossly biased response or non-response, and it is therefore recorded indirectly via a battery of questions related to life conditions. The second, more frequent, situation is when a continuous underlying variable is a convenient mental construct, but it is not observable in principle; examples are intelligence, risk aversion and abilities of various kind. In these cases, the search for one or possibly a few continuous variables as the latent structure underlying a set of observed categorical variables is especially natural.

To formalize the problem, our exploration starts by considering a simple idealized situation where a continuous variable is not observable directly, but only through a set of dichotomous manifest variables derived from the latent one. This simplification should not mislead the attention from the actual focus of this work: although we have in mind categorical variables, polytomous or even ordered, as well as discrete variables, we are now considering the binary case as this represents the most extreme departure from continuous data, hence presumably the most challenging situation to be tested. Additionally, our aim is to use density-based methods for clustering data of heterogeneous type, then not only categorical. However, since density-based methods essentially requires continuous variables to work with, we shall focus for now on the case of non-continuous variables and discuss how to handle mixed data afterwards.

Consider a continuous random variable Z, which is only assumed to admit a density function $f(z)$, $z \in \mathbb{R}$ and let $Z^{(1)}, \dots, Z^{(L)}$ be a set of noisy version of Z:

$$Z^{(l)} = Z + \epsilon^{(l)}, \quad l = 1, \dots, L, \tag{1}$$

where $\epsilon^{(1)}, \dots, \epsilon^{(L)}$ are all $\mathrm{N}(0, \sigma_\epsilon^2)$, mutually independent and independent of Z. In place of the variable of interest, Z, we observe instead

$$X^{(l)} = \mathrm{sign}(Z^{(l)}), \quad l = 1, \dots, L. \tag{2}$$

Step (1) is necessary in this construction otherwise, if $\sigma_\epsilon = 0$, all $X^{(l)}$ would coincide, and increasing L would not increase the available information. On the reverse side, if $\sigma_\epsilon \to \infty$, the amount of information on Z carried by each $X^{(l)}$ vanishes.

A question of interest is the following. How does the amount of information on Z increase as the number L of binary representations $X^{(l)}$ increases? To answer this question, we may measure the error of approximating Z with the sequence $X^{(1)}, \dots, X^{(L)}$ via mean square error:

$$\mathrm{MSE} = E\left[\left(Z - \sum_{l=1}^{L} \frac{X^{(l)}}{L}\right)^2\right]. \tag{3}$$

Some simple but tedious algebraic work leads to:

$$\mathrm{MSE} = E\left(Z^2\right) + \frac{1}{L}E\left(X^{(1)2}\right) + \frac{L-1}{L}E\left(X^{(1)}X^{(2)}\right) - 2E\left(ZX^{(1)}\right) \tag{4}$$

where the expectations depend on $\mathrm{var}(Z)$ and σ_ϵ^2.

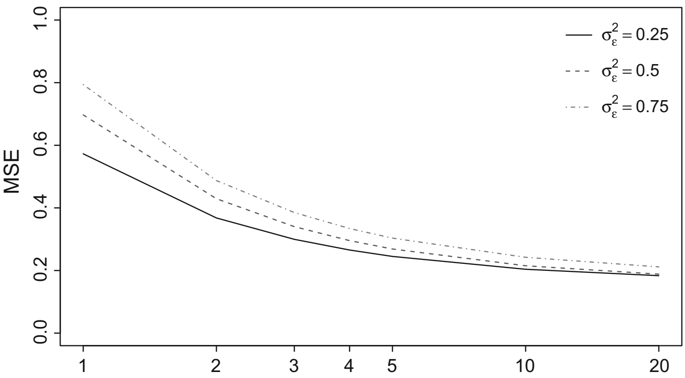


Fig. 1 Plot of MSE versus L, on a log-scale, for some values of σ_ϵ, when $Z \sim N(0, 1)$

To get a perception of the MSE behaviour, Fig. 1 displays its value versus L, on the logarithmic scale, for some values of σ_ϵ when $Z \sim N(0, 1)$. In this case, the moments involved in (4) take the form

$$E\left(Z^2\right) = 1, \quad E\left({X^{(1)}}^2\right) = 1,$$

$$E\left(X^{(1)}\,X^{(2)}\right) = \frac{2}{\pi}\arcsin\left(\frac{1}{1+\sigma_\epsilon^2}\right), \quad E(Z\,X^{(1)}) = \sqrt{\frac{2}{\pi(1+\sigma_\epsilon^2)}}\,.$$

Details about these computations are provided in the accompanying Supplementary Material. The main message conveyed by Fig. 1 is that the MSE decreases appreciably only for small values of L. In the approximate range 3 to 5 the rate of decrease slows down and beyond about $L = 5$ is quite limited. A precise statement cannot be made, also because of the concurrent effect of σ_ϵ, but the behaviour appears essentially of this sort. Conversely, the plot indicates that, when a set of manifest dichotomous variables is used to reconstruct a continuous latent structure, a drastic reduction in number can be legitimate.

In practical work, several latent variables plausibly co-exist, each with the role of our Z here, and subject-matter considerations are of key importance in deciding how many latent variables one could reasonably attempt to reconstruct.

2.2 Continuous variables from multidimensional scaling

Since density-based methods essentially requires continuous variables to work with, we must convert non-continuous variables into continuous ones. The constructed variables do not need to be equal in number to the original non-continuous variables.

The essence of our proposal to tackle the above-stated problem is outlined in the following paragraph.

- The first step is to construct a dissimilarity matrix, D say, of the observed units. In general, the dissimilarity matrix can be used for merging information from all variables, whether they are continuous, discrete, ordered or unordered categorical, to quantify the dissimilarity between any given pair of units.

- Once D has been obtained, we make use of Multidimensional scaling (MDS) to construct a set of continuous variables coherent with D at least approximately, that is, the distances between units as measured on the newly created MDS variables are as close as possible to the original dissimilarities.
- At this stage, all available information is coded into continuous variables and density-based methods can be applied.

The above broad scheme gives rise to several variants, for three reasons: (a) the actual construction of the matrix D offers a large variety of options; (b) the MDS stage can be carried out in different ways, although the set of alternatives for this step is relatively more limited; (c) the transformation using dissimilarity and MDS may be applied to all the observed variables or to the non-continuous ones only, keeping the continuous variables unchanged;

We shall not be concerned with aspect (a), since this stage is the same which would be accomplished by someone who is tackling the clustering problem using traditional distance-based methods, which represent a reference methodology for our comparisons. Indeed, the choice of the dissimilarity metric has been historically a controversial matter which cannot be tackled without considering the nature of the data, the goal of the analysis, and subject-matter knowledge. We refer the reader to the existing literature for a discussion about the specification of D (e.g., Kaufman and Rousseeuw 1990, Chapter 1).

As for question (b), MDS was developed to produce multidimensional geometric representations of data, where quantitative or qualitative relationships in the data are made to correspond with geometric relationships in the representation; in this sense, MDS lends itself to our purpose of extracting the continuous latent structure underlying a set of data of mixed nature. More specifically, MDS starts with information about the pairwise dissimilarity between the elements of a set of objects, and identifies a new configuration of data defined in a metric space, where the pairwise distances between the new data are the best approximation of the originally observed dissimilarities (metric MDS), or possibly of some monotone increasing function of them (non-metric MDS). We recall that Multidimensional scaling is widely documented in standard texts; see for instance Mardia et al. (1979, Chapter 14).

Typically, the number of MDS variables is reduced, even substantially, compared to the original ones, at least when the latter are in the form of categorical variables. Concerning the actual use of MDS, we have to decide about the following aspects: (i) choose between the so-called metric (which comprises various forms itself) and non-metric version of MDS, which depends only on the ranking of dissimilarities, hence it is invariant over monotonic transformations of the dissimilarities; (ii) decide how many MDS variables to construct. These issues are explored through the simulations of Sect. 3; its final subsection deals with the case of mixed variable and question (c) above.

A possible remark is that the proposed method is an instance of the so-called 'tandem method', in which data undertake some preliminary processing before entering a clustering procedure (Arabie and Hubert 1994). This scheme has been criticized as the pre-clustering stage, often principal component analysis, can destroy the grouping structure of the original data. In the present case, however, the nature of MDS, which by construction preserves distances at least approximately, provides a form of safeguard against such a danger.

Also, it is appropriate to state that the route examined here is not the only possibility for tackling the problem. An alternative is provided by the concept of factor analysis for categorical data, introduced by Bartholomew (1980) and examined further by Bartholomew and Knott (1999) under the heading of latent trait models. This would represent the natural route to implement the theoretical framework described in Sect. 2.1 and it would also have the advantage of transferring information from categorical variables directly to continuous ones, without going through the intermediate step of the dissimilarity matrix. The present choice has been adopted on the ground of simplicity and flexibility, since the dissimilarity matrix can easily incorporate information also from discrete and ordered categorical variables; in addition the corresponding software tools are more widely available and more familiar to the community working in cluster analysis. Finally, it is worth to stress that the use of MDS appears as the most appropriate for the subsequent application of clustering methods. This choice does not mean to rule out the potential usefulness of latent trait models, a route which might deserve a separate exploration.

Another possible direction, in the case of binary or ordinal categorical variables, could be to introduce a latent continuous multivariate distribution whose marginal components are observed only in the form of intervals to which the units belong. The latent-variable formulation is commonly employed in the context of categorical variables but it becomes rapidly cumbersome if the additional aspect of clustering is superimposed. Moreover this formulation would not apply to unordered categorical variables and it would be feasible only within the mixture-model approach to clustering, not with modal clustering. The latter restriction is also shared by Oh and Raftery (1998), who introduce a Bayesian model to cluster objects based on their dissimilarities. The authors do not explicitly consider to apply the model on categorical or mixed data after computing their distances, but this seems to us quite natural to extend the usability of the method. Once more, we are motivated to explore the current proposal which, as already remarked, combines wide flexibility and simplicity.

3 Numerical exploration via simulations

3.1 Simulation design and specification of the methods

To pursue the overall task stated in Sect. 1, under the specifications described in Sect. 2, we have run a quite extensive simulation study.

Similarly to the setting of Sect. 2.1, we consider simulation of samples of size n from a continuous and unobservable random variable Z, whose density function is denoted $f_Z(\cdot)$, except that now $Z = (Z_1, \ldots, Z_d)'$ can be multidimensional. The distribution of Z is characterized by a structure which comprises G clusters, in a range of possible forms described below.

The various methods are applied under the assumption that $Z = (Z_o', Z_u')'$ is formed by two blocks of continuous variables with d_o and d_u components, respectively, such that Z_o is observable while Z_u is unobservable. In place of Z_u, we observe a set of binary variables, supposed to be simplified categorical representations of its underlying components Z_j:

$$X_j^{(l)} = \text{sign}\left(Z_j + \epsilon_j^{(l)}\right), \quad \epsilon_j^{(l)} \sim N\left(0, \sigma^2_{\epsilon_j}\right), \tag{5}$$

where $l = 1, \dots, L$, $j = 1, \dots, d_u$. Therefore, it is assumed that observations on the set of variables

$$X = \left(X_1^{(1)}, \dots, X_1^{(L)}, \dots, X_j^{(1)}, \dots, X_j^{(L)}, \dots, X_{d_u}^{(1)}, \dots, X_{d_u}^{(L)}\right)', \tag{6}$$

are available for each of n units, along with the possible observations of Z_o. There is no compelling reason for keeping L constant across the d_u component variables; this choice has been made for mere simplicity.

The task is to cluster these n units into homogeneous groups by methods recalled in Sect. 1 combined with the preliminary process described in Sect. 2.2. More specifically, we explore the following situations:

(i) $d_o = 0$ (and $d_u = d$), i.e. the observed data are of purely qualitative (binary) nature. Here, the key steps are the following: (1) from the X variables, compute a dissimilarity matrix D among the n units; (2) from D, apply MDS to obtain a numerical configuration $Z^* = (Z_1^*, \dots, Z_{d^*}^*)'$; (3) perform density-based clustering on Z^*.

(ii) $d_o > 0$, i.e. mixed categorical and continuous data are observed. To handle data of this sort, two options have been explored, as follows.
   1. MDS is applied to the dissimilarity matrix computed from all observed variables, $Z^* = (Z_o', X')' \in \mathbb{R}^{d^*}$, and clustering is performed on these MDS variables.
   2. The observed continuous components Z_o are retained in their original form and clustering is applied to the variables $Z^* = (Z_o', Z_u^{*\prime})'$, where Z_u^* denotes the set of MDS variables extracted from the dissimilarity matrix of the binary data X.

It stands to reason that information about the true underlying clustering structure, as well any other information on Z_u not provided by X is pretended not to be known, and is used for assessing results only.

In evaluating the quality of the performances, we are interested in distinguishing the possible sources of erroneous partitioning. This may be due to an awkward true clustering structure of the latent vector Z, or to the unavoidable loss of information when the underlying continuous variables Z_u give rise to the categorical variables X, or possibly to a disruption of the clustering structure caused by the MDS transformation. To get some insight on these possible effects, the true clustering structure of Z is compared with the following outcomes: (a) the partitions detected by density-based clustering on Z^*; (b) the partitions detected by the same techniques on the latent variables Z; (c) the clusters detected by some benchmark clustering methods for mixed data.

For each of these outcomes, the agreement with the real underlying grouping is assessed via the adjusted Rand index (Hubert and Arabie 1985), ARI. This index takes the value 1 when there is perfect agreement between the two partitions, while its expected value is 0 under random allocation; negative values are also possible, indicating a classification worse than what would be expected under a random allocation.

The actual implementation of the numerical work has taken place within the computing environment R (R Development Core Team 2011), complemented by some of its packages. Details are provided next, along with a description of the specific methods employed.

– Dissimilarities between binary variables can be specified in several ways. Our option is simply to compute the Euclidean distance between the vectors of binary values using the R function `dist`.
– Among the several available MDS variants, we consider the two most commonly in use, that is, its 'classical' version as the representative of metric MDS, computed in R by function `cmdscale`, and the non-metric version, computed by function `isoMDS` of the R package `MASS`, version 7.3–37 (Venables and Ripley 2002).
– We have to select some specific density-based clustering methodologies on which trying-out our procedure. Only very few of them can be examined, considering that they must be tested in a variety of situations and that some of the proposed theoretical formulations have not yet reached an operational stage. In particular, we consider the model-based clustering methodology implemented by package `mclust` version 4.4 (Fraley et al. 2012), which is based on mixtures of normal densities. Concerning the modal approach, we consider one representative of the level set formulation proposed by Azzalini and Torelli (2007) and implemented by package `pdfCluster` version 1.0–1 (Azzalini and Menardi 2014). Optional parameters of the two methods are left to their default values. In both cases, selection of the number of clusters is integral part of the procedure: in model-based clustering this is based on the Bayesian information criterion; in modal clustering this corresponds to the number of the detected modes which, in turn, depends on the amount of smoothing to estimate the density. This latter value has been selected as asymptotically optimal for normal data and then shrinked by a value of 3/4 to reduce oversmoothing. While the criteria adopted for both formulations are indeed not optimal, these are standard choices as also correspond to the default value of associated arguments in the adopted packages.
– The number of operations required to run the whole methodology is simply the sum of the number of operations required to run its single steps. Both MDS and density-based clustering have a computational complexity which strongly depends on the specific method adopted. In principle, classical MDS is computationally more burdensome than non-metric MDS (e.g., Tzeng et al. 2008). In practice, we experienced that the computing time for non-metric MDS is more than 10 times larger than for classical MDS and several thousands observations are easily handled. Concerning the clustering step, the adopted modal clustering method requires $O\left(\frac{n^{\lfloor p/2 \rfloor}}{\lfloor p/2 \rfloor}\right)$ operations to run on a $n \times p$ matrix when $p > 3$, and $O(n \log n)$ operations for smaller p; model-based clustering requires $O(Kp^2n)$ operations to estimate a mixture of K normal components. Whatever option is selected, the time to run the whole procedure has an order of magnitude of the seconds when applied to 5-variate samples of size up to 500.

- As the dissimilarity between many pairs of units can happen to be the same when the number of binary variables is small, the resulting numerical configuration is usually characterized by a large number of ties. In this case, non-parametric density estimation may be problematic, due to the small variability in the sample. For this reason, in addition to the application of modal clustering on the Z^* vectors, we also run the same methods on a jittered version of Z^*, obtained by adding a small amount of uniformly distributed noise to the data. Jittering has been used also in connection with `mclust` but in this case it made no real difference; therefore, the corresponding results have not been reported.
- We include two reference methods for comparison: dissimilarity-based clustering, namely the K-medoids method implemented by function `pam` of package `cluster` version 1.15–3 (Maechler et al. 2013), and latent class analysis. The latter approach is tested on the mixed data settings (ii) only, where some difficulties may result in the specification of a joint distribution involving heterogeneous data. We formulate a mixture of multivariate Bernoulli and Gaussian distributions for the binary, and respectively, continuous variables and we assume the distributions to be conditional independent between the different types of variables. Latent class analysis has been implemented by function `flexmix` of package `flexmix` version 2.3-12 (Gruen and Leisch 2008; Leisch 2004). Both the reference methods are given a head start by setting the number of clusters to the true number G. For a comprehensive assessment and a comparison of the two approaches see Anderlucci and Hennig (2014).

3.2 The cases examined

Simulations have been run with the following settings: $n = 100$ (except in cases D5C and D5D specified below), $L \in \{1, 2, 3, 4, 5, 10, 20\}$, $d \in \{1, 2, 5\}$, $\sigma_\epsilon = 0.25\ \hat{\text{var}}(Z_j)^{1/2}$, $d^* \in \{1, \ldots d\}$, where $\hat{\text{var}}(Z_j)$ denotes the sample variance of the n values drawn from of Z_j. For each case considered, $N = 5000$ replicates have been examined.

Next, we describe the selected cases for the distribution of Z from which the samples are drawn. The following notation is adopted: $N_d(\mu, \Sigma)$ denotes the d-dimensional normal distribution with mean μ and variance Σ, $U_{\mathcal{S}}$ denotes a uniform distribution defined over the set $\mathcal{S}$, 1_d is the unitary vector with d components, I_d is the identity matrix of dimension d and $\text{vech}(\Sigma)$ is the vector formed by the lower triangle of a symmetric matrix Σ.

d = 1 D1A well separated groups
$\sum_{g=1}^{2} \pi_g N(\mu_g, \sigma_g)$
$\mu_1 = -2.5, \mu_2 = 2.5, \sigma_1 = \sigma_2 = 1, \pi_1 = \pi_2 = 0.5;$
D1B less separated groups
$\sum_{g=1}^{2} \pi_g N(\mu_g, \sigma_g)$
$\mu_1 = -1.5, \mu_2 = 1.5, \sigma_1 = \sigma_2 = 1, \pi_1 = \pi_2 = 0.5;$
d = 2 D2A well separated groups
$\sum_{g=1}^{2} \pi_g N(\mu_g, \Sigma_g)$
$\mu_1 = 2.5 \cdot 1_2, \mu_2 = -2.5 \cdot 1_2, \Sigma_1 = \Sigma_2 = I_2, \pi_1 = \pi_2 = 0.5;$

D2B less separated groups

$\sum_{g=1}^{2} \pi_g N(\mu_g, \Sigma_g)$

$\mu_1 = 1.5 \cdot 1_2, \mu_2 = -1.5 \cdot 1_2, \Sigma_1 = \Sigma_2 = I_2, \pi_1 = \pi_2 = 0.5;$

D2C nonconvex groups

$\sum_{g=1}^{2} \pi_g U_{\mathcal{S}_g}$

$\mathcal{S}_1 = \{z > 0 : \sum_{j=1}^{d} z_j^2 = 1\}, \mathcal{S}_2 = \{z < 0 : \sum_{j=1}^{d} z_j^2 = 1\},$

$\pi_1 = \pi_2 = 0.5;$

D2D groups whose structure is only distinguishable in a multidimensional space.

$\sum_{g=1}^{3} \pi_g N(\mu_g, \Sigma_g)$

$\mu_1 = (-1.5, 1.5)', \mu_2 = (1.5, -1.5)', \mu_3 = (-2.5, -2.5)', \pi_1 = \pi_2 = \pi_3 = 1/3, \text{vech}(\Sigma_1) = \text{vech}(\Sigma_2) = (0.8, 0.8, 1)', \text{vech}(\Sigma_3) = (0.8, -0.8, 1)';$

d = 5 D5A well separated groups $\sum_{g=1}^{2} \pi_g N(\mu_g, \Sigma_g)$

$\mu_1 = 2.5 \cdot 1_5, \mu_2 = -2.5 \cdot 1_5, \Sigma_1 = \Sigma_2 = I_5, \pi_1 = \pi_2 = 0.5;$

D5B less separated groups

$\sum_{g=1}^{2} \pi_g N(\mu_g, \Sigma_g)$

$\mu_1 = 1.5 \cdot 1_5, \mu_2 = (-1.5) \cdot 1_5, \Sigma_1 = \Sigma_2 = I_5, \pi_1 = \pi_2 = 0.5;$

D5C non-convex groups

$\sum_{g=1}^{2} \pi_g U_{\mathcal{S}_g},$

where $\mathcal{S}_1$ is the portion of a 5-dimensional unit torus having positive coordinates, and $\mathcal{S}_2$ is similar with negative coordinates, $\pi_1 = \pi_2 = 0.5$.

Due to the more complex structure and to a rather high dimension, samples of size $n = 300$ have been drawn from this distribution. See Fig. 2 for an illustration of the 3-d analogue of this structure having major radius $3/2$ and minor radius 1.

D5D groups whose structure is distinguishable only in a multidimensional space

$\sum_{g=1}^{5} \pi_g N(\mu_g, \Sigma_g)$

$\mu_1 = (-3, 3, 3, 3, 3)', \mu_2 = (3, -3, 3, 3, 3)', \mu_3 = (3, 3, -3, 3, 3)', \mu_4 = (3, 3, 3, -3, 3)', \mu_5 = (3, 3, 3, 3, -3)', \pi_j = 1/5,$ and $\Sigma_j = I_5$ for $j = 1, \ldots, 5$.

Also in this complex case $n = 300$ has been adopted.

All these settings, except D2D and D5D, have an essentially unidimensional structure as far as clustering is concerned, in the sense that projecting the data on a suitably selected unidimensional subspace is enough to reveal the existence of clusters, although with various degree of separations among these clusters in the different cases. Clusters in cases D2D and D5D require instead at least two and four dimensions, respectively, to be discerned.

3.3 Outcomes of clustering binary data

Results of the simulations based on binary data only ($d_o = 0$) are displayed in Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13, where the average ARI of each of the various tested

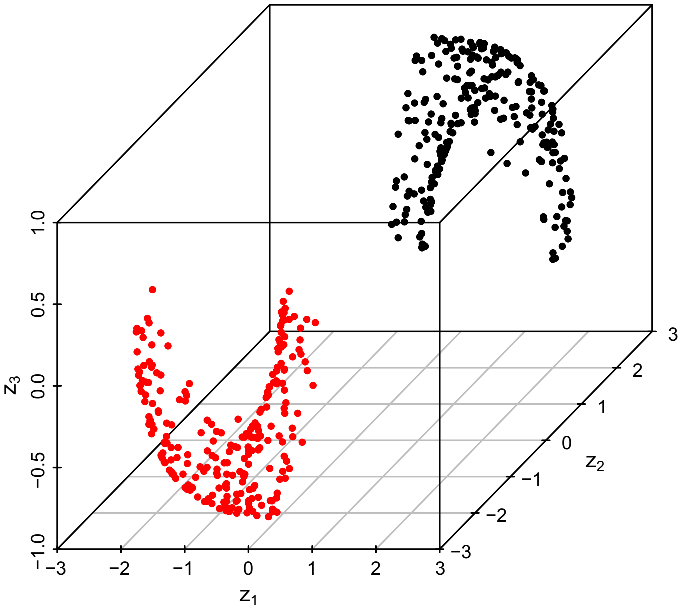


Fig. 2 An illustration of the 3-d analogue of the D5C clustering structure, with non-convex-shaped clusters

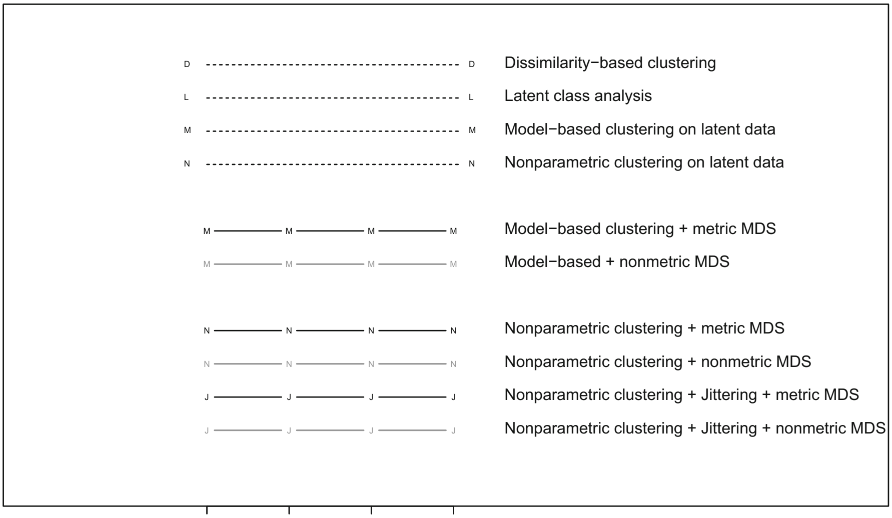

D ---------- D Dissimilarity−based clustering
L ---------- L Latent class analysis
M ---------- M Model−based clustering on latent data
N ---------- N Nonparametric clustering on latent data
M — M — M — M Model−based clustering + metric MDS
M — M — M — M Model−based + nonmetric MDS
N — N — N — N Nonparametric clustering + metric MDS
N — N — N — N Nonparametric clustering + nonmetric MDS
J — J — J — J Nonparametric clustering + Jittering + metric MDS
J — J — J — J Nonparametric clustering + Jittering + nonmetric MDS

Fig. 3 Legend of letters, grey levels and line types adopted in Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15

clustering methods and the true clustering structure is plotted versus L on a $\log_{10}$ scale. The letters and the line types adopted for identifying the various types of clustering are described in Fig. 3; the coding of lines remains the same across all plots. Additional

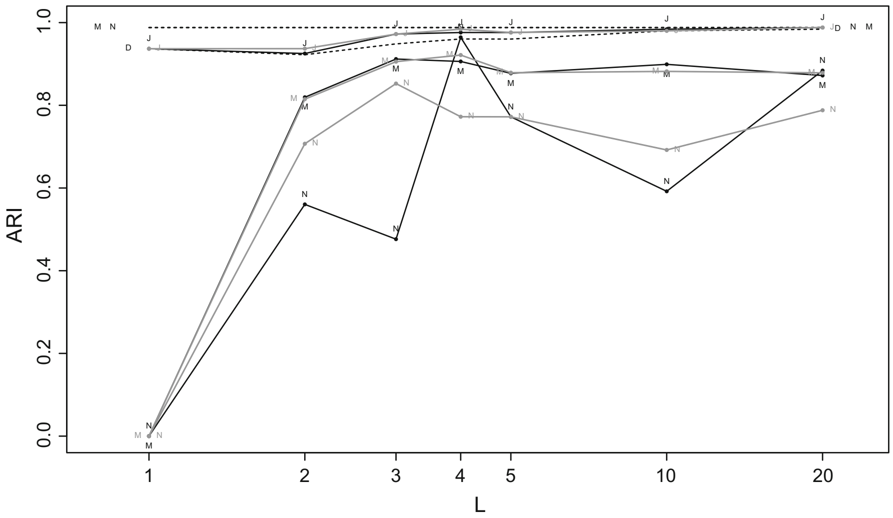


Fig. 4 Average ARI across simulations of detected clustering and true clustering structure in the D1A setting, as *L* varies (on the logarithmic scale)

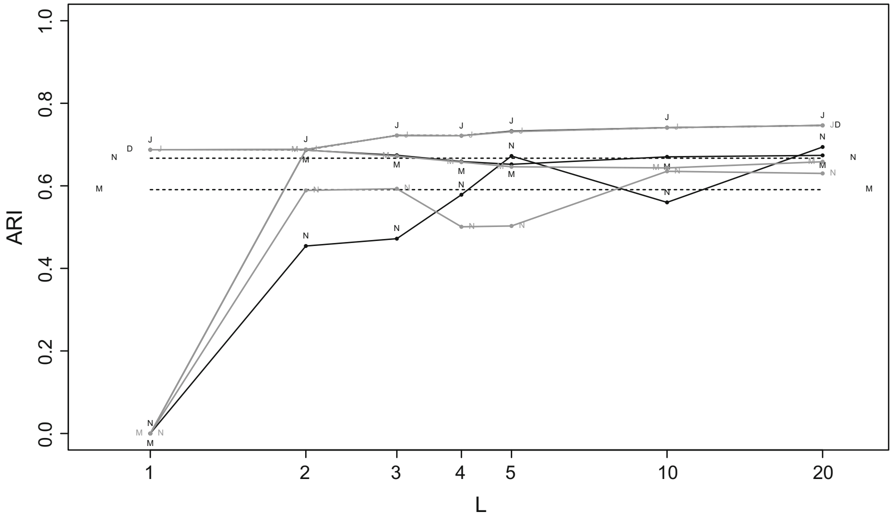


Fig. 5 Average ARI across simulations of detected clustering and true clustering structure in the D1B setting, as *L* varies (on the logarithmic scale)

results reporting the Monte Carlo distributions of the ARI across simulations are included in the supplementary material accompanying this paper.

The general indication emerging from the set of available plots is that the proposed methodology appears to work satisfactorily in most of the cases which have been considered, as long as *L* is at least 2, but an increase of *L* beyond 4 or 5 produces a limited improvment, if any. These indications are broadly in agreement with

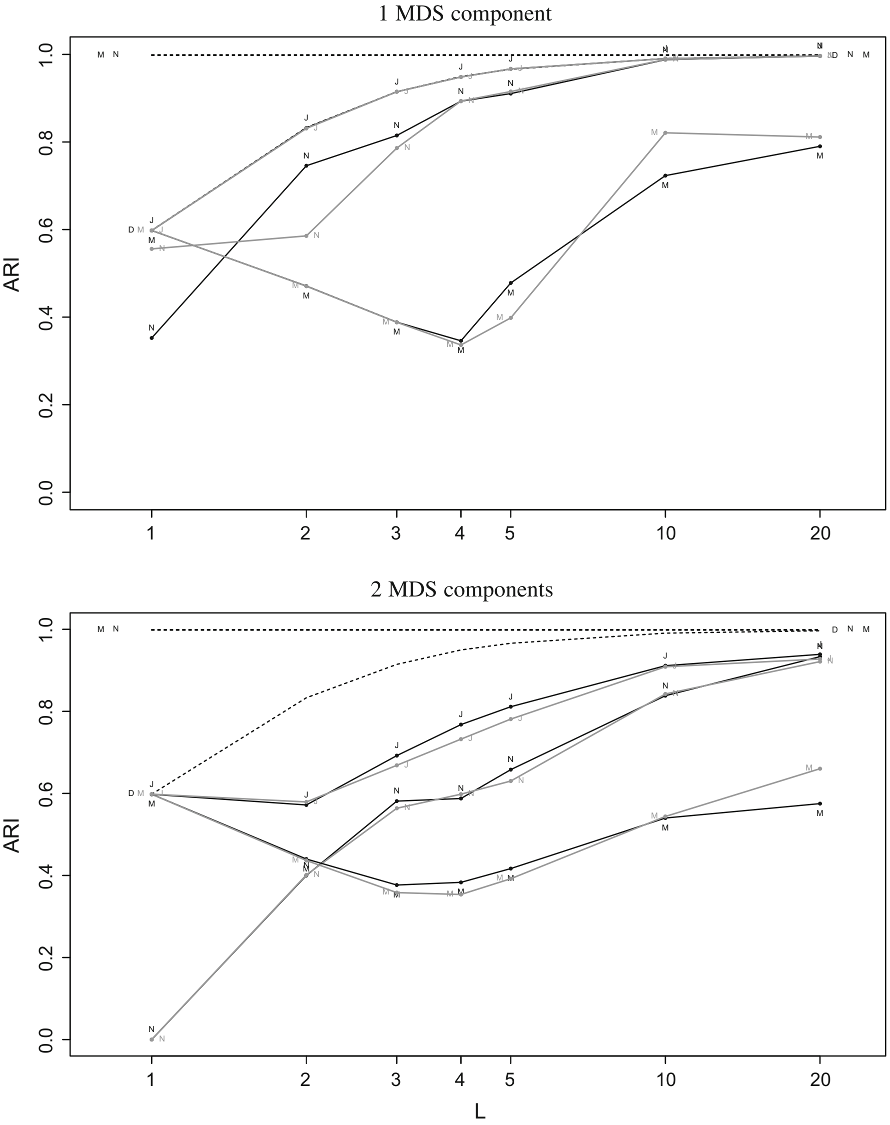


Fig. 6 Average ARI across simulations of detected clustering and true clustering structure in the D2A setting, as L varies (on the logarithmic scale)

those obtained in Sect. 2.1, summarized by Fig. 1. However, there are exceptions to this overall behaviour; several reasons may concur to the lack of a clear-cut pattern. First, as already mentioned, density estimation is problematic in the presence of many ties, as it occurs frequently with binary data. This is especially true for nonparametric clustering, which is then better used in combination with jittering to avoid ties. Therefore, from now on, we shall focus of this variant of the method. Moreover,

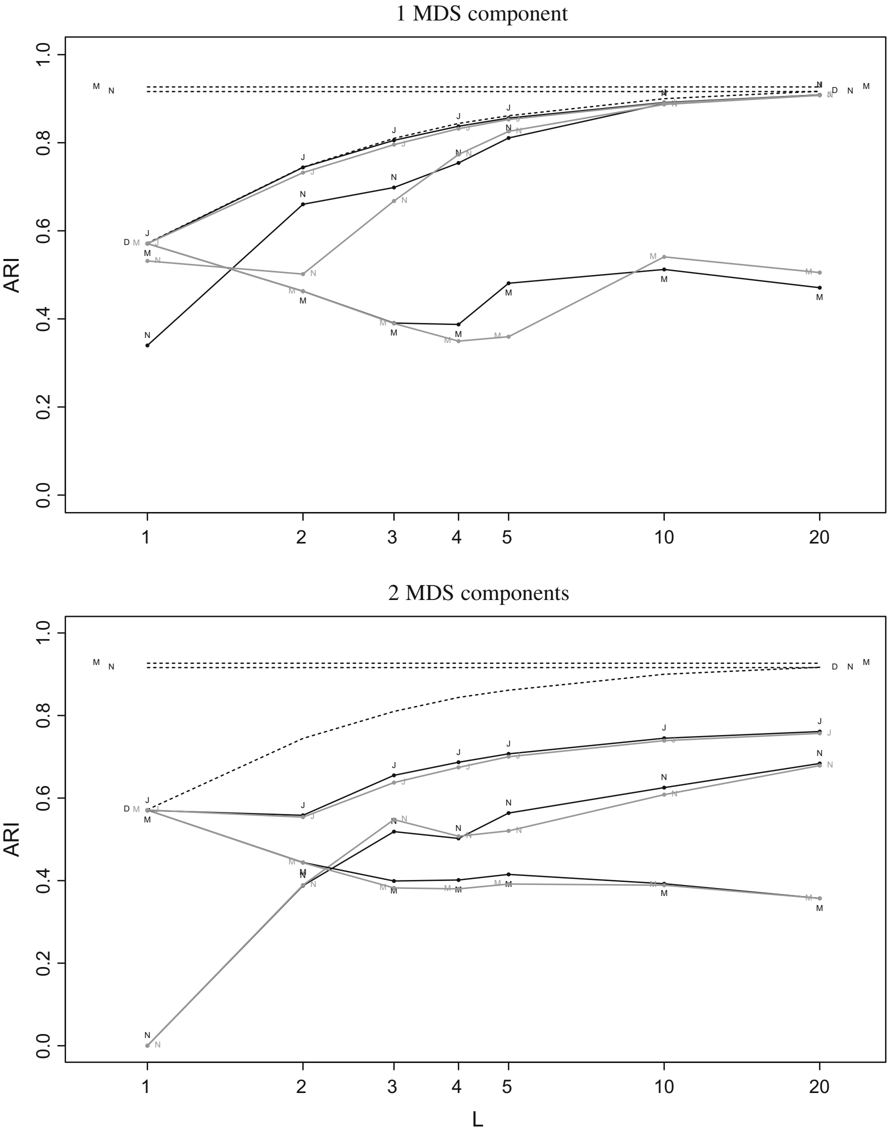


Fig. 7 Average ARI across simulations of detected clustering and true clustering structure in the D2B setting, as L varies (on the logarithmic scale)

some of the simulation settings present a non-trivial clustering structure, and the two density-based approaches have different strengths and weaknesses depending on this underlying structure. For instance, modal clustering is somewhat compromised in the D5D setting, since non-parametric density estimate, in a moderately high-dimensional case with $d = 5$ and $n = 300$, is too poor to cope with $G = 5$ groups. Conversely, the model-based approach is impaired by densities with non-convex contour level

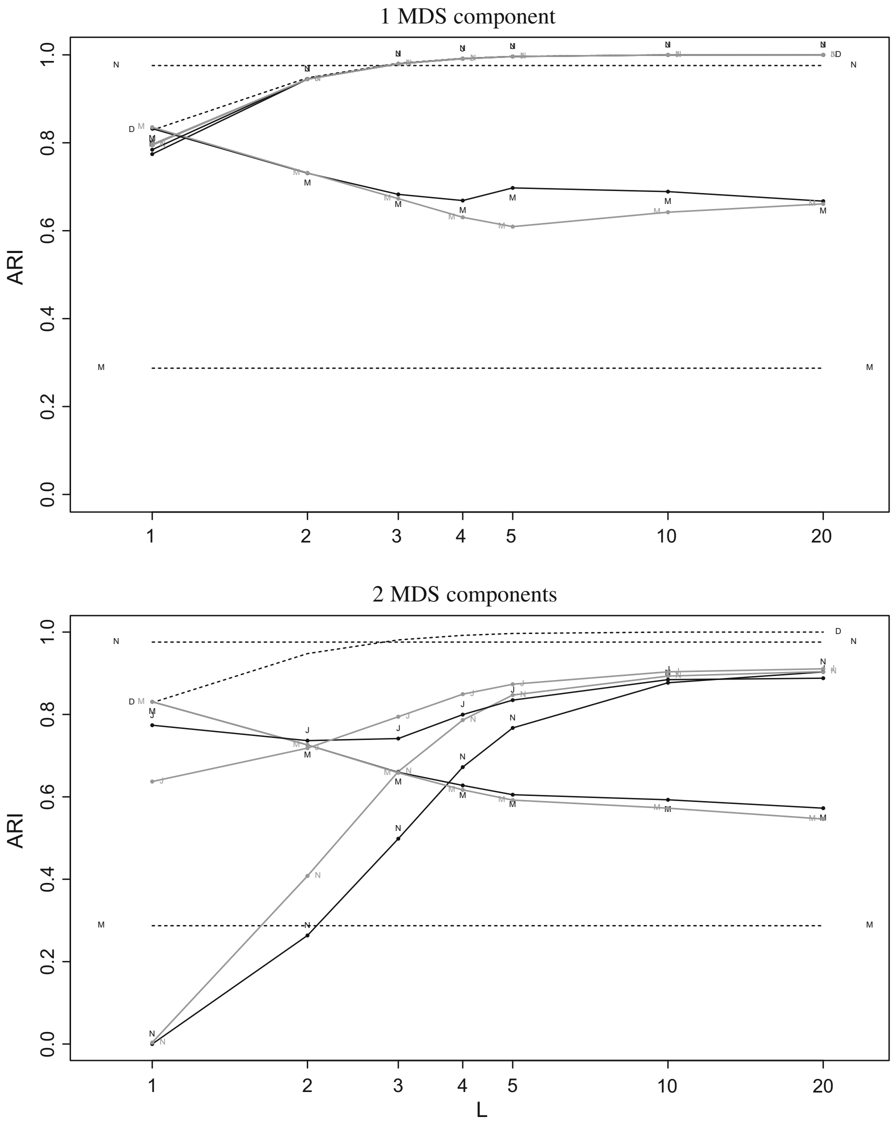
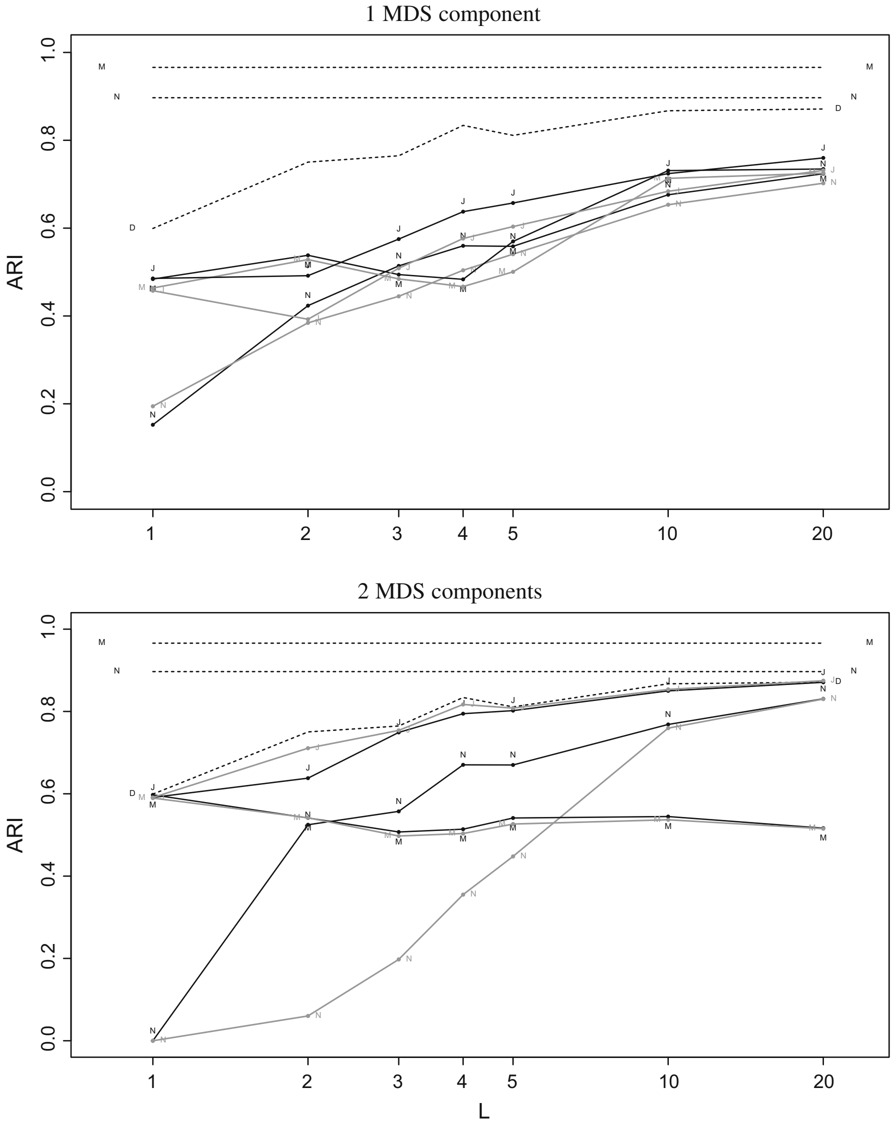


Fig. 8 Average ARI across simulations of detected clustering and true clustering structure in the D2C setting, as *L* varies (on the logarithmic scale)

curves. However, this fact often holds also with normal mixtures, where the average ARI behaviour is hindered by a large varge variability of its distributions (see the Supplementary Material). Additionally, an odd behavior of model-based clustering may derive from an overestimation of the number of clusters, a problem from which the BIC is known to suffer. Also, in making the comparisons, one must consider the relative values of the continuous curves and the corresponding dashed lines or curves, not simply the gross value of the ARI. In other words, for each given case we must



1 MDS component

2 MDS components

Fig. 9 Average ARI across simulations of detected clustering and true clustering structure in the D2D setting, as L varies (on the logarithmic scale)

examine the relative values of the ARI values taking into account the corresponding ARI of the procedures based on the latent variables Z^* (which of course would be not observable in real applications). As a general indication, in the majority of case the J curves lie above the corresponding M curves, if $L > 1$.

In general, results obtained from metric or non-metric MDS are largely equivalent, which warrants a choice based on computational considerations. Conversely, the num-

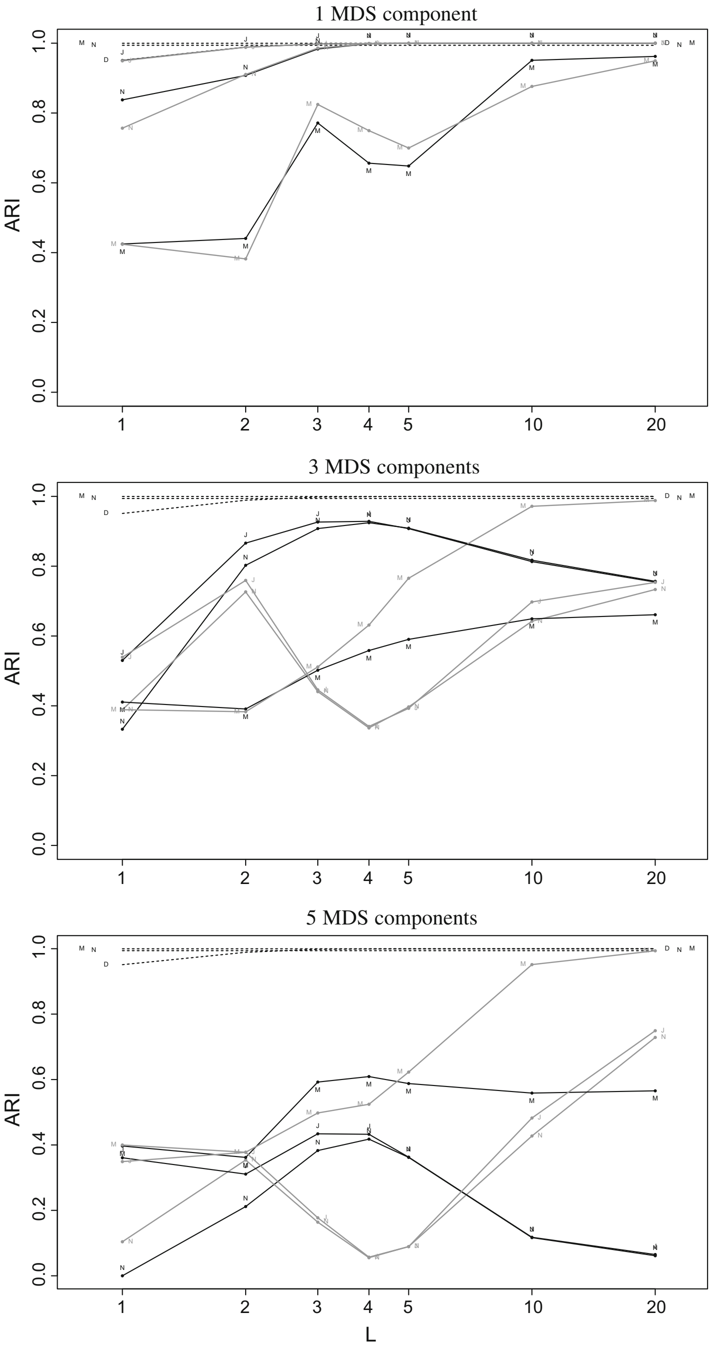
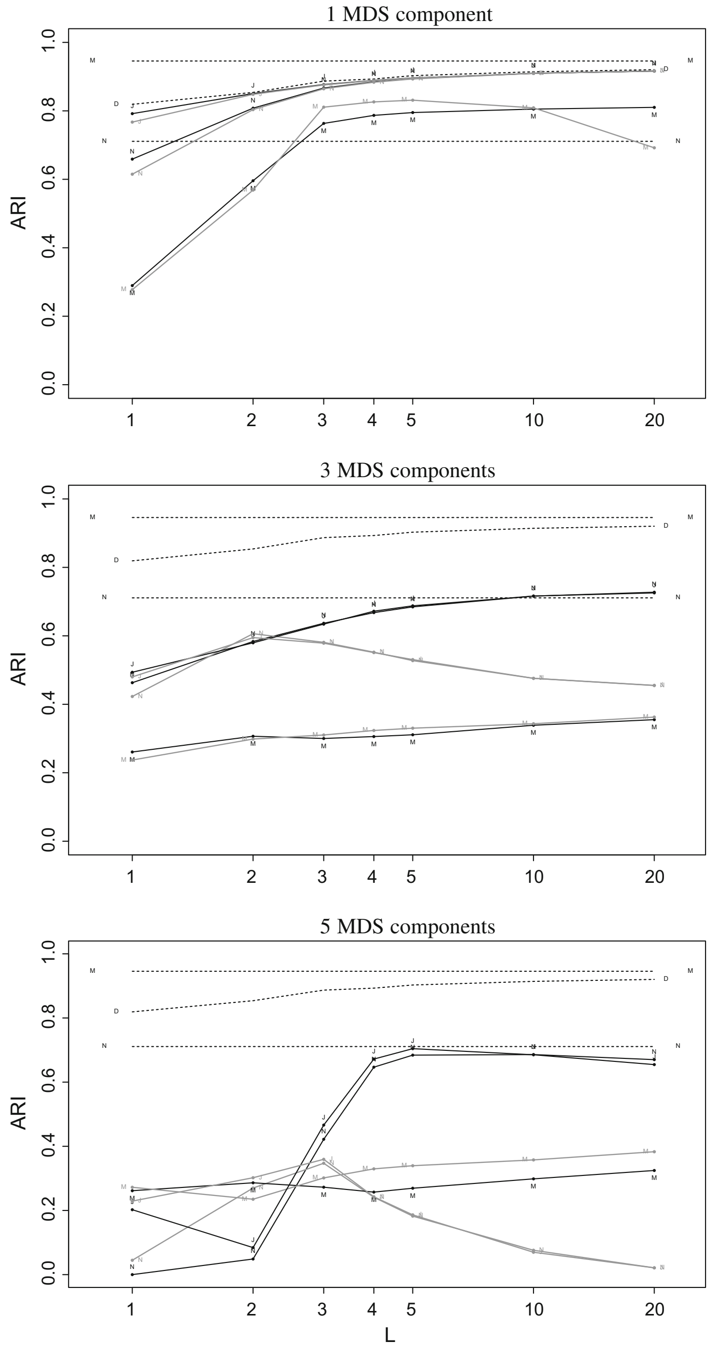


Fig. 10 Average ARI across simulations of detected clustering and true clustering structure in the D5A setting, as L varies (on the logarithmic scale). Results deriving from the use of $d^* = 2$ and $d^* = 4$ are not reported for brevity


1 MDS component

3 MDS components

5 MDS components

Fig. 11 Average ARI across simulations of detected clustering and true clustering structure in the D5B setting, as L varies (on the logarithmic scale). Results deriving from the use of $d^* = 2$ and $d^* = 4$ are not reported for brevity

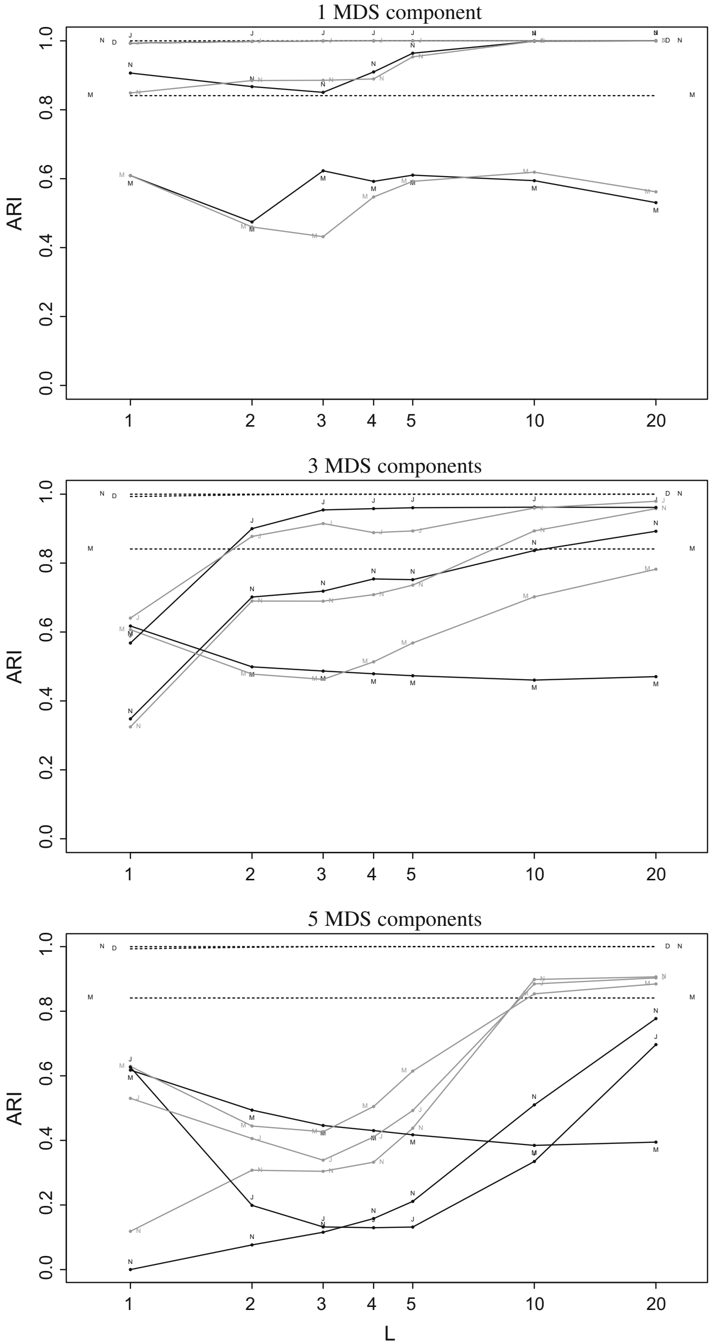


Fig. 12 Average ARI across simulations of detected clustering and true clustering structure in the D5C setting, as L varies (on the logarithmic scale). Results deriving from the use of $d^* = 2$ and $d^* = 4$ are not reported for brevity

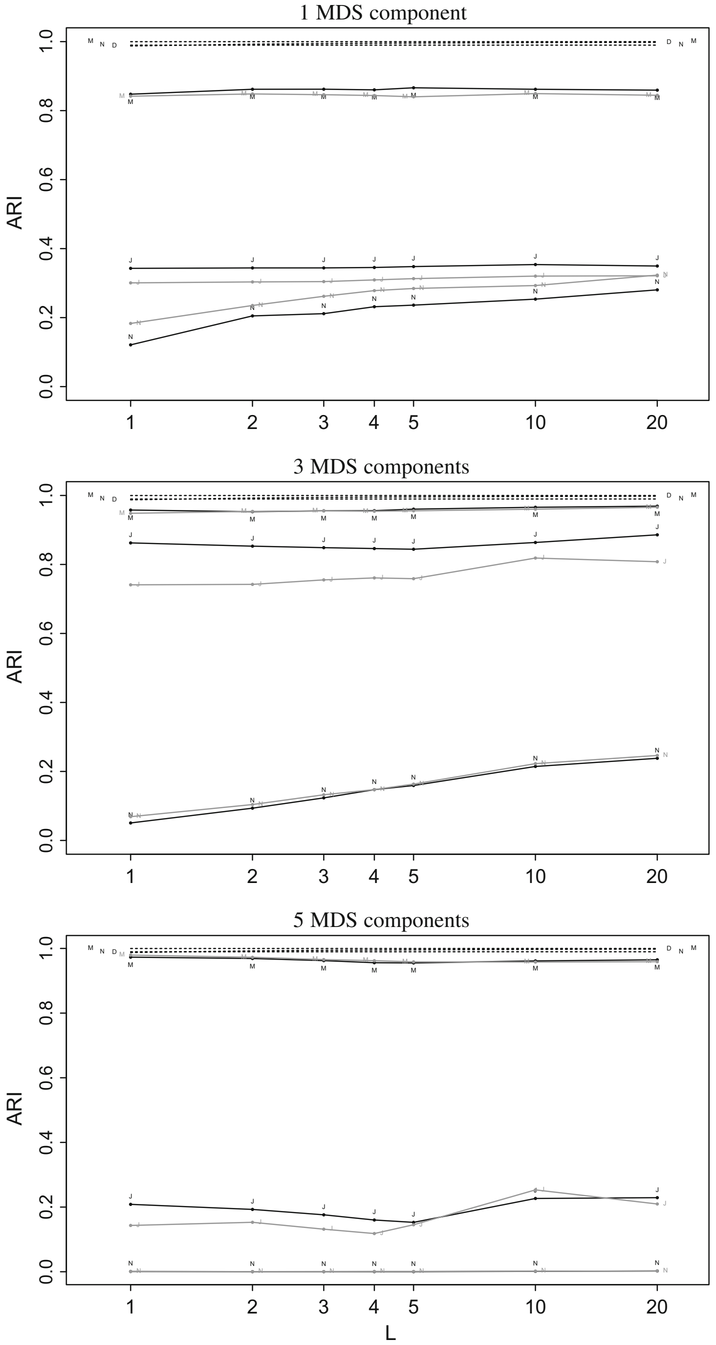


Fig. 13 Average ARI across simulations of detected clustering and true clustering structure in the D5D setting, as L varies (on the logarithmic scale). Results deriving from the use of $d^* = 2$ and $d^* = 4$ are not reported for brevity

ber of selected MDS components has definitely an impact on the quality of partitions. Recall that most of the clustering structures considered are intrinsically unidimensional and the use of multiple MDS variables has then a negative effect. The D2D and D5D cluster settings are conversely multidimensional and the use of two MDS components works well here. In real data applications, one has to resort on background information on the plausible number of latent continuous variables to decide about the number of MDS variables to consider. In general, evidence from the simulations suggests not to exceed a 3- or 4-dimensional MDS configuration, unless there are reasons to believe that the clustering structure is really complex.

3.4 Handling mixed data

Explorations about the use of mixed data refer to two of the settings considered before, D2D and D5D. In the former case, we set $Z_o = Z_1$ and $Z_u = Z_2$, while in the latter case we select the two blocks as $Z_o = (Z_1, Z_2)$ and $Z_u = (Z_3, Z_4, Z_5)$. The choice of these two distributions, within the large set considered earlier, relates to the feature already mentioned of an intrinsically two- and four-dimensional clustering structure, where both the unobservable and the observed components of Z contribute to this structure. Both options referred to as '1' and '2' in Sect. 2.1 have been examined.

Figures 14 and 15 display the average ARI values as the number of binary variables varies and for different values of d^* MDS components; specifically, d^* spans in $\{1, 2\}$ in the first case and $\{1, 2, 3\}$ in the second case. Taking into account the results from the previous simulations, non-parametric clustering has been run only on the jittered data.

In interpreting the plots, one must bear in mind that the two options are not directly comparable side by side, because of the different number of variables involved. Consider, for example, the left and right panels of Fig. 14: in both cases one to two MDS variables have been extracted form the dissimilarity matrix; however, according to option 1 clustering has been performed on these MDS scores, while according to option 2 clustering has been applied on the set of data obtained by merging the MDS scores with the continuous variables Z_o. Thus, the top-left panel in Fig. 14 is based on two continuous variables, like in the bottom-right panel of the same figure and these two plots are broadly comparable. A similar match holds for Fig. 15. From this viewpoint, the two options appear to work quite similarly, with a little superiority of the first option in Fig. 15, yet counterbalanced by a larger variability (see Supplementary Material).

As for the two forms of density-based clustering, the non-parametric one appears again preferable when a small number of MDS components is used, but its advantage decreases or even disappears when the MDS dimension increases inappropriately, since the model-based method is less affected in this circumstance.

Concerning the benchmark clustering methods, K-medoids appears effective in recovering the true groups while latent class analysis tends to behave worse and similarly to model-based clustering. It should be borne in mind that both methods benefit from setting the number of clusters equal to the true one.

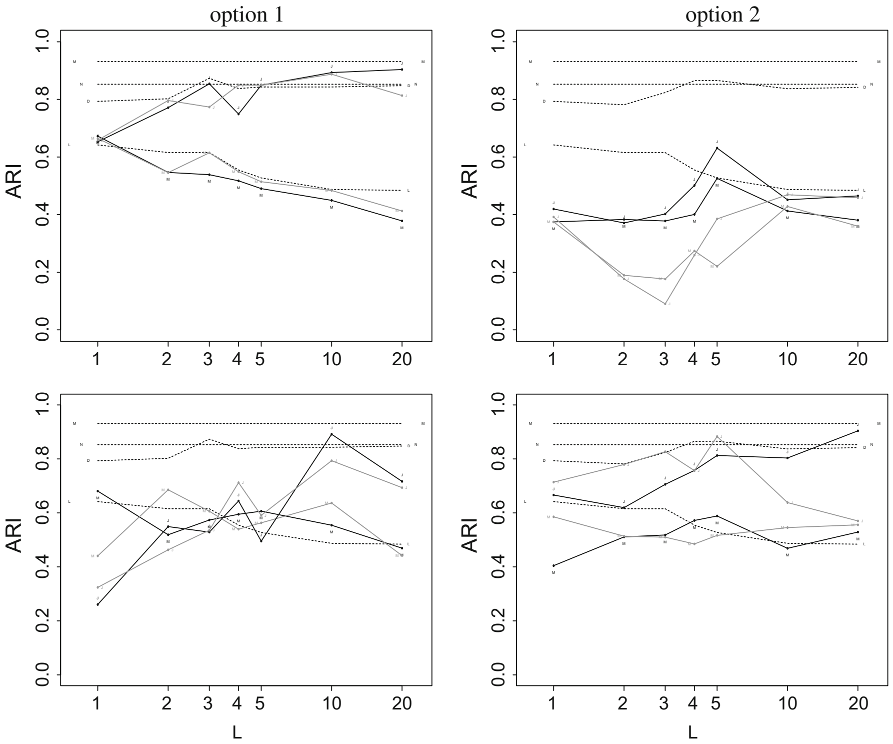


Fig. 14 Average ARI across simulations of detected clustering and true clustering structure in the D2D setting, as the number of binary variables varies. Clustering methods have been applied on the MDS numerical configuration drawn from all the observed variables (*right panels*) or on a data set obtained by merging the continuous observed components with the MDS numerical configuration extracted from the dissimilarity matrix of the non-continuous data (*left panel*). The *top plots* refer to use of one MDS variable, the *bottom* ones to two of them

4 Real data examples

In this section, we provide an illustration of the proposed methodology on some real data examples, which are publicly available at the UCI machine learning repository (Asuncion and Newman 2010).

On each considered set of data, of categorical or mixed type, we first combine the different variables into a single dissimilarity matrix by using a generalization of the Gower coefficient described by Kaufman and Rousseeuw (1990, Sect. 2.6). Then, we apply MDS to extract the continuous latent configuration assumed to underlie the observed data. Mixed data have been handled according to option ‘1’, i.e. the dissimilarity matrix is computed on the whole set of variables. Since we do not have strong reasons to consider the observed variables to be the expression of a multidimensional latent structure, we take $d^* = 1$ in all the examples. Finally, we apply model-based clustering on the continuous score reconstructed, and modal clustering on the same

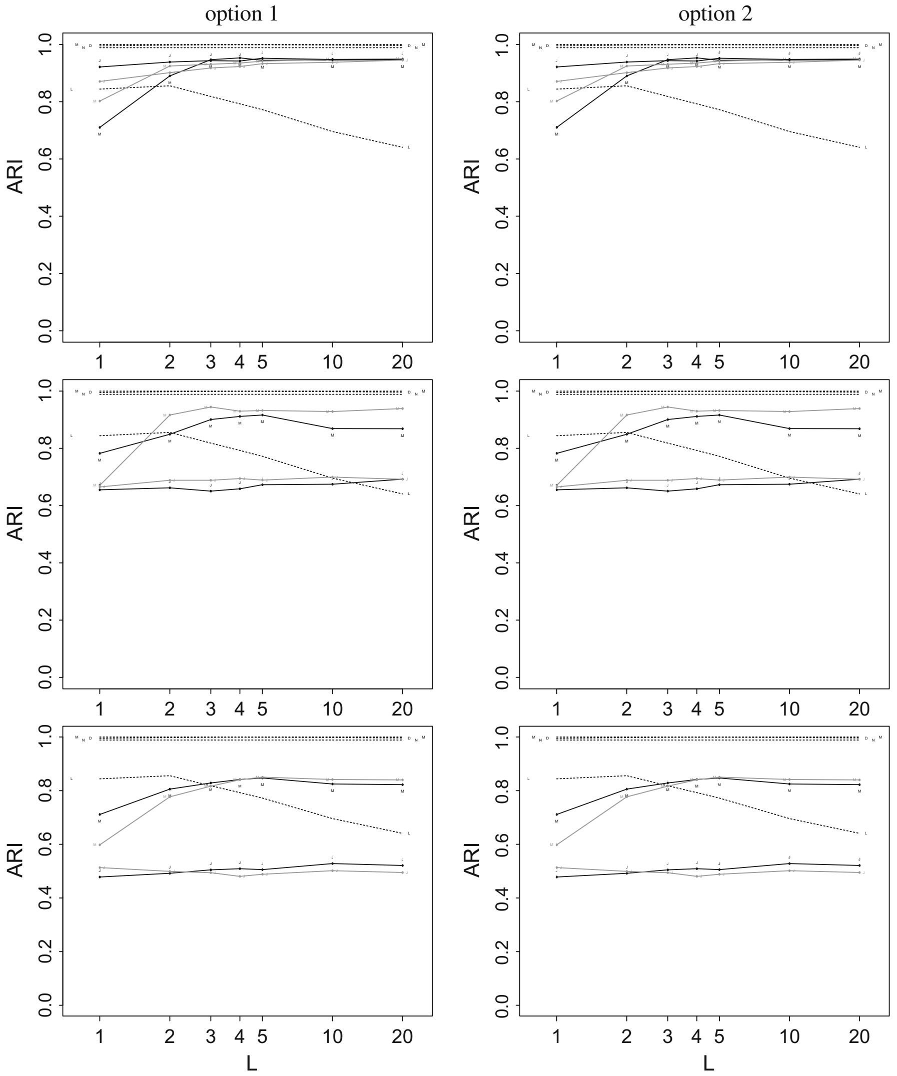


Fig. 15 Average ARI across simulations of detected clustering and true clustering structure in the D5D setting, as the number of binary variables varies. Clustering methods have been applied on the MDS numerical configuration drawn from all the observed variables (*right panels*) or on a data set obtained by merging the continuous observed components with the MDS numerical configuration extracted from the dissimilarity matrix of the non-continuous data (*left panel*). The *plots* in the *top row* refer to use of one MDS variable, the *middle row* to three, the *bottom row* to five

jittered score. Again, we do not force the procedures to detect a prespecified number of clusters, but allow them for estimating it from the data.

For comparison purposes, we also apply latent class analysis for mixed data on the original observations and K-medoids on the computed dissimilarity matrix. Both the benchmarks are given a head start by setting the number of clusters to the actual one.

The first example concerns the votes of 435 US House of Representatives congressmen in 1984 for 16 key issues identified by the Congressional Quartely Almanac. The data represent a straight instance of the framework described in Sect. 2.1, as all the 16 variables are of binary nature, describing the vote in favour or against social or economic questions (e.g. immigration, education spending, crime, budgeting etc.), as a whole providing information about the underlying political position of the congressmen. The aim of clustering is to reconstruct the political side of each voter, either Democrat or Republican.

The second example refers to 15 characteristics of 690 credit card applicants, and can be used to predict the bank decision about either approving or rejecting the credit card application. The data represent quite a challenging example, as they comprise a heterogeneous set of variables: categorical both dichotomous and polichotomous, and continuous variables. Additionally, some instances of the latter ones are of peculiar type, as most of their probability mass is concentrated on a very few values. Also, two continuous variables present a very skew distribution, which has been treated by taking the logarithmic transformation.

The third example includes 14 demographical, physical, and clinical attributes collected by the Hungarian Institute of Cardiology (under the responsibility of Andras Janosi) within a experimental study involving 294 individuals. The observed variables are used to diagnostic the presence or absence of some heart diseases. Similarly to the credit card example, there are both dichotomous and polytomous nominal variables, and continuous ones.

Before proceeding with the analysis, we have first removed from the datasets the variables presenting a very large proportion of missing values and, afterwards, all the observations presenting missing values. In fact, this step is only required for the application of latent class analysis, at least operationally.

Results are displayed in Tables 1, 2 and 3. While looking at the values of the adjusted Rand index might convey a discouraging message about the effectiveness of the two density-based methods, a careful observation of the cross-frequency tables gives a general indication of success in the identification of the true clustering. Low values of the ARI are associated to the cases where the two methods over-estimate the number of groups (Table 2 for modal clustering and Tables 1, 2 for model-based clustering) and an appropriate aggregation of these groups into two macro-clusters leads to quite satisfactory results. In fact, these partitions are consistent with the assumption about the existence of a continuous score underlying the observed variables: in the Congressional voting data, for example, such score may be interpreted as the political ideology of each Congressman. In this perspective, model-based clusters can be read as three different levels of conservatism, a moderate position, and three different levels of progressivism. Similarly, in the credit example, different clusters correspond to different levels of credit solvency.

As far as the reference methods concern, clustering based on the dissimilarities show a satisfactory behaviour in all the examples, as the misclassification error keeps lower than 20 % in all the examples. In fact, we should recall that the good performance are favoured by the prespecified number of clusters set equal to the actual one.

Latent class analysis shows satisfactory results in the third example only (Table 3) while it fails resoundingly in the two other datasets. In fact, while an unarguable

Table 1 US Congressional voting data: true groups (democratic, republican) compared with clusters identified by modal clustering, model-based clustering, dissimilarity-based clustering and latent class analysis

Modal		Model-based							K-medoids		Latent class	
1	2	1	2	3	4	5	6	7	1	2	1	2
Democratic												
101	23	35	31	30	18	8	1	1	102	22	66	58
Republican												
4	104	0	0	4	18	30	29	27	9	99	54	54
ARI = 0.587		ARI = 0.170							ARI = 0.535		ARI = 0.000	

Table 2 Credit card data: true groups (card denial, card approval) compared with clusters identified by modal clustering, model-based clustering, dissimilarity-based clustering and latent class analysis

Modal				Model-based				K-medoids		Latent class	
1	2	3	4	1	2	3	4	1	2	1	2
Denyal											
86	171	78	22	75	187	71	24	294	63	118	239
Approval											
10	37	68	181	8	45	57	186	64	232	153	143
ARI = 0.223				ARI = 0.232				ARI = 0.372		ARI = 0.039	

Table 3 Heart disease data: true groups (absence or presence of heart disease) compared with clusters identified by modal clustering, model-based clustering, dissimilarity-based clustering and latent class analysis

Modal		Model-based		K-medoids		Latent class	
1	2	1	2	1	2	1	2
Absence							
148	15	146	17	145	18	142	21
Presence							
34	64	30	68	30	68	25	73
ARI = 0.382		ARI = 0.403		ARI = 0.393		ARI = 0.414	

strength of this method is that it is the only existing clustering approach for mixed data which relies on the data distribution, we have experienced a certain instability of the Expectation Maximization algorithm used to estimate the parameters, probably due also to the concentration of the probability mass on a very few values. Certainly one could improve the estimation by regularizing the component-specific parameter estimates, but our feeling is that the specification of such a complicated model does not worth, on the whole, when the simple tandem-approach here discussed tends to provide better results.

5 Concluding remarks

The overall indication emerging from our numerical exploration is that the proposed technique appears to work quite satisfactorily, provided some requisites are fulfilled.

For any underlying latent variable, it is advisable to have available a number $L > 1$ of binary representations. As L increases, there is generally an improvement in performance; however, beyond $L = 4$ or 5, there is generally little further gain. The number of constructed MDS scores should ideally not exceed the number of underlying continuous latent variables, or at least not by a sizeable number. In general, it seems unlikely that more than three or four MDS variables are required, except in very complex situations. The two options for handling mixed data work in a broadly similar way, and the choice between them is a matter of preference.

The above indications are formulated as if we knew the number of latent variables, which in practical work is not the case, at least not exactly. In practice one must take into account subject-matter considerations to formulate a judgement on this aspect, like for instance in the illustrative applications of Sect. 4.

The two variant forms of MDS worked very similarly. Therefore the choice can be addressed on the basis of computational convenience.

As for the relative performance of the two density-based clustering methods, the modal one performs better than the model-based one in the majority of cases, especially so if $L > 1$ and the binary variables are jittered. The model-based version is less affected than the non-parametric by a misjudgement of the number of MDS variables.

All our numerical work and the above comments refer to binary observations. Consideration of other forms of categorical variables, possibly of ordered type, or discrete variables would have increased further the already large amount of numerical outcomes. Since the case of binary variables is qualitatively the one more distant from continuous variables, it seems reasonable to expect that the indications obtained here hold in a broad sense also in these other cases.

References

Anderlucci L, Hennig C (2014) Clustering of categorical data: a comparison of a model- based and a distance-based approach. Commun Stat Theory Methods 43(4):704–721

Arabie P, Hubert L (1994) Cluster analysis in marketing research. In: Bagozzi R (ed) Handbook of marketing research. Blackwell, Oxford

Asuncion A, Newman D (2010) UCI machine learning repository. School of Information and Computer Sciences, University of California, Irvine

Azzalini A, Menardi G (2014) Clustering via nonparametric density estimation: the R package pdfCluster. J Stat Softw 57(11):1–26

Azzalini A, Torelli N (2007) Clustering via nonparametric density estimation. Stat Comput 17:71–80

Bartholomew DJ (1980) Factor analysis for categorical data. J R Stat Soc Series B 42:293–321

Bartholomew DJ, Knott M (1999) Latent variable models and factor analysis, 2nd edn. Arnold Publisher, London

Browne RP, McNicholas PD (2012) Model-based clustering, classification, and discriminant analysis of data with mixed type. J Stat Plan Inference 142:2976–2984

Fraley C, Raftery A (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput J 41:578–588

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97:611–631

Fraley C, Raftery AE, Murphy B, Scrucca L (2012) Mclust version 4 for R: normal mixture modeling and model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington

Fukunaga K, Hostetler LD (1975) The estimation of the gradient of a density function, with application in pattern recognition. IEEE Trans Inf Theory 21:32–40

Goodman LA (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61:215–231

Gruen B, Leisch F (2008) FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J Stat Softw 28(4):1–35. http://www.jstatsoft.org/v28/i04/

Hartigan JA (1975) Clustering algorithms. Wiley, New York

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Hunt L, Jorgensen M (2003) Mixture model clustering for mixed data with missing information. Comput Stat Data Anal 41:429–440

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Leisch F (2004) FlexMix: a general framework for finite mixture models and latent class regression in R. J Stat Softw 11(8):1–18. http://www.jstatsoft.org/v11/i08/

Lin TI (2010) Robust mixture modeling using multivariate skew t distributions. Stat Comput 20(3):343–356

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2013) Cluster: cluster analysis basics and extensions. R package version 1.14.4

Marbac M, Biernacki C, Vandewalle V (2015) Model-based clustering for conditionally correlated categorical data. J Classif 32(2):145–175

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press, Cambridge

Menardi G, Azzalini A (2014) An advancement in clustering via nonparametric density estimation. Stat Comput 24:753–767

Oh M, Raftery AE (1998) Model-based clustering with dissimilarities: a Bayesian approach. J Comput Graph Stat 16:559–585

R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0

Stuetzle W (2003) Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. J Classif 20:25–47

Stuetzle W, Nugent R (2010) A generalized single linkage method for estimating the cluster tree of a density. J Comput Graph Stat 19:397–418

Tzeng J, Lu HH, Li WH (2008) Multidimensional scaling for large genomic data sets. BMC Bioinformatics 9(1):179

Venables VN, Ripley BD (2002) Modern applied statistics with S. Springer, New York. http://www.stats.ox.ac.uk/pub/MASS4

Vermunt JK, Magidson J (2002) Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL (eds) Applied latent class analysis. Cambridge University Press, Cambridge, pp 89–106

Wishart D (1969) Mode analysis: a generalization of nearest neighbor which reduces chaining effects. In: Cole AJ (ed) Numerical taxonomy. Academic Press, Cambridge, pp 282–308

Wolfe JH (1970) Pattern clustering by multivariate mixture analysis. Multivar Behav Res 5:329–350