

A Comparison Between Methods to Cluster Mixed-Type Data: Gaussian Mixtures Versus Gower Distance



Monia Ranalli and Roberto Rocci

Abstract In this paper, we compare through a simulation study two approaches to cluster mixed-type data, where some variables are continuous and some others ordinal. The first is model-based, according to which the variables are assumed to follow a Gaussian mixture model, where, as regards the ordinal variables, it is only partially observed. In order to overcome computational issues, the parameter estimation is carried out through an EM-like algorithm maximizing a composite log-likelihood based on low-dimensional margins. In the second approach, the Gower distance matrix is computed, then the PAM algorithm is used for clustering.

Keywords Mixture models · Composite likelihood · EM algorithm · Mixed-type data · Gower's distance · PAM algorithm

1 Introduction

The aim of cluster analysis is to partition the data into meaningful homogeneous groups which should differ considerably from each other. The problem is made more difficult by the presence of mixed-type data: ordinal and continuous variables. In order to find a solution, mainly two different approaches exist, based on a model describing the data generation process or a distance able to capture the dissimilarity between two entities. Before to summarize the main features of the two approaches, let us specify that when we use the word categorical data, we are still referring to the ordinal variables. Following the definition given in [1], ordinal variables are categorical variables with ordered categories.

As regards the model-based approach, the literature on clustering for continuous data is rich and wide; the most commonly clustering model-based used is the finite

M. Ranalli (✉) · R. Rocci

Department of Statistics, Sapienza University of Rome, Rome, Italy
e-mail: monia.ranalli@uniroma1.it

R. Rocci

e-mail: roberto.rocci@uniroma1.it

© Springer Nature Switzerland AG 2021

S. Balzano et al. (eds.), *Statistical Learning and Modeling in Data Analysis*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-69944-4_17

mixture of Gaussians [17]). Differently, that one developed for categorical data is still limited. In the Underlying Response Variable (URV), mainly developed in the SEM framework (see, e.g., [11, 14, 20] approach, the ordinal variables are seen as a discretization of continuous latent variables jointly distributed as a finite mixture (see [5, 16, 23]). However, this makes the maximum likelihood estimation rather complex because it requires the computation of many high-dimensional integrals. The problem is usually solved by approximating the likelihood function by a surrogate one. In this regard we mention some useful surrogate functions, such as the variational likelihood [7] or the composite likelihood [21, 23, 24]. The problem arises when we consider the joint distribution between continuous and ordinal variables. By assuming the local independence assumption, the issue can be easily solved by factorizing the joint density into the product of univariate marginals. However, this assumption is unrealistic and too restrictive.

Following the URV approach, [5, 23] proposed a model according to which the variables follow a Gaussian mixture model, where some variables, the ordinal ones, are only partially observed through their discretization. As a side note, at this stage, nominal variables cannot be included in the model, since there is no type of proximity among the unordered categories.

Besides these methods, there are others based on the Gower's distance [8]. This is computed as the average of partial dissimilarities across subjects (or entities), where the type of partial dissimilarity used depends on the specific type of the variable. To cluster the data then a k -medoids algorithm can be used (PAM algorithm, [13, 25]). However, these clustering methods are not the only ones existing in literature. Indeed there are many techniques for mixed-type data and many reviews. See, for example, [2, 6, 10]. Comparing clustering techniques is extremely useful and benchmarking in cluster analysis has been increasing. A good discussion on it can be found in [18].

The paper aims at exploring and comparing the behavior of the mixture model for mixed-type data with the distance-based methods, and some more naive approaches, according to which ordinal data are treated as metric.

The plan of the paper is as follows. In Sect. 2, we describe the model-based approach to cluster mixed-type data. The Gower distance method followed by the PAM algorithm is described in Sect. 3. In Sect. 4, we compare these clustering techniques through a simulation study. In the last section, some concluding remarks are pointed out.

2 The Model-Based Approach

Let $\mathbf{x} = [x_1, \dots, x_O]'$ and $\mathbf{y}^{\bar{O}} = [y_{O+1}, \dots, y_P]'$ be O ordinal and $\bar{O} = P - O$ continuous variables, respectively. The associated categories for each ordinal variable are denoted by $c_i = 1, 2, \dots, C_i$ with $i = 1, 2, \dots, O$.

Following the Underlying Response Variable (URV) approach, the ordinal variables \mathbf{x} are considered as a categorization of a continuous multivariate latent variable

$\mathbf{y}^O = [y_1, \dots, y_O]'$. The latent relationship between \mathbf{x} and \mathbf{y}^O is explained by the threshold model,

$$x_i = c_i \Leftrightarrow \gamma_{c_i-1}^{(i)} \leq y_i < \gamma_{c_i}^{(i)},$$

where $-\infty = \gamma_0^{(i)} < \gamma_1^{(i)} < \dots < \gamma_{C_i-1}^{(i)} < \gamma_{C_i}^{(i)} = +\infty$ are the thresholds defining the C_i categories collected in a set Γ . To accommodate both cluster structure and dependence within the groups, we assume that $\mathbf{y} = [\mathbf{y}^O, \mathbf{y}^{\bar{O}}]'$ follows a heteroscedastic Gaussian mixture, $f(\mathbf{y}) = \sum_{g=1}^G \tau_g \phi_p(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where the τ_g 's are the mixing weights and $\phi_p(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a P -variate normal distribution with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$.

Let us set $\boldsymbol{\psi} = \{\tau_1, \dots, \tau_G, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_G, \Gamma\} \in \Psi$, where Ψ is the parameter space. For a random i.i.d. sample of size N : $(\mathbf{x}_1, \mathbf{y}_1^{\bar{O}}), \dots, (\mathbf{x}_N, \mathbf{y}_N^{\bar{O}})$, the log-likelihood is

$$\ell(\boldsymbol{\psi}) = \sum_{n=1}^N \log \left[\sum_{g=1}^G \tau_g \phi_{\bar{O}}(\mathbf{y}_n^{\bar{O}}; \boldsymbol{\mu}_g^{\bar{O}}, \boldsymbol{\Sigma}_g^{\bar{O}\bar{O}}) \pi_n(\boldsymbol{\mu}_{n;g}^{O|\bar{O}}, \boldsymbol{\Sigma}_g^{O|\bar{O}}, \Gamma) \right], \quad (1)$$

where with obvious notation

$$\begin{aligned} \pi_n(\boldsymbol{\mu}_{n;g}^{O|\bar{O}}, \boldsymbol{\Sigma}_g^{O|\bar{O}}, \Gamma) &= \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \dots \int_{\gamma_{c_O-1}^{(O)}}^{\gamma_{c_O}^{(O)}} \phi_O(\mathbf{u}; \boldsymbol{\mu}_{n;g}^{O|\bar{O}}, \boldsymbol{\Sigma}_g^{O|\bar{O}}) d\mathbf{u} \\ \boldsymbol{\mu}_{n;g}^{O|\bar{O}} &= \boldsymbol{\mu}_g^O + \boldsymbol{\Sigma}_g^{O\bar{O}} (\boldsymbol{\Sigma}_g^{\bar{O}\bar{O}})^{-1} (\mathbf{y}_n^{\bar{O}} - \boldsymbol{\mu}_g^{\bar{O}}), \\ \boldsymbol{\Sigma}_g^{O|\bar{O}} &= \boldsymbol{\Sigma}_g^{OO} - \boldsymbol{\Sigma}_g^{O\bar{O}} (\boldsymbol{\Sigma}_g^{\bar{O}\bar{O}})^{-1} \boldsymbol{\Sigma}_g^{\bar{O}O}. \end{aligned}$$

$\pi_n(\boldsymbol{\mu}_{n;g}^{O|\bar{O}}, \boldsymbol{\Sigma}_g^{O|\bar{O}}, \Gamma)$ is the conditional joint probability of response pattern $\mathbf{x}_n = (c_{1;n}, \dots, c_{O;n})$ given the cluster g and the values $\mathbf{y}_n^{\bar{O}}$ for the continuous variables. Finally, τ_g is the probability of belonging to group g subject to $\tau_g > 0$ and $\sum_{g=1}^G \tau_g = 1$.

The presence of multidimensional integrals makes the maximum likelihood estimation computationally demanding and infeasible as the number of ordinal variables increases. To overcome this, a composite likelihood approach is adopted [15]. It allows us to simplify the problem by replacing the full likelihood with a surrogate function. As suggested in [21, 23, 24] within a similar context, the full log-likelihood could be replaced by $O(O - 1)/2$ marginal distributions each of them composed of a pair of ordinal variables and the \bar{O} continuous variables. In this way, the computational complexity is greatly decreased because the evaluation of the new function requires the calculation of bivariate, rather than O -variate, integrals. This leads to the following surrogate function

$$c\ell(\boldsymbol{\psi}) = \sum_{n=1}^N \sum_{i=1}^{O-1} \sum_{j=i+1}^O \sum_{c_i=1}^{C_i} \sum_{c_j=1}^{C_j} \delta_{nc_i c_j}^{(ij)} \log \left[\sum_{g=1}^G \tau_g \phi_{\bar{O}}(\mathbf{y}_n^{\bar{O}}; \boldsymbol{\mu}_g^{\bar{O}}, \boldsymbol{\Sigma}_g^{\bar{O}\bar{O}}) \pi_{c_i c_j}^{(ij|\bar{O})}(\boldsymbol{\mu}_{n;g}^{(ij|\bar{O})}, \boldsymbol{\Sigma}_g^{(ij|\bar{O})}, \boldsymbol{\Gamma}^{(ij)}) \right],$$

where $\delta_{nc_i c_j}^{(ij)}$ is a dummy variable assuming 1 if the n th observation presents the combination of categories c_i and c_j for variables x_i and x_j , respectively, 0 otherwise; $\pi_{c_i c_j}^{(ij|\bar{O})}(\boldsymbol{\mu}_{n;g}^{(ij|\bar{O})}, \boldsymbol{\Sigma}_g^{(ij|\bar{O})}, \boldsymbol{\Gamma}^{(ij)})$ is the conditional probability of the pair $(x_i = c_i, x_j = c_j)$ obtained by integrating the density of a bivariate normal distribution with parameters $(\boldsymbol{\mu}_{n;g}^{(ij|\bar{O})}, \boldsymbol{\Sigma}_g^{(ij|\bar{O})})$ between the corresponding threshold parameters contained in the set $\boldsymbol{\Gamma}^{(ij)}$. The parameter estimates are carried out through an EM-like algorithm that works in the same manner as the standard EM. Likewise, it suffers from the problem of local optima.

In the simulation study, the partition has been initialized randomly. The output of a mixture model for continuous data has been considered as a good rational starting point for the component parameters. On the other hand, the initial values for the thresholds have been computed as follows: for each variable, we have considered the empirical relative frequency of each category and then we have minimized the quadratic difference between this frequency and the corresponding quantile of the mixture.

2.1 Classification, Model Selection, and Identifiability

The classification is obtained by assigning the observations to the component with the maximum scaled composite fit, i.e., the CMAP criterion [23, 24]. As regards model selection, the best model is chosen by minimizing the composite version of penalized likelihood selection criteria like BIC or CLC (see [22] and references therein). Finally, as regards identifiability, adopting a composite likelihood approach, the sufficient condition should be reformulated by investigating the Godambe information matrix, that is, the analogous of the information matrix. However, as far as we know, such modification has not been formally investigated yet. About the necessary condition, we note that the number of essential parameters in the block of ordinal variables equals the number of parameters of a log-linear model with only two-factor interaction terms. Thus, it means that we can estimate a lower number of parameters compared to a full maximum likelihood approach. Furthermore, under the underlying response variable approach, the means and the variances of the latent variables are set to 0 and 1, respectively, because they are not identified. This identification constraint individualizes uniquely the mixture components (ignoring the label switching problem), as well described in [19]. This is sufficient to estimate both thresholds and component parameters if all the observed variables have three

categories at least and when groups are known. Given the particular structure of the mean vectors and covariance matrices, it is preferable to adopt an alternative, but equivalent, parametrization. This is analogous to that one used by [12]; it consists in setting the first two thresholds to 0 and 1, respectively, without constraining means and variances. This means that there is a one-to-one correspondence between the two sets of parameters. If there is a binary variable, then the variance of the corresponding latent variable is set equal to 1 (while its mean should be still kept free).

3 The Gower Distance Method

Gower distance is computed as the average of partial dissimilarities across observations (subjects or objects), where the computation of the partial dissimilarities depends on the specific type of the variable. For the continuous variables, a range-normalized Manhattan distance is used; for the ordinal variables, they are first ranked, then Manhattan distance is used with a special adjustment for ties. Then, a weighted sum is calculated to create the final distance matrix. However, it is important to note that as the sample size increases, its storage becomes infeasible.

One of the popular partitioning algorithms for mixed-type data is k -medoids (PAM algorithm [13, 25]), which is based on the Gower's distance. The k -means and the PAM algorithm are briefly described in Sects. 3.1 and 3.2. Both suffer from reaching local optima; indeed different initializations can lead to different partitions. Finally, the choice of the number of cluster can be made based on different criteria; the most commonly used is choosing the number of clusters corresponding to an elbow of the scree plot of the within deviance versus the number of clusters.

3.1 k -means

By letting $\mathbf{X} = \{\mathbf{x}_n : n = 1, \dots, N\}$ be the sample of P -dimensional observations, k -means is based on the minimization of the loss function

$$\ell_{km}(\psi, \mathbf{Z}; \mathbf{X}) = \sum_{n=1}^N \sum_{g=1}^G z_{ng} d^2(\mathbf{x}_n, \boldsymbol{\mu}_g), \quad (2)$$

where $d^2(\mathbf{x}_n, \boldsymbol{\mu}_g)$ is the squared distance, usually the classical unweighted Euclidean between \mathbf{x}_n and $\boldsymbol{\mu}_g$, $\mathbf{Z} = [z_{ng}]$ is a binary membership matrix, with rows that sum to 1, such that $z_{ng} = 1$ if observation n belongs to cluster g and 0 otherwise, and $\psi = \{\mu_1, \dots, \mu_G\}$ is the set of cluster centroids.

3.2 *k-medoids*

The PAM algorithm is an iterative algorithm composed of the following steps:

1. choose k random entities to become the medoids;
2. assign every entity to its closest medoid using the distance matrix computed;
3. for each cluster, the observation with the lowest average distance is re-assigned as the medoid;
4. if at least one medoid has changed, repeat steps 2–4, otherwise the algorithm reaches convergence.

Both k -means and k -medoids are partitioning algorithms and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. However, k -means has cluster centers defined by Euclidean distance (i.e., centroids), while cluster centers for PAM are restricted to be the observations themselves (i.e., medoids). Furthermore, k -medoids can be based on an arbitrary dissimilarity matrix. As a consequence, k -medoids is more robust because it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances.

4 Simulation Study

To evaluate empirically the performance of the different clustering methods, a simulation study has been conducted. We compare: a mixture of Gaussians treating all variables as continuous (Naive), a mixture model for mixed-type data (Mixed), PAM algorithm, and k -means, treating all variables as continuous. The performance has been evaluated in terms of recovering the true cluster structure using the Adjusted Rand Index (ARI) [9] between the true hard partition matrix and the estimated one. The ARI counts the pairs of entities that are assigned to the same or different clusters under both partition matrices. The index has expected value zero for independent clusterings and maximum value 1 for identical clusterings.

We simulated 250 samples from a latent mixture of Gaussians with three components. We considered 8 scenarios given by three different experimental factors: the sample size ($N = 100, 500$), the separation between clusters (well separated or not), and number of ordinal variables (3 ordinal and 5 continuous variables or the other way around).

In order to have approximately the same computational time for each method, the model-based approaches (Naive and Mixed) were initialized using only one good rational starting point described in Sect. 2, while for the remaining ones, 10 different random starting points were used.

Data were generated from a three-component mixture model partially observed with 3 or 5 ordinal variables (5 categories) and 5 or 3 continuous variables. In Table 1, we report the true values that are used to generate the data. The overlap between groups is measured by the Bhattacharyya distance [3, 4]. The Bhattacharyya

Table 1 True values of the observed/latent three-component mixture model and thresholds under different scenarios

Common parameters	
Mixture weights	$p_1 = 0.25$
	$p_2 = 0.35$
	$p_3 = 0.40$
Covariance matrixes	$\Sigma_1 = \begin{bmatrix} 2.50 & 0.60 & 1.50 & 0.50 & 0.20 & 0.70 & 0.40 & 0.40 \\ 0.60 & 1.00 & 0.40 & 0.40 & 0.65 & 0.40 & 0.50 & 0.20 \\ 1.50 & 0.40 & 2.00 & 0.30 & 0.25 & 0.50 & 0.4 & 0.30 \\ 0.50 & 0.40 & 0.30 & 1.00 & 1.00 & 0.40 & 0.25 & 0.50 \\ 0.20 & 0.65 & 0.25 & 1.00 & 2.00 & 0.70 & 0.65 & 0.20 \\ 0.70 & 0.40 & 0.50 & 0.40 & 0.70 & 1.50 & 0.30 & 0.40 \\ 0.40 & 0.50 & 0.40 & 0.25 & 0.65 & 0.30 & 1.75 & 0.25 \\ 0.40 & 0.20 & 0.30 & 0.50 & 0.20 & 0.40 & 0.25 & 1.00 \end{bmatrix}$ $\Sigma_2 = \begin{bmatrix} 1.875 & 0.450 & 1.125 & 0.375 & 0.150 & 0.5250 & 0.375 & 0.300 \\ 0.450 & 0.750 & 0.300 & 0.300 & 0.4875 & 0.300 & 0.300 & 1.125 \\ 1.125 & 0.300 & 1.500 & 0.225 & 0.1875 & 0.375 & 0.450 & 0.750 \\ 0.375 & 0.300 & 0.225 & 0.750 & 0.750 & 0.300 & 0.5250 & 0.150 \\ 0.150 & 0.4875 & 0.1875 & 0.750 & 1.500 & 0.525 & 0.375 & 0.225 \\ 0.525 & 0.300 & 0.375 & 0.300 & 0.525 & 1.125 & 0.750 & 0.1875 \\ 0.375 & 0.300 & 0.450 & 0.525 & 0.375 & 0.750 & 1.000 & 0.500 \\ 0.300 & 1.125 & 0.750 & 0.150 & 0.225 & 0.1875 & 0.500 & 1.75 \end{bmatrix}$ $\Sigma_3 = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$
Thresholds	[0, 1, 2, 3]
Separated groups	
Mean Vectors	$\mu_1 = [-1, 3.5, 1.5, 0, -2, 3, 3, 5]$ $\mu_2 = [2, 0, 4.5, 5, 3, 7, -2, 0]$ $\mu_3 = [0, -2, -1, -2, 5, -3, 0, -3]$
Non-separated groups	
Mean Vectors	$\mu_1 = [-1, 3.5, 1.5, 0, -2, 3, 0, 5]$ $\mu_2 = [2, 1, 3, 1.5, 0, 2, -2, 2]$ $\mu_3 = [0, -1, 0, -0.5, 2, -1, 1.5, -1]$

Table 2 Simulation results: ARI values for different clustering methods across the eight scenarios with $N = 100, 500$, groups with high (H) or low (L) level of separation and number of ordinal variables equal to 3 or 5 with $G = 3$. The Gower distance methods, Gower + PAM (G-PAM) and k -means were initialized using 10 (10) random starting points

3 Ordinal Variable and 5 Continuous Variables				
N = 100 & H	Mixed	Naive	G-PAM (10)	k -means (10)
Mean	0.7997	0.2310	0.5966	0.6566
Median	0.7684	0.1886	0.5947	0.6539
Std	0.1235	0.2209	0.0091	0.0085
N = 500 & H	Mixed	Naive	G-PAM (10)	k -means (10)
Mean	0.9444	0.2618	0.5962	0.6538
Median	0.9663	0.2925	0.5967	0.6544
Std	0.0517	0.1917	0.0064	0.0092
N = 100 & L	Mixed	Naive	G-PAM (10)	k -means (10)
Mean	0.6322	0.1456	0.5824	0.6501
Median	0.6202	0.1066	0.5865	0.6532
Std	0.1096	0.1164	0.0280	0.0121
N = 500 & L	Mixed	Naive	G-PAM (10)	k -means (10)
Mean	0.8953	0.2235	0.5957	0.6543
Median	0.8957	0.1046	0.5962	0.6550
Std	0.0832	0.2416	0.0064	0.0090
5 Ordinal Variable & 3 Continuous Variables				
N = 100 & H	Mixed	Naive	G-PAM (10)	k -means (10)
Mean	0.6895	0.2223	0.5921	0.6125
Median	0.6354	0.1437	0.5891	0.6095
Std	0.1547	0.2271	0.0124	0.0124
N = 500 & H	Mixed	Naive	G-PAM (10)	k -means (10)
Mean	0.8181	0.3725	0.5898	0.6124
Median	0.8435	0.3511	0.5882	0.6089
Std	0.1096	0.2735	0.0088	0.0151
N = 100 & L	Mixed	Naive	G-PAM (10)	k -means (10)
Mean	0.6073	0.1080	0.5545	0.6458
Median	0.5634	0.0113	0.5553	0.6438
Std	0.1321	0.1877	0.0254	0.0120
N = 500 & L	Mixed	Naive	G-PAM (10)	k -means (10)
Mean	0.8069	0.2027	0.5454	0.6432
Median	0.8150	0.1255	0.5413	0.6423
Std	0.1442	0.2342	0.0130	0.0080

distance is equal to: 19.00 considering $g = 1, 2$, 26.27 considering $g = 1, 3$ and 34.27 considering $g = 2, 3$ when the groups are well separated; 5.96 considering $g = 1, 2$, 12.98 considering $g = 1, 3$ and 11.24 considering $g = 2, 3$ when the groups are not well separated. In the simulation study, the number of groups is kept fixed. Indeed, the purpose of the study is to assess the ability of the algorithm to capture the cluster structure. In Table 2 we report the simulation results.

Analyzing the results in Table 2, we note that all clustering methods improve their performances as N increases and the level of separation between groups is higher, as expected. In almost all scenarios, the mixture model for mixed-type data seems to behave better than others. Indeed, we note that in terms of mean or median the mixture model for mixed-type data is the best, followed by the k -means and PAM based on the Gower distance matrix. The poorest performances are shown by the naive approach. In terms of mean or median, the mixture model for mixed-type data is not always the best compared to the non-model-based approaches. More specifically, when $N = 100$ and the groups are not well separated, it seems that it is more affected by the issue of local maxima. Furthermore, we note that when there are more ordinal variables than continuous variables, ARI values decrease, although when N increases the worsening is not significant. This is expected, since more ordinal variables we have, more information is losing about the cluster structure underlying the data. Finally, although it is still common to treat ordinal data as metric, we have shown that it can lead to wrong results, especially when the groups are not well separated.

5 Concluding Remarks

In this paper, we compared the model-based approach and Gower distance methods to cluster mixed-type data. From the simulation study, it is possible to conclude that when the groups are less separated, the clustering performances of the Gower distance methods seem to be more affected by the choice of the random starting points. The model-based for mixed type of data as N increases becomes the best one both in terms of means and median. However, it is important to note that larger sample sizes could cause some computational problems. On one hand, for larger N it is possible to compute the Gower matrix, but its storage may become infeasible. On the other hand, this leads to a higher number of bivariate integrals involved in the composite likelihood. However, this increase remains linear, and thus still feasible.

References

1. Agresti, A.: Analysis of Ordinal Categorical Data, vol. 656. Wiley (2010)
2. Ahmad, A., Khan, S.S.: Survey of state-of-the-art mixed data clustering algorithms. IEEE Access 7, 31883–31902 (2019)

3. Bagnato, L., Greselin, F., Punzo, A.: On the spectral decomposition in normal discriminant analysis. *Commun. Stat. - Simul. Comput.* **43**(6), 1471–1489 (2014)
4. Bhattacharyya, A.: On a measure of divergence between two multinomial populations. *Sankhya: Ind. J. Stat. (1933-1960)* **7**(4), 401–406 (1946)
5. Everitt, B.: A finite mixture model for the clustering of mixed-mode data. *Stat. Prob. Lett.* **6**(5), 305–309 (1988)
6. Foss, A.H., Markatou, M., Ray, B.: Distance metrics and clustering methods for mixed-type data. *Int. Stat. Rev.* **87**(1), 80–109 (2019)
7. Gollini, I., Murphy, T.: Mixture of latent trait analyzers for model-based clustering of categorical data. *Stat. Comput.* **24**(4), 569–588 (2014)
8. Gower, J.C.: A general coefficient of similarity and some of its properties. *Biometrics* **27**(4), 857–871 (1971)
9. Hubert, L., Arabie, P.: Comparing partitions. *J. Classif.* **2**(1), 193–218 (1985)
10. Hunt, L., Jorgensen, M.: Clustering mixed data. *WIREs Data Min. Knowl. Disc.* **1**(4), 352–361 (2011)
11. Jöreskog, K.G.: New developments in lisrel: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity* **24**(4), 387–404 (1990)
12. Jöreskog, K.G., Sörbom, D.: LISREL 8: User's Reference Guide. Scientific Software (1996)
13. Kaufman, L., Rousseeuw, P.J.: Clustering by means of medoids (1987)
14. Lee, S.Y., Poon, W.Y., Bentler, P.: Full maximum likelihood analysis of structural equation models with polytomous variables. *Stat. Prob. Lett.* **9**(1), 91–97 (1990)
15. Lindsay, B.: Composite likelihood methods. *Contemp. Math.* **80**, 221–239 (1988)
16. Lubke, G., Neale, M.: Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behav. Res.* **43**(4), 592–620 (2008)
17. McLachlan, G., Peel, D.: *Finite Mixture Models*. Wiley (2000)
18. Mechelen, I., Boulesteix, A., Dangl, R., Dean, N., Guyon, I., Hennig, C., Leisch, F., Steinley, D.: Benchmarking in cluster analysis: A white paper. [arXiv: Other Statistics](https://arxiv.org/abs/1808.08867) (2018)
19. Millsap, R.E., Yun-Tein, J.: Assessing factorial invariance in ordered-categorical measures. *Multivariate Behav. Res.* **39**(3), 479–515 (2004)
20. Muthén, B.: A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49**(1), 115–132 (1984)
21. Ranalli, M., Rocci, R.: Mixture models for ordinal data: a pairwise likelihood approach. *Stat. Comput.* 1–19 (2016). <https://doi.org/10.1007/s11222-014-9543-4>
22. Ranalli, M., Rocci, R.: Standard and novel model selection criteria in the pairwise likelihood estimation of a mixture model for ordinal data. In: Adalbert, F.X., Hans, W., Kestler, A. (eds.) *Analysis of Large and Complex Data. Studies in Classification, Data Analysis and Knowledge Organization* (2016). <https://doi.org/10.1007/978-3-319-25226-1>
23. Ranalli, M., Rocci, R.: Mixture models for mixed-type data through a composite likelihood approach. *Comput. Stat. Data Anal.* **110**(C), 87–102 (2017). <https://doi.org/10.1016/j.csda.2016.12.01>
24. Ranalli, M., Rocci, R.: A model-based approach to simultaneous clustering and dimensional reduction of ordinal data. *Psychometrika* (2017). <http://orcid.org/10.1007/s11336-017-9578-5>
25. Steinley, D.: Handbook of Cluster Analysis, chap. *K-Medoids and Other Criteria for Crisp Clustering*. Chapman and Hall/CRC, New York (2016)