Journal of Classification
https://doi.org/10.1007/s00357-025-09516-3

ORIGINAL RESEARCH

Automatic Dendrogram Slicing for Mixed-Type Data Clustering

Lucio Palazzo[1] · Alfonso Iodice D'Enza[2] · Domenico Vistocco[2] · Francesco Palumbo[2]

Accepted: 16 July 2025

Abstract
Clustering is one of the most ubiquitous unsupervised learning tasks, with applications to a wide variety of domains. Genomics data analysis makes no exception, yet genomics datasets combine features of different natures (continuous and categorical) and are, therefore, of mixed type. The hierarchical clustering of a set of genomics data observations requires pairwise distances or dissimilarities, and it returns a sequence of nested clustering partitions represented as a tree graph. The choice of the dissimilarity or distance measure affects the obtained sequence of cluster partitions. Furthermore, selecting the reference partition out of the nested sequence is up to the user: this is done by setting a threshold, that is, by cutting the tree-based graph horizontally. A permutation test-based procedure has been proposed in the literature to select the final partition based on more than a single threshold (non-horizontal cut). This paper introduces a novel top-down implementation of such permutation test-based procedure to identify the final partition out of a hierarchy of solutions. Different approaches for distance computations are considered to extend the procedure's applicability to the mixed data case.

Keywords Hierarchical clustering · Mixed data · Permutation test

1 Introduction

Clustering is a widely used unsupervised learning technique aimed at identifying heterogeneous groups of homogeneous observations. Different application domains often require tailored approaches to clustering. In genomics (and multi-omics) data analysis, clustering is a common task: genomics datasets frequently encompass information on genetic variations, gene expression levels, epigenetic modifications, and more. These datasets often involve a

✉ Lucio Palazzo
lucio.palazzo@unior.it

✉ Francesco Palumbo
fpalumbo@unina.it

[1] Department of Social Sciences and Humanities, University of Naples L'Orientale, Largo San Giovanni Maggiore 30, 80134 Naples, NA, Italy

[2] Department of Political Sciences, University of Naples Federico II, Via Leopoldo Rodin 22, 80133 Naples, NA, Italy

Published online: 30 July 2025

combination of diverse feature types, including continuous, categorical, and sequence-based data. This inherent diversity makes genomics data inherently mixed in terms of data types (McMurdie and Holmes, 2012; Ritchie et al., 2015). Such mixed-type genomics data are commonly referred to as multi-omics.

This paper focuses on the best-known and used approach, which is distance-based, albeit other alternative methods to clustering exist (e.g., model-based, Bouveyron and Brunet-Saumard, 2014; and density-based, Bhattacharjee and Mitra, 2021). In particular, distance-based clustering methods can be roughly divided into partitive (e.g., k-means, Lloyd, 1982; MacQueen et al., 1967) and hierarchical: the former approach is computationally more efficient than the latter, but it requires the user to set the number of clusters in advance. Choosing the number of clusters is a delicate task that can be seen as a hyperparameter tuning process: because of the unsupervised nature of clustering, the performance metric of choice critically affects the tuning process, and different metrics may lead to different numbers of clusters.

In hierarchical clustering, there is no need to choose the number of clusters in advance; in fact, the observations are progressively aggregated (bottom-up) into larger groups or progressively divided (top-down) into smaller groups: the output is, therefore, a sequence of cluster partitions, often represented as an upside-down tree graph referred to as dendrogram. Another difference between hierarchical and partitive approaches is in the input: in hierarchical clustering, the input is a dissimilarity or distance matrix.

In general, hierarchical clustering is well suited for mixed-type data (Ahmad and Dey, 2007; Chavent et al., 2014; Ahmad and Khan, 2019), and it does not require cluster mean (or centroid) computation: in the case of mixed data, the number of options to choose from when it comes to centroid calculation increases considerably.

Enhanced approaches to mixed-type data clustering have been proposed in the literature and have been comparatively reviewed: notable contributions are the reviews by McNicholas (2016) and Foss et al. (2019) for model-based methods and the review by Van de Velden et al. (2019) for distance-based methods. Surveys that compare techniques for mixed data can be found in the literature (Ahmad and Dey, 2007; Ahmad and Khan, 2019; Van de Velden et al., 2019; Chu et al., 2024): a comparative overview is beyond the scope of the paper, and we constrain ourselves to distance-based hierarchical clustering for mixed data.

Pairwise distance computation for mixed-type data is not straightforward. In practice, data are often re-coded to homogenize the features (by quantifying the categorical features or by discretizing the continuous features) and to process them as all continuous or all categorical. An alternative approach to distance computation is to one-hot encode the categorical features and then apply L2 (Euclidean) or L1 (city-block) metrics to the whole dataset: using L1 or L2 metrics on one-hot encoded features corresponds to using the simple matching dissimilarity index. Dimension reduction for mixed data is another option to homogenize the features: in particular, Factor Analysis of Mixed Data (FAMD, independently proposed by different authors (Hill and Smith, 1976; De Leeuw and Van Rijckevorsel, 1980; Kiers, 1991; Pagès, 2004)) is applied to obtain low-dimensional scores that are then clustered. Finally, another approach is to use ad hoc distance for mixed-type data, taking into account the different nature of the features explicitly. Among ad hoc distance/dissimilarity indexes for mixed-type data, Gower's index (Gower, 1971) is arguably the most common choice. In particular, the Gower dissimilarity index is the convex combination of matching dissimilarity and range-normalized city-block distance. The weight of the convex combination depends on the proportion of categorical and continuous features at hand. The Gower index does not take into account possible between-feature associations/correlations since it is purely additive,

that is, it assumes independence between the features (the value is the sum of by-feature differences).

Fairly recent reviews on distances for non-continuous data (Ross, 2014; Ring et al., 2015) argued in favor of distances that explicitly take into account inter-dependencies among features. Among them, Mousavi and Sehhati (2023) propose a spectral clustering procedure that takes as input entropy-based distances that consider both inter- and intra-feature differences. The entropy-based distance has also been applied in combination with k-prototypes (Huang, 1997), but not yet in combination with hierarchical clustering.

We assessed the hierarchical clustering performances with respect to all four aforementioned distance computation approaches, on real and synthetic data, in Section 4 and Appendix 2.

The selection of a clustering solution out of the hierarchy of partitions depicted in the dendrogram requires the user to specify a threshold (dendrogram cut). The selected threshold is unique; therefore, the dendrogram cut is horizontal. A horizontal cut implies that some of the possible partitions are ignored, that is, partitions that might be obtained using different thresholds for different branches of the dendrogram. The DEndrogram Slicing through a PermutatiOn Test Approach (DESPOTA (Bruzzese and Vistocco, 2015)) is a procedure that extends the spectrum of possible clustering solutions. In DESPOTA, the dendrogram branches can be cut at different heights: a permutation test (Good, 2013) is applied to progressively lower-level nodes to decide whether to split them or not.

In the original DESPOTA implementation, the search space of partitions (consisting of the nodes of the dendrogram) is explored using a top-down approach. However, the permuted values of the test statistic are computed via a bottom-up approach, which incurs significant computational costs, even for moderately large datasets.

This paper aims to propose a top-down implementation of DESPOTA that aligns with the top-down exploration of dendrogram nodes while reducing the computational burden. We illustrate the proposed procedure via an application on multi-omics data, consisting of the most relevant features for methylation and gene expression (continuous) data, combined with categorical clinical data of non-small cell lung cancer patients. Different approaches to distance computations are compared with each other and with the corresponding solutions obtained via horizontal cuts. Finally, the obtained clusters are evaluated by means of survival analysis.

The paper is structured as follows: in Section 2, the key elements of a dendrogram growth are described: the pairwise distances computation and the construction of the clusters hierarchy; in Section 3, the non-horizontal cut procedure is described alongside with its proposed variant. The proposed procedure is then applied to multi-omics data in Section 4, while Section 5 concludes the paper.

2 Growing the Dendrogram

The key elements that determine the hierarchy of clustering solutions are (*i*) the way pairwise distances are measured, which depends mostly, but not exclusively, on the nature of the data at hand, and (*ii*) the way clusters are obtained (either by progressive aggregations or recursive splits).

In Section 2.1, we briefly describe the evaluated approaches to distance computation: Euclidean with the one-hot encoding of the categorical features, FAMD-based dimension reduction, Gower and entropy-based distances. All of the above approaches are presented

in terms of the general framework for computing distances between non-continuous data proposed by Van de Velden et al. (2024a). In Section 2.2, we discuss the linkage functions, define a distance-based computation of the minimum increase of the sum of squares, and review some computationally efficient divisive hierarchical clustering procedures.

2.1 Distance Computation for Mixed-Type Data

Consider a set of n observations described by a set of Q_u numerical features $X_j, j = 1\ldots, Q_u$ and by a set of Q_c (ordered or unordered) categorical features C_j, each with q_j categories, $j = 1\ldots, Q_c$. Let the numerical features be stored in the $n \times Q_u$ data matrix $\mathbf{X}$; let $\mathbf{Z}_j$ be the one-hot encoding of C_j, and let $\mathbf{Z} = \left[\mathbf{Z}_1, \ldots, \mathbf{Z}_{Q_c}\right]$ the $n \times Q_c^*$ super-indicator matrix, with $Q_c^* = \sum_{j=1}^{Q_c} q_j$.

In the continuous data case, consider $\mathbf{X}^*$ to be a pre-transformed version of $\mathbf{X}$, e.g., if the features are scaled to unit variance, then $\mathbf{X}^* = \mathbf{X}(\Sigma_{\mathbf{x}}^*)^{-1/2}$, where $\Sigma_{\mathbf{x}}^*$ is the diagonal part of $\Sigma_{\mathbf{x}}$, the covariance matrix of $\mathbf{X}$. The Euclidean distance matrix $\mathbf{D}$ is

$$\mathbf{D} = \left[\mathbf{g}\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\mathbf{g}^\mathsf{T} - 2\mathbf{X}^*\mathbf{X}^{*\mathsf{T}}\right]^{1/2}, \tag{1}$$

where $\mathbf{g} = \text{diag}(\mathbf{X}^*\mathbf{X}^{*\mathsf{T}})$ and $\mathbf{1}_n$ an n-dimensional vector of 1's.

A straightforward way to extend the distance matrix computation to the mixed data case is to one-hot encode the categorical features and juxtapose them to the continuous features, obtaining the general data matrix $\mathbf{Y} = [\mathbf{X}\,\mathbf{Z}]$. Such an approach is referred to naive as in Van de Velden et al. (2024b): we simply refer to it as Euclidean distance for mixed data. Upon defining $\mathbf{X}^* = \mathbf{Y}(\Sigma_{\mathbf{y}}^*)^{-1/2}$, one can compute the pairwise distances matrix as in Eq. 1.

The dimension reduction approach is based on the computation of the FAMD principal component scores: in particular, the data matrix $\mathbf{Y}$ is transformed as follows:

$$\mathbf{Y}^* = \mathbf{X}\Sigma_{\mathbf{x}}^{*-1/2}\,\mathbf{D}_z^{-1/2}\mathbf{Z}\mathbf{M}, \tag{2}$$

where $\mathbf{M} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$ is centering operator, $\mathbf{I}_n$ is an $n-$dimensional identity matrix, and $\mathbf{D}_z$ the diagonal part of $\mathbf{Z}^\mathsf{T}\mathbf{Z}$. The FAMD loss function is as follows:

$$\min_{\mathbf{A},\mathbf{B}} \left\| n^{-1/2}\mathbf{Y}^* p^{-1/2} - n^{1/2}\mathbf{A}\mathbf{B}^\mathsf{T} p^{1/2} \right\|^2 \quad \text{s.t.} \quad p\mathbf{B}^\mathsf{T}\mathbf{B} = \mathbf{I}_d. \tag{3}$$

The $n \times d$ matrix of row principal scores is $\mathbf{A} = n^{1/2}\mathbf{U}\mathbf{D}_\alpha$ where $\mathbf{U}$ and $\mathbf{D}_\alpha$ are the first d left singular vectors and values resulting from the singular value decomposition of $\mathbf{Y}^*$. By setting $\mathbf{A} = \mathbf{X}^*$, the distance matrix $\mathbf{D}$ is obtained via Formula Eq. 1.

The Gower dissimilarity index is an additive measure that combines the range-normalized L1 distance (for continuous features) and the simple matching coefficient (for categorical features). The distance matrix is, therefore, the sum of two components:

$$\mathbf{D} = \mathbf{D}_{L1} + \mathbf{D}_m. \tag{4}$$

In particular, the D_{L1} distance matrix has general element $\mathbf{D}_{L1}(i, i^{'})$, with $i, i^{'} = 1, \ldots, n$ and $i \neq i^{'}$ given by

$$\mathbf{D}_{L1}(i, i^{'}) = \sum_{j=1}^{Q_u} \frac{|x_{ij} - x_{i^{'}j}|}{\text{range}(X_j)}; \tag{5}$$

the matching coefficient dissimilarity matrix $\mathbf{D}_m$ is given by

$$\mathbf{D}_m = \sum_{j=1}^{Q_c} \mathbf{Z}_j \Delta_j^{(m)} \mathbf{Z}_j^\mathsf{T} = \mathbf{Z}\Delta^{(m)}\mathbf{Z}^\mathsf{T}, \tag{6}$$

where $\Delta_j^{(m)} = \mathbf{1}_{q_j}\mathbf{1}_{q_j}^\mathsf{T} - \mathbf{I}_{q_j}$, and $\Delta^{(m)}$ is a $Q_c^* \times Q_c^*$ block-diagonal matrix whose j^{th} diagonal block is $\Delta_j^{(m)}$.

The entropy-based distance considers features interdependence: the differences due to a single feature depend on both the observed values of the feature itself and its interdependence with the other features. For continuous data, an example of interdependence-based is the Mahalanobis distance, which is computed via Formula Eq. 1 upon defining $\mathbf{X}^* = \mathbf{X}(\Sigma_\mathbf{x})^{-1/2}$. According to the considered entropy-based approach (Mousavi and Sehhati, 2023), for continuous features, a modified version of the Mahalanobis distance is used that weighs the entries of $\Sigma_\mathbf{x}^{-1}$ based on the mutual information (MI). For categorical features, the pairwise distances can be defined using Formula Eq. 6: the difference is in the definition of the Δ matrix, that differs from $\Delta^{(m)}$ in that it is not strictly 0/1 valued. Given two features i and j, the $(a, b)^{th}$ element of Δ_j is the linear combination

$$\delta^j(a, b) = \sum_{i \neq j} w_{ji} \Phi^{ji}, \tag{7}$$

where Φ^{ji} is the normalized entropy computed on the conditional distributions of each feature $i \neq j$ given categories a and b, respectively; the weight w_{ji} of the linear combination is the MI index between features j^{th} and i^{th}. Therefore, for non-continuous features, the distance is

$$\mathbf{D}_c = \mathbf{Z}\Delta\mathbf{Z}^\mathsf{T}, \tag{8}$$

whereas the overall distance is the weighted sum of the distances computed on the continuous and categorical features, that is

$$\mathbf{D} = \frac{1}{Q_c + 1}\mathbf{D}_u + \frac{Q_c}{Q_c + 1}\mathbf{D}_c. \tag{9}$$

The definitions of the quantities of the linear combination w_{ji} and Φ^{ji} in Formula Eq. 7 depend on the nature of features j and i, and we refer the reader to the Appendix 1 for details.

2.2 Building the Hierarchy: Bottom-Up vs Top-Down

The hierarchy of partitions can be defined by two alternative approaches: bottom-up (agglomerative) and top-down (divisive). The former generates a sequence of partitions from n (one observation per cluster) to 1 (all observations clustered together); the top-down approach starts from a single cluster, containing all the observations, that is split recursively until obtaining n singletons.

Whether the procedure is agglomerative or divisive, at some point, it is required to compute distances between subgroups of observations: this is done via a user-defined linkage function.

There are multiple types of linkage functions to choose from, such as single, complete, average and Ward. Each linkage defines the distance between sets (clusters) of observations.

Given two sets L and R, of size n_L and n_R, respectively, and their union $F = L \cup R$, a linkage $h(F)$ measures the distance between the two sets of points, L and R.

The single linkage between L and R is given by the minimum distance between an observation i from the L set and an observation i' from the R set, formally

$$h(F) = \min\left(|\mathbf{x}_i - \mathbf{x}_{i'}|^2 : \forall\, i \in L, i' \in R\right); \tag{10}$$

by replacing min() with max() and mean(), the complete and average linkages are obtained. The Ward linkage refers to the Minimum Increase of Sum of Squares (MISSQ):

$$\begin{aligned}\text{MISSQ:}\frac{n_L \cdot n_R}{n_L + n_R}\|\mu_L - \mu_R\|^2 = &\sum_{\mathbf{x}_i \in F}\|\mathbf{x}_i - \mu_F\|^2 + \\ &- \sum_{\mathbf{x}_i \in L}\|\mathbf{x}_i - \mu_L\|^2 - \sum_{\mathbf{x}_i \in R}\|\mathbf{x}_i - \mu_R\|^2.\end{aligned} \tag{11}$$

A further linkage available is the centroid linkage but, in this paper, the focus is on pairwise distance-based linkages and not distance-to-centroid linkages.

The Ward linkage in Eq. 11 does involve centroids, but each element of the right-hand side sum can be rewritten in terms of pairwise distances. In fact, for the general set of observations A, the following relation holds.

$$\sum_{x_i \in A}\|x_i - \mu_A\|^2 = \frac{1}{n_A}\sum_{i=1}^{n_A}\sum_{i'=1}^{n_A}\|x_i - x_{i'}\|^2. \tag{12}$$

According to Eq. 12, the MISSQ can be rewritten as

$$\begin{aligned}\text{MISSQ:}\frac{1}{n_L + n_R}&\sum_{i=1}^{n_L+n_R}\sum_{i'=1}^{n_L+n_R}\|x_i - x_{i'}\|^2 + \\ &- \frac{1}{n_L}\sum_{i=1}^{n_L}\sum_{j=1}^{n_L}\|x_i - x_{i'}\|^2 - \frac{1}{n_R}\sum_{i=1}^{n_R}\sum_{j=1}^{n_R}\|x_i - x_{i'}\|^2.\end{aligned} \tag{13}$$

Recalling that the aim of the paper is to compare two alternative DESPOTA implementations, the linkage of choice can be fixed: from here on we consider the Ward linkage since its computation in the proposed top-down procedure is not straightforward and, therefore, is worth describing (see Section 3).

In hierarchical clustering, the agglomerative approach is more often used because of the computational burden that comes with the divisive approach: in fact, there are $2^{n-1} - 1$ ways to bipartite n observations in two groups, and that is just a single split. Fairly recent contributions in the literature proposed procedures that combine clever splits and evaluation criteria. In Principal Direction Divisive Partitioning (PDDP (Boley, 1998)), the observations are projected on the principal direction identified by the eigenvector associated with the largest eigenvalue of the pairwise distance matrix and then split into two clusters according to the sign of their projection (that is, the split threshold is at 0). The PDDP split is applied recursively until the sequence of clustering solution is obtained, or a user-defined early stop condition is met. Note that the quality of PDDP clustering will depend on whether or not the largest eigenvalue is dominant. Further variants of PDDP have been proposed, e.g., the interval PDDP (iPDDP (Tasoulis et al., 2010)): the projections are obtained as in PDDP, yet the splitting criterion depends on the largest distance between any two projections; in the k-means PDDP (km-PDDP (Zeimpekis and Gallopoulos, 2008)), a k-means procedure (with fixed $k = 2$) is applied to split the projections up. An alternative and popular approach is the bisecting k-means (Savaresi and Boley, 2001), which is a variant of the k-means for $k = 2$. The bisecting k-means is implemented in the `scikit-learn` python library (Pedregosa et

al., 2011), and it is based on a careful seeding of the starting cluster centroids. Unlike PDDP, however, bisecting k-means does require the computation of centroids. In a fairly recent review, Roux (2018) assessed the performance in terms of computational time of several agglomerative versus most of the aforementioned divisive approaches: the PDDP and its variants proved to outperform the agglomerative counterparts.

3 Cutting the Dendrogram: DESPOTA

The horizontal cut of the dendrogram is usually chosen by visual inspection, although some criteria have been proposed in the literature (Milligan and Cooper, 1985; Jung et al., 2003; Ran et al., 2023). Using a single threshold implies discarding some of the possible clustering solutions that could, instead, be explored by using different thresholds for different branches (non-horizontal cuts). Moreover, a subjective dendrogram cut corresponds to a subjective selection of the number of clusters.

The DEndrogram Slicing through a PermutatiOn Test Approach (DESPOTA (Bruzzese and Vistocco, 2015)) evaluates a set of candidate clustering solutions not limited to those accessible via horizontal cut: in DESPOTA, each branch of the dendrogram can be cut at a different height. The exploration of the solutions space starts at the top of the dendrogram and moves downward, to evaluate each node: a permutation test (Good, 2013) is used to decide whether or not the set of observations within that node should be split any further. In particular, the null hypothesis of the permutation test is that the two sets of observations resulting from the split are part of a single cluster. Therefore, the father node in question should not be split, and it becomes a terminal node. If the null hypothesis is rejected, the node is split, and the procedure recursively evaluates the two children nodes to state whether they should be labeled as father or as terminal. The algorithm stops when no further splits are left to evaluate, that is, there are no father nodes left. The pseudo-code for the DESPOTA procedure is reported in Algorithm 1.

Let F_k be the set of observations within the node k, and F_1 is then the full set of observations; similarly, L_k and R_k, respectively of size n_{L_k} and n_{R_k}, are the set of observations in the children nodes originating from F_k. Figure 1a shows that the quantities involved in the computation of the observed value of test statistic are the heights of the dendrogram at F_k, L_k, and R_k, that is, the linkage function values $h(F_k)$, $h(L_k)$, and $h(R_k)$. In particular, the DESPOTA test statistics is

$$rc_k = \frac{|h(L_k) - h(R_k)|}{h(F_k) - \min\{h(L_k), h(R_k)\}} \in [0, 1], \tag{14}$$

and it represents the ratio between the minimum merging cost and the actual merging cost computed on the initial dendrogram (Bruzzese and Vistocco, 2015). The permutation test distribution under the null hypothesis of the rc_k statistics in Eq. 14 is computed on M rearrangements of the children nodes of F_k, $L_k^{(m)}$, and $R_k^{(m)}$ $(m = 1, \ldots, M)$, that are defined by permuting observations from L_k to R_k, preserving the original sizes n_{L_k} and n_{R_k}. The Monte Carlo p-value is defined as

$$p_{rc_k}^{MC}(L_k, R_k) = \frac{1}{M} \sum_{m=1}^{M} I(rc_k^{(m)} \leq rc_k), \tag{15}$$

where $I(\cdot)$ is the indicator function. The null hypothesis is rejected if $p_{rc_k}^{MC}(L_k, R_k) \leq \alpha$, given a user-defined significance level α. When the null hypothesis is rejected, L_k and R_k are investigated to state whether they correspond to a father or terminal node.

```
Input: dendrogram, pairwise distances, α
terminal_node_list ← [ ]
father_node_list ← F_1
repeat
    if is_terminal(father_node_list[1], L, R, α) is true then
        add father_node_list[1] to terminal_node_list
    else
        if h(L) ≥ h(R) then
            add L and R to the father_node_list
        else
            add R and L to the father_node_list
    remove the first element from father_node_list
until father_node_list is empty
return terminal_node_list
```

```
input: F and its children nodes (L and R), pairwise distances, α
M = 1000
father_node_list ← F_1
compute the test statistic rc from F, L and R ▷ see Formula (14)
for m in 1 to M do
    randomly permute L and R to obtain L^(m) and R^(m)
    agglomerative clustering on L^(m) to obtain h(L^(m))
    agglomerative clustering on R^(m) to obtain h(R^(m))
    compute the test statistic rc^(m)
compute p_rc^MC(L, R) ▷ see Formula (15)
if p_rc^MC(L, R) ≤ α then
    terminal_tf = false
else
    terminal_tf = true
return terminal_tf
```

```
input: F and its children nodes (L and R), pairwise distances, α
M = 1000
father_node_list ← F_1
compute the test statistic rc from F, L and R ▷ see Formula (14)
for m in 1 to M do
    randomly permute L and R to obtain L^(m) and R^(m)
    single split of L^(m) to obtain h(L^(m))
    single split of R^(m) to obtain h(R^(m))
    ▷ using Equations in Formula (16) ◁
    compute the test statistic rc^(m)
compute p_rc^MC(L, R) ▷ see Formula (15)
if p_rc^MC(L, R) ≤ α then
    terminal_tf = false
else
    terminal_tf = true
return terminal_tf
```

In DESPOTA, the dendrogram nodes evaluation process follows a top-down strategy. Yet, the linkages $h(F_k)$ (that is fixed), $h(L_k^{(m)})$, and $h(R_k^{(m)})$ are computed according to a bottom-up strategy. In other words, an agglomerative procedure is applied to each of the two sets of

observations in $L_k^{(m)}$ and $R_k^{(m)}$. For each node, $2 \times M$ dendrograms have to be computed to obtain $p_{rc_k}^{MC}$ from Formula Eq. 15. The computational burden becomes cumbersome for even moderately large datasets, more so if several nodes are evaluated.

The novel top-down implementation of DESPOTA proposed in this paper is coherent with the exploration of the clustering solutions hierarchy, and it alleviates the computational burden that comes with the original DESPOTA implementation. To compute rc_k^m for each permutation m, the linkage functions $h(F_k)$, $h(L_k^{(m)})$, and $h(R_k^{(m)})$ are needed: F_k stays the same in each permutation, which is why no m superscript is used. Hereafter, we refer, without loss of generality, to the computations for the Ward linkage case, computed as in Formula Eq. 13. A similar solution is easily generalized to other linkages. In the original DESPOTA implementation, two full dendrograms must be grown on the set of observations in $L_k^{(m)}$ and in $R_k^{(m)}$, respectively: such dendrograms are needed just to obtain $h(L_k^{(m)})$ and $h(R_k^{(m)})$ (see Fig. 1a). The latter quantities, however, can be obtained as follows:

$$
\begin{aligned}
h(L_k^{(m)}) = & \frac{1}{n_{L_k}} \sum_{i,j \in L_k^{(m)}} \|x_i - x_j\|^2 + \\
& - \frac{1}{n_{L_{k,l}^{(m)}}} \sum_{i,j \in L_{k,l}^{(m)}} \|x_i - x_j\|^2 - \frac{1}{n_{L_{k,r}^{(m)}}} \sum_{i,j \in L_{k,r}^{(m)}} \|x_i - x_j\|^2, \\
h(R_k^{(m)}) = & \frac{1}{n_{R_k}} \sum_{i,j \in R_k^{(m)}} \|x_i - x_j\|^2 + \\
& - \frac{1}{n_{R_{k,l}^{(m)}}} \sum_{i,j \in R_{k,l}^{(m)}} \|x_i - x_j\|^2 - \frac{1}{n_{R_{k,r}^{(m)}}} \sum_{i,j \in R_{k,r}^{(m)}} \|x_i - x_j\|^2.
\end{aligned}
\tag{16}
$$

where $\left\{L_{k,l}^{(m)}, L_{k,r}^{(m)}\right\}$ are subsets of $L_k^{(m)}$ and $\left\{R_{k,l}^{(m)}, R_{k,r}^{(m)}\right\}$ are subsets of $R_k^{(m)}$. Note that the number of observations n_{L_k} and n_{L_k} are constant because, for all m, it results that $L_k^{(m)}$ and $R_k^{(m)}$ are of the same size of L_k and R_k, respectively. We propose a top-down procedure, a single step of a divisive approach, to respectively split up $L_k^{(m)}$ into $L_{k,l}^{(m)}$ and $L_{k,r}^{(m)}$, and $R_k^{(m)}$ into $R_{k,l}^{(m)}$ and $R_{k,l}^{(m)}$ (see Fig. 1b). We then use the equations in Formula Eq. 16 to compute $h(L_k^{(m)})$ and $h(R_k^{(m)})$ and obtain rc_k^m. It is worth remarking that the proposed top-down implementation to compute rc_k^m only requires a single divisive split to obtain $h(L_k^{(m)})$ and $h(R_k^{(m)})$, whereas the original implementation of DESPOTA requires a complete agglomerative procedure to be applied on $L_k^{(m)}$ and $R_k^{(m)}$. We compared the computational time required to achieve a two clusters solution using an agglomerative versus some of the divisive approaches mentioned in Section 2.2 that are implemented in the Hipart python library (Anagnostou et al., 2022). The considered PDDP-like methods proved to outperform the agglomerative approach when the number of observations is in the few hundreds: for smaller datasets, the computational time is negligible for all the considered methods. Detailed results are available as supplementary material[1].

[1] https://alfonsoiodicede.github.io/blogposts_archive/divisive_sim.html

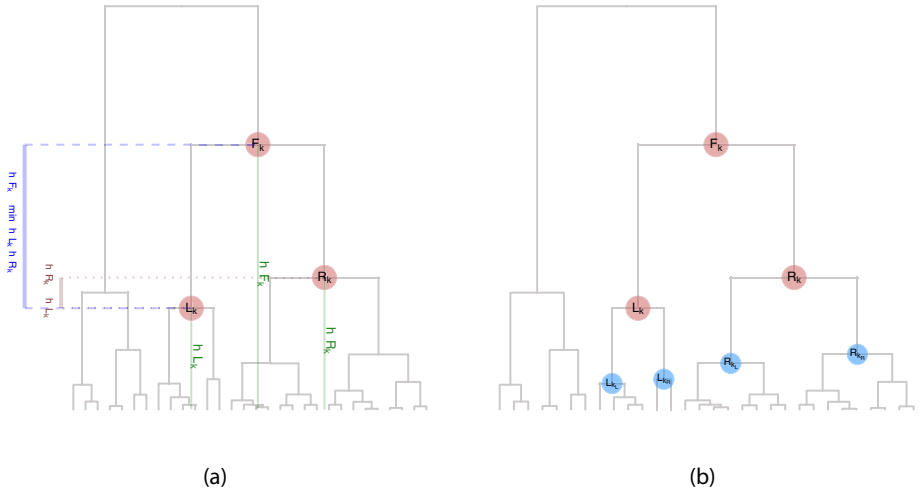

Fig. 1 Computation of rc_k for each permutation: linkage quantities needed (**a**); extra linkage quantities needed in the top-down DESPOTA (no full sub-dendrogram needed, just a single further binary split of the nodes of interest) (**b**)

Algorithms 2 and 3 report the old and the new implementations of DESPOTA's core function (`is_terminal()`) that states whether a node should be further split. Lines 7 and 8 are highlighted as they differentiate the two algorithms from one another.

To show that the bottom-up and the top-down approaches are substantially equivalent in terms of suggested splits, we estimated the distributions of the rc_k test statistics under the null hypothesis computed as in Algorithm 2 and as in various options for Algorithm 3.

We considered the well-known Palmer penguins dataset (Horst et al., 2020) and evaluated the first node (from the top) of the dendrogram. In particular, the dataset consists of four features (bill length, bill depth, flipper length, and body mass) and 333 observations. Eleven rows containing missing values are omitted from the data. Figure 2 reports the histograms of the distribution of the rc_k test statistic under the null hypothesis testing different approaches along with the bottom-up (black histogram). The observed value of the test statistics is 0.23 (vertical red dashed line). We then estimated the test statistics distribution under the null hypothesis computing $rc_k^{(m)}$ as in Eq. 16 on $M = 5000$ permutations. To obtain the subsets of $L_k^{(m)}$ and $R_k^{(m)}$, we used a single step of the divisive algorithms described in Section 2 and compared the results to the bottom-up approach. The aim of such an experiment is to check whether the null distribution produced by each method leads to the same decision, that is, whether or not to split the evaluated node. As a result, for all the considered methods, the $p_{rc_k}^{MC}$ was found to be smaller than 0.05, correctly suggesting splitting up the considered set of observations.

4 Application on Multi-omics Data

Lung cancer is one of the most prevalent malignant tumors worldwide. In particular, non-small cell lung cancer can be divided into two major variations: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Identifying the mechanisms underlying LUAD and LUSC is needed to improve diagnosis and design therapeutic interventions. Recent advances

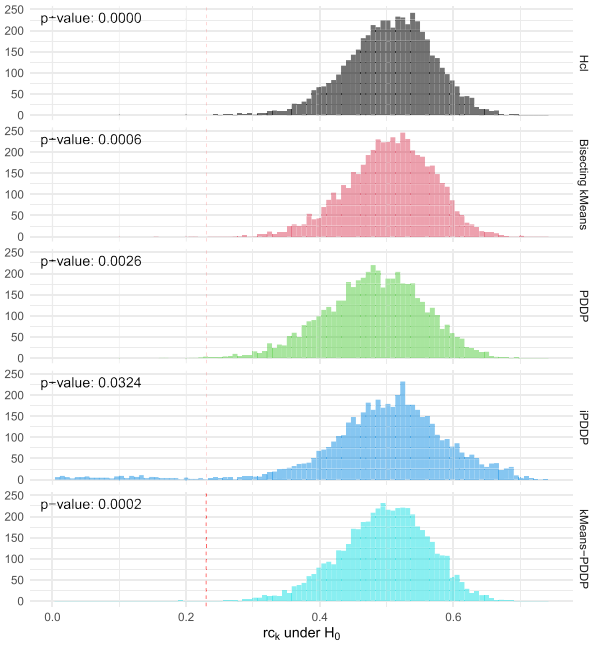

Fig. 2 Histogram of the distribution of the rc_k test statistic under the null hypothesis ($M = 5000$ permutations), as obtained via Algorithm 2 (bottom-up) and Algorithm 3 (top-down). The Hcl plot shows the results of the bottom-up hierarchical clustering, while different divisive approaches were tested for the top-down case: bisecting k-means, PDDP, iPDDP, and km-PDDP. The observed value for the test statistic is $rc_k = 0.23$, and the corresponding $p^{MC}_{rc_k}$ are reporting in each panel: in all but iPDDP, it results that $p^{MC}_{rc_k} \approx 0$

in high throughput omics technologies have led to publicly accessible descriptions of cancer cells' genetic, epigenetic, and transcriptional profiles. In this section, we show the application of the proposed procedure to lung cancer data. In particular, we considered the next-generation sequencing, survival, and clinical data from The Cancer Genome Atlas[2] (TCGA) repository.

Accurately classifying patients to predict prognosis is crucial to help physicians in the decision process, but it remains a difficult task. For example, some studies revealed that relevant cuproptosis-related lncRNAs were associated with the prognosis of patients with LUAD and LUSC, helping to predict prognosis with accuracy (Wang et al., 2023); in other studies, cancer is classified using both novel and traditional machine learning or applying multiple feature selection methods (Chen and Dhahbi, 2021). A recent paper by Ellen et al. (2023) used the same data source to perform multimodal survival analysis and predict non-small cell lung cancer.

In this paper, we apply the same pre-processing pipeline as in Ellen et al. (2023) for the linear feature selection step. In particular, the feature engineering steps were performed separately for each biological modality (mRNA and methylation) associated with LUAD and LUSC data. The feature selection is based on a combination of Wald significance tests and univariate Cox proportional hazards models to evaluate how each feature affects patient survival time. A total of 100 continuous features were selected for each modality, 50 for

[2] Source: www.cancer.gov/ccg/research/genome-sequencing/tcga

Table 1 Description of the considered clinical features

Gender	Gender of the patient (female, male)
Prior malignancy	Indicates whether the patient has a history of malignancies (no, yes)
Number pack years smoked	Lifetime tobacco exposure of the patient, defined as the number of cigarettes smoked per day multiplied by the number of years smoked and divided by 20. This feature has been dichotomized by using the first quartile as the threshold: light smoker if $\leq$ 30, heavy smoker otherwise
Volume	Tumor volume in cm^3. This feature has been dichotomized by using the median as the threshold: low if ≤ 0.32, high otherwise
Stage	Pathological stages as defined by the American Joint Committee on Cancer[6] (I, Ia, Ib, II, IIa, IIb, III, IIIa, IIIb, IV)

mRNA and 50 for methylation. The pre-processing pipeline code is available as supplementary material[3].

A principal component analysis (PCA: see, e.g., Greenacre et al., 2022 for an overview) has been applied to the selected features, and the 20 most informative components (in terms of explained variability) were retained. Clinical data categorical features, described in Table 1, were also considered, as complementary information. We evaluated LUAD and LUSC data for a total of 586 patients, of which 360 (61.4%) survived and 226 (38.6%) died during the study period.

The four mixed-type data distance measures introduced in Section 2.1 are considered. The linkage function of choice is the distance-based Ward, defined in Formula Eq. 13. A comparison of the considered distances and the clustering solutions between the horizontal and the DESPOTA-based (non-horizontal) cut is presented in Fig. 3. In particular, we considered a grid of values for α, from 0.01 to 0.5, with a step of 0.005 for $\alpha \in [0.01, 0.2]$ and 0.1 for $\alpha \in [0.2, 0.5]$. For each value of α, we considered the partition identified by DESPOTA and compared it to the partition with the same number of clusters obtained via the horizontal cut of the dendrogram. The two partitions were compared by means of the average silhouette width (ASW (Rousseeuw, 1987)). For each panel of Fig. 3, the number of clusters is mapped to the horizontal axis and the ASW is mapped to the vertical axis.

For a fixed number of clusters, dendrograms resulting from the entropy-based distance provide better performances in terms of ASW compared to the corresponding horizontal cut. In particular, DESPOTA solutions in the Euclidean case provide ASW values greater than or equal to ASW values resulting from horizontal cuts. The ASW resulting from FAMD-based and Gower horizontal cuts increase with the size of the partition and are larger than the ASW resulting from non-horizontal cuts. It is worth noting that, when the number of selected clusters is low, there are few partitions alternative to the horizontal cut, as depicted in Fig. 3 for lower values of k.

The ASW of the entropy-based dissimilarity on DESPOTA clusters is higher than that of the alternative approaches. The entropy-based ASW indicates the best solution to be $k = 4$ clusters; Euclidean, Gower, and FAMD-based suggested solutions are $k = 4$, $k = 3$, and $k = 6$ clusters, respectively, with the Euclidean ASW being the second best. We further analyze the obtained partitions to assess the interpretability of the solutions obtained via the entropy-based distance (best) and Euclidean distance (second best).

[3] https://alfonsoiodicede.github.io/blogposts_archive/SR_despota_mix_supplementary.html

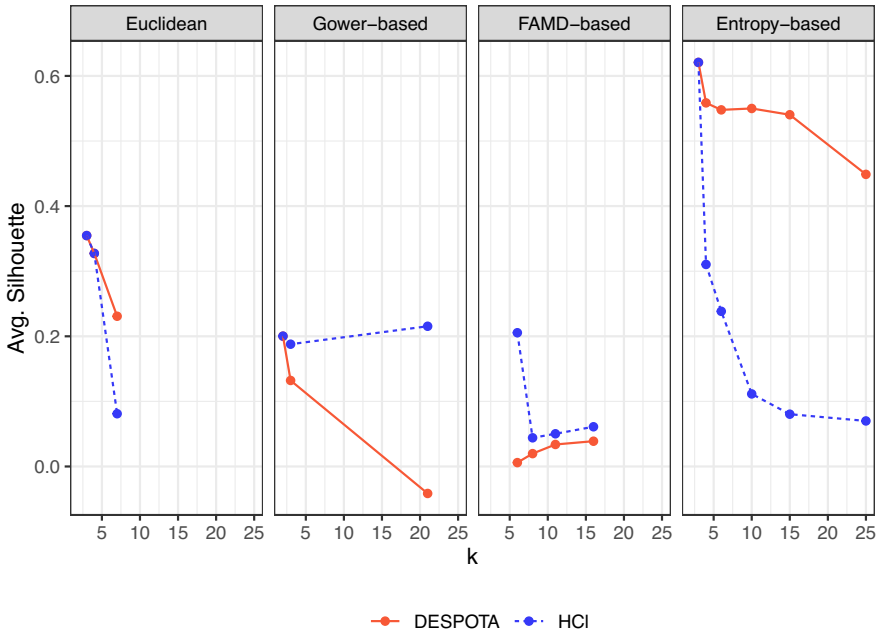

Fig. 3 Comparison between DESPOTA (bold line) and horizontal cut (HCl, dashed line). The partitions are automatically identified by DESPOTA with a varying significance level α. The average silhouette widths of DESPOTA clusters are compared with the partition of the same size obtained via horizontal cut

In multi-omics data clustering, one of the aims of the analysis is to identify clusters with survival rates that are similar within each cluster and different among clusters. To this end, for each of the partitions considered, we compare the by-cluster Kaplan-Meier curves. Figures 4a and 4b depict the Kaplan-Meier curves of the cluster solutions obtained through the entropy-based and Euclidean distance, respectively, comparing both DESPOTA and horizontal cluster solutions. The curves in panel Fig. 4a show different survival profiles among the groups, and the two cuts produce different clustering allocations, while survival curves of the Euclidean case (Fig. 4b) generate the same result, as expected. All the considered scenarios present a significant difference in survival between the groups, with p-values lower than a confidence level of 0.05.

The partition based on the Euclidean distance consists of clusters with 469, 93, 23, and 1 patients, respectively: the singleton being characterized by a very low survival rate (see Fig. 4b). In the entropy-based dendrogram case, the cut suggested by DESPOTA is non-horizontal, which leads to a different partition. In particular, the two partitions differ from each other and lead to different survival curves (Fig. 4a). Table 2 depicts the main characteristics of the DESPOTA-based partition: most patients with high two-year restricted mean survival times (RMST*) belong to the first cluster, for both LUAD and LUSC subtypes. In the second cluster, the 1-year survival probability of LUAD and LUSC patients differ considerably (0.25 and 0.47). Clusters obtained with the horizontal cut show similar features. Patients from clusters one and two have similar RMST* and survival probabilities; they mainly differ in median survival days, as the cluster 1 corresponding value is twice the cluster 2 value. Clusters three and four differ in the survival probability for LUAD patients, which is higher in cluster three (0.58). The Euclidean-based partition, characterized by the presence of a singleton, is

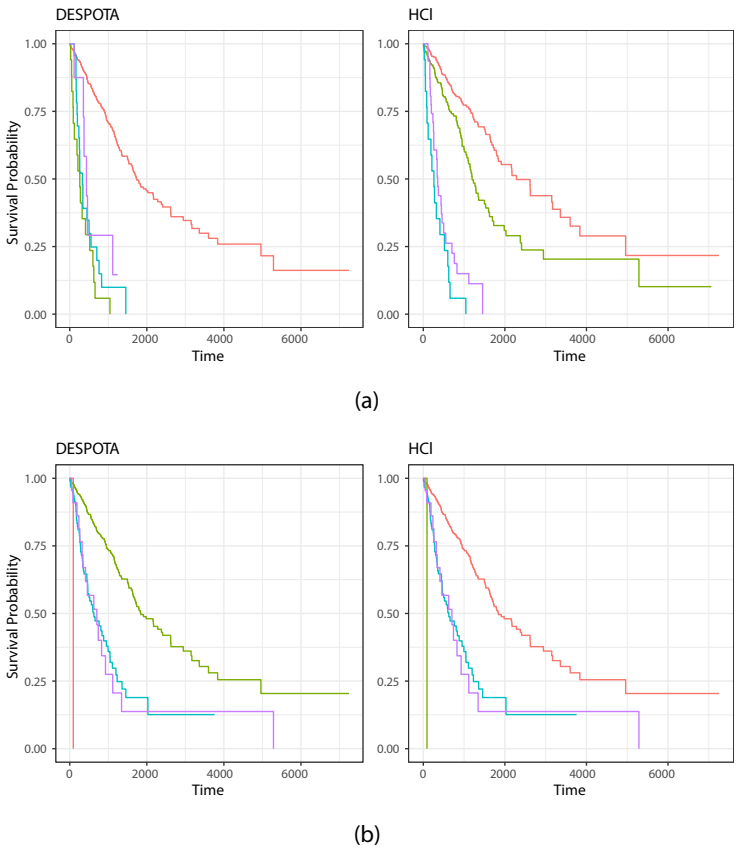

Fig. 4 Survival curves between clusters of patients identified by DESPOTA versus horizontal (HCl) cut, both for the entropy-based (**a**) and Euclidean dissimilarities (**b**)

also characterized by a larger cluster of patients with a high survival probability. Clusters two and three are characterized by a low survival probability of LUSC and LUAD patients, respectively.

The v-test assesses whether a cluster is characterized by specific features. For continuous features, the v-test statistic measures the standardized deviation between the cluster-specific mean and the overall mean. For categorical features, the v-test evaluates whether the proportions of a specific category differ significantly across clusters (see, e.g., Husson et al., 2017).

The test is implemented using the `FactoMineR` R package. In the analyzed application, v-tests are used to determine whether each subgroup can be characterized by the clinical features listed in Table 1, with the cancer subtype included as a supplementary feature.

In addition, we grouped the AJCC pathological stages into four macro-categories (I, II, III, IV). It should be noted that this test can only make inferential conclusions if the categorical features were not used to determine the clusters. Nonetheless, we calculated the v-test statistic as a heuristic criterion. Figure 5 summarizes the v-test statistics of the clinical features evaluated for each considered approach, along with their significance levels. As

Table 2 Survival analysis results of the proposed methods stratified by cancer subtypes

		n	Deaths	RMST*	Median	Survival 1-y
Entropy-based (DESPOTA)						
1 ($n_c = 537$)	LUAD	296	94	659 (10.2)	1632	0.90 (0.02)
	LUSC	241	88	649 (12.1)	1912	0.89 (0.02)
2 ($n_c = 24$)	LUAD	15	13	417 (60.4)	340	0.47 (0.13)
	LUSC	9	8	339 (67.3)	282	0.25 (0.15)
3 ($n_c = 17$)	LUAD	9	9	363 (76.4)	275	0.33 (0.16)
	LUSC	8	8	243 (77.1)	148	0.38 (0.17)
4 ($n_c = 8$)	LUAD	7	5	513 (64.3)	447	0.83 (0.15)
	LUSC	1	1	113 (—)	113	1.00 (—)
Entropy-based (Horizontal)						
1 ($n_c = 312$)	LUAD	162	46	680 (11.2)	2617	0.93 (0.02)
	LUSC	150	46	664 (13.9)	2284	0.92 (0.02)
2 ($n_c = 225$)	LUAD	134	48	633 (17.9)	1288	0.87 (0.03)
	LUSC	91	42	625 (22.2)	1107	0.84 (0.04)
3 ($n_c = 32$)	LUAD	22	18	448 (47.5)	434	0.58 (0.11)
	LUSC	10	9	314 (64.3)	236	0.22 (0.14)
4 ($n_c = 17$)	LUAD	9	9	363 (76.4)	275	0.33 (0.16)
	LUSC	8	8	243 (77.1)	148	0.38 (0.17)
Euclidean						
1 ($n_c = 469$)	LUAD	238	68	675 (10.1)	1790	0.93 (0.02)
	LUSC	231	82	646 (12.6)	1984	0.89 (0.02)
2 ($n_c = 93$)	LUAD	81	46	538 (27.1)	807	0.71 (0.05)
	LUSC	12	12	333 (70.7)	270	0.33 (0.14)
3 ($n_c = 23$)	LUAD	7	6	506 (83.6)	624	0.57 (0.19)
	LUSC	16	11	530 (66.1)	740	0.72 (0.12)
4 ($n_c = 1$)	LUAD	1	1	91 (—)	91	1.00 (—)
	LUSC	0	0	— (—)	—	— (—)

Note: RMST*, restricted mean survival time (in days) in each of the groups with an upper limit equal to 730 days (two years); *Median*, median survival time; *Survival 1-y*, the survival probability of a patient for at least 1 year. Standard errors in parentheses

expected, the pathological stage is one of the most influencing features; nonetheless, the light smoker, gender, and cancer subtype play a role in characterizing some clusters.

Regarding the outcomes of the entropy-based distance and DESPOTA cut, patients from cluster 1 exhibited low pathological stages, while clusters 3 and 4 are characterized by the most severe pathological stages (the v-test scores of stages III and IV are high and positive). In cluster 2, it is possible to notice a higher concentration of male patients, while cluster 4 is unbalanced towards LUAD patients. This is also a consequence of the small cluster sizes, that is, 17 patients for the third and 8 for the fourth cluster. Looking at the horizontal cut of the entropy-based distance, pathological stages seem to be more influential in all the clusters. In particular, the first two clusters (312 and 225 patients, respectively) have opposite v-test scores for pathological stages, prior malignancy and cancer volume. Clusters 3 and 4 show similar results, confirming the results noticed in Table 2. In conclusion, the disease

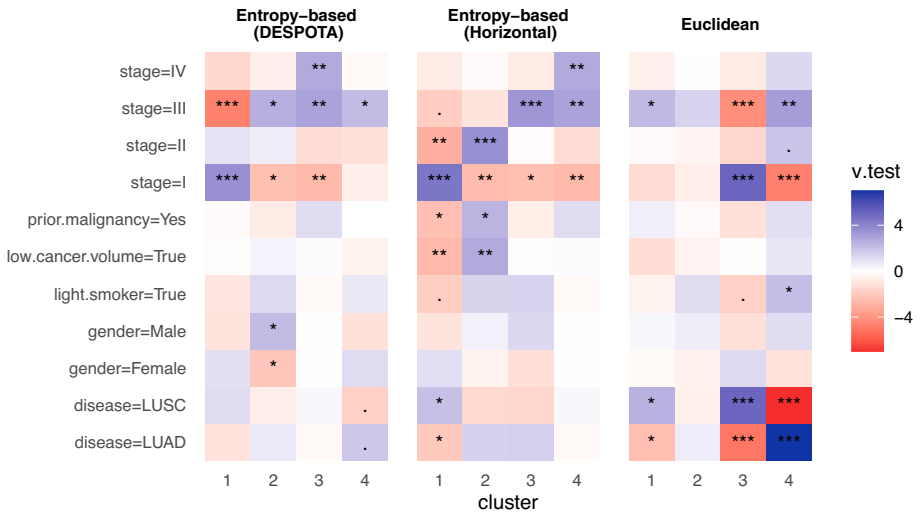

Fig. 5 Plot of the v-tests of clinical features for the clustering results. Inferential conclusions can be made only for features that are not actively used to determine the clusters (i.e., disease). Legend of the p-values: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

subtype and the light smoker feature show a weak influence in cluster 1. For what concerns the Euclidean case, patients in cluster 1 are characterized by a higher presence of stage III LUSC disease; on the contrary in group 3, there is a higher concentration of Stage I LUSC patients. Cluster 2 shows v-test scores close to zero, indicating a weak relationship with the considered clinical features.

5 Conclusion and Future Work

In unsupervised learning tasks, such as clustering, the absence of an observed response feature makes the performance assessment a particularly challenging task. The method of choice inherently optimizes a specific criterion and so does the performance assessment metric. The best solution is, therefore, the one providing the most meaningful insights, which are often application- and domain-specific. It is crucial, however, to explore appropriate methods, measures and tools that align with the nature of the data at hand. There is no one-size-fits-all clustering method, nor a validity measure that universally identifies the optimal clustering solution. All decisions within an unsupervised learning pipeline should prioritize the interpretability of results. In other words, the clustering pipeline should be inherently iterative: methods and measures should be continuously refined to achieve practical interpretation while keeping their mathematical validity.

In hierarchical clustering, different solutions can be obtained, depending on the distance measure of choice, on the linkage function and on the selection of the final partition out of a hierarchy of partitions. Even within a chosen distance measure, hyperparameter settings may exist that ultimately affect the clustering solution (e.g., in the entropy-based distance, the choice of a different bandwidth or the kernel density estimator affects the final distance being computed). In this paper, we aim to explore a wide spectrum of clustering for mixed data solutions: in particular, a novel implementation of a permutation test-based procedure

enabling non-horizontal cuts of the dendrogram is proposed; furthermore, we explored four alternative approaches to compute pairwise distances in the case of mixed-type features. The application of multi-omics data pointed out how the choice of distance metric and of the clustering partition may affect the results and favor (or, undermine) the identification of informative insights.

A current line of research (see, e.g., Van de Velden et al., 2024b) concerns the comparative overview of distance computation for mixed-type data, including a systematic review of all the aspects that may lead: *(i)* to under/overemphasize the contribution to the overall distance of some of the features just because of their nature and *(ii)* to ignore or undermine the inter-dependencies among features of different type.

Future work will involve the evaluation of alternative test statistics for exploring a hierarchy of partitions: for example, using selective inference for hierarchical clustering (Gao et al., 2022), in particular, for the Monte Carlo approximation to the p-value. A further line of work is the development of CRAN-R packages for the implementation of efficient divisive clustering procedures (some implementations are available in Python (Anagnostou et al., 2022)) and for the re-sampling methods-based automatic identification of optimal cluster allocations.

Appendix 1

In this section, given the categorical feature j, we define ϕ_{ji} and w_{ji} quantities when i is categorical and when i is continuous. When both j and i are categorical, the empirical joint probability distributions of each categorical features pair are stored in the off-diagonal blocks of

$$\mathbf{P} = \frac{1}{n}\begin{bmatrix} \mathbf{Z}_1^\top\mathbf{Z}_1 & \mathbf{Z}_1^\top\mathbf{Z}_2 & \dots & \mathbf{Z}_1^\top\mathbf{Z}_{Q_c} \\ \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}_{Q_c}^\top\mathbf{Z}_1 & \mathbf{Z}_{Q_c}^\top\mathbf{Z}_2 & \dots & \mathbf{Z}_{Q_c}^\top\mathbf{Z}_{Q_c} \end{bmatrix}. \tag{17}$$

Let $\mathbf{p}_a^{ji}$ and $\mathbf{p}_b^{ji}$ be rows of $\mathbf{P}_{ji}$, off-diagonal block of $\mathbf{P}$; $\Phi_{ne}^{ji}(\mathbf{p}_a^{ji}, \mathbf{p}_b^{ji})$ is the normalized entropy of the joint distribution of the categories (a, b) of the j^{th} feature with the i^{th} feature. In particular,

$$\Phi_{ne}^{ji}(\mathbf{p}_a^{ji}, \mathbf{p}_b^{ji}) = \frac{\sum_{\ell=1}^{q_i} (\mathbf{p}_{a\ell}^{ji} + \mathbf{p}_{b\ell}^{ji}) log_2(\mathbf{p}_{a\ell}^{ji} + \mathbf{p}_{b\ell}^{ji})}{\log_2(q_i)}, \tag{18}$$

if the j^{th} attribute is categorical, and

$$\Phi_{ne}^{ji}(\mathbf{p}_a^{ji}, \mathbf{p}_b^{ji}) = \frac{\sum_{\ell=\min(a,b)}^{\max(a,b)-1} (\mathbf{p}_{a\ell(\ell+1)}^{ji} + \mathbf{p}_{b\ell(\ell+1)}^{ji}) log_2(\mathbf{p}_{a\ell(\ell+1)}^{ji} + \mathbf{p}_{b\ell(\ell+1)}^{ji})}{\log_2(q_i)},$$

if the j^{th} attribute is ordinal.

The quantity w_{ji} corresponds to the MI index between j and i, that is

$$w_{ji} = \sum_{\upsilon=1}^{q_j}\sum_{\ell=1}^{q_i} \mathbf{p}_{\upsilon\ell}^{ji} \log_2\left(\frac{\mathbf{p}_{\upsilon\ell}^{ji}}{\mathbf{p}_{\upsilon\cdot}^{ji}\mathbf{p}_{\cdot\ell}^{ji}}\right), \tag{19}$$

where $\mathbf{p}_{\upsilon.}^{ji}$ and $\mathbf{p}_{.\ell}^{ji}$ indicate the υ^{th} row margin and the ℓ^{th} column margin of $\mathbf{P}^{ji}$, respectively.

When the i^{th} feature is continuous, $i = 1, \ldots, Q_u$, the difference between the categories a and b of the attribute j depends on $f_a(X_i)$ and $f_b(X_i)$, the distributions of X_i conditional to the categories a and b, respectively. The Jensen-Shannon Distance is used, defined as

$$\Phi_{JS}^{ji}(f_a(X_i), f_b(X_i)) = \frac{1}{4}\sqrt{\Phi_{KL}^{ji}(f_a(X_i), f_{ab}(X_i)) + \Phi_{KL}^{ji}(f_{ab}(X_i), f_b(X_i))}, \quad (20)$$

where $f_{ab}(X_i) = \frac{f_a(X_i)+f_b(X_i)}{2}$ and $\Phi_{KL}^{ji}(f_a(X_i), f_b(X_i))$ is the Kullback-Leibler (KL) divergence

$$\Phi_{KL}^{ji}(f_a(X_i), f_b(X_i)) = \int f_a(x) \log_2 \frac{f_a(x)}{f_b(x)} dx. \quad (21)$$

The probability density functions $f_a(X_i)$ and $f_b(X_i)$ have to be estimated from the data, and a (normalized) kernel density estimator can be used (Mousavi and Sehhati, 2023). The MI-based quantity w_{ji} is defined as

$$w_{ji} = \frac{1}{n}\sum_{\ell=1}^{n}\left(\psi(n) - \psi(n_\ell^j) + \psi(k) - \psi(m_\ell)\right), \quad (22)$$

where $c_{j\ell}$ is the category of C_j of the ℓ^{th} observation, and n_ℓ^j its frequency; d_ℓ^i is the distance of the ℓ^{th} observation to the k^{th} neighbor (k is used defined) with respect to X_i; m_ℓ is the number of observations within a d_ℓ^i distance from the ℓ^{th} observation; finally, $\psi(\cdot)$ is the digamma distribution. Details on the derivation of Formula Eq. 22 are in Ross (2014). The Formula Eq. 7 is therefore rewritten as

$$\delta^j(a,b) = \sum_{i=1}^{Q_c} w_{ji}\Phi_{ne}^{ji}(\mathbf{p}_a^{ji}, \mathbf{p}_b^{ji}) + \sum_{i=Q_c+1}^{Q_c+Q_u} w_{ji}\Phi_{JS}^{ji}\left(\hat{f}_b(X_i), \hat{f}_a.(X_i)\right), \quad (23)$$

that is, the $(a,b)^{th}$ entry of the Δ_j, that is the j^{th} diagonal block of the Δ that is plugged in Formula Eq. 8 to obtain $\mathbf{D}_c$.

Appendix 2

In this section, hierarchical clustering is applied on both simulated and real datasets to assess the performance when used on the different mixed-type distances defined in Section 2.1. As the ground truth is available, we assess performances via the adjusted Rand Index (ARI).

The simulated data consist of two continuous and two categorical features divided into two clusters G_1, G_2 with varying sizes: in the balanced case $|G_1| = |G_2| = 500$ observations; in the unbalanced case $|G_1| = 400$ and $|G_2| = 100$. Different data-generating processes were considered for continuous and categorical features, more specifically:

- Continuous observations $\mathbf{X}_u = (X_1, X_2)$ are drawn from a bivariate Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Gaussian case) or from a skewed Gaussian $SN(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ (skewed Gaussian case (Sahu et al., 2003)).
- Observations from the categorical feature (C_j), $j = 1, 2$ are drawn from Multinomial distribution $Mult(n, \boldsymbol{\pi}_j)$, where $\boldsymbol{\pi}_j$ is a 4-dimensional vector.

The combinations of features/clusters dependencies are as follows:

a) Informative Continuous, Gaussian:

$$G_1: \ \mathbf{X}_u \sim N\left((0,0), \begin{pmatrix} 1.00 & 0.50 \\ 0.50 & 1.00 \end{pmatrix}\right),$$
$$G_2: \ \mathbf{X}_u \sim N\left((3,3), \begin{pmatrix} 0.66 & -0.50 \\ -0.50 & 0.66 \end{pmatrix}\right).$$

b) Uninformative Continuous, Gaussian:

$$G_1, G_2: \ \mathbf{X}_u \sim N\left((0,0), \begin{pmatrix} 1.00 & 0.50 \\ 0.50 & 1.00 \end{pmatrix}\right).$$

c) Informative Continuous, Skewed Gaussian:

$$G_1: \ \mathbf{X}_u \sim SN\left((0,0), \begin{pmatrix} 1.00 & 0.50 \\ 0.50 & 1.00 \end{pmatrix}, \begin{pmatrix} 30.00 & 0.00 \\ 0.00 & -30.00 \end{pmatrix}\right),$$
$$G_2: \ \mathbf{X}_u \sim SN\left((0,0), \begin{pmatrix} 0.66 & -0.50 \\ -0.50 & 0.66 \end{pmatrix}, \begin{pmatrix} -30.00 & 0.00 \\ 0.00 & 30.00 \end{pmatrix}\right).$$

d) Uninformative Continuous, Skewed Gaussian:

$$G_1, G_2: \ \mathbf{X}_u \sim SN\left((0,0), \begin{pmatrix} 1.00 & 0.50 \\ 0.50 & 1.00 \end{pmatrix}, \begin{pmatrix} 30.00 & 0.00 \\ 0.00 & -30.00 \end{pmatrix}\right).$$

e) Informative Categorical:

$$\begin{aligned} G_1: \ & C_1 \sim Mult\left(|G_1|, (0.7, 0.1, 0.1, 0.1)\right), \\ & C_2 \sim Mult\left(|G_2|, (0.1, 0.7, 0.1, 0.1)\right). \\ G_2: \ & C_1 \sim Mult\left(|G_1|, (0.1, 0.1, 0.7, 0.1)\right), \\ & C_2 \sim Mult\left(|G_2|, (0.1, 0.1, 0.1, 0.7)\right). \end{aligned}$$

f) Uninformative Categorical:

$$G_1, G_2: \ C_c \sim Mult\left(|G_c|, (0.25, 0.25, 0.25, 0.25)\right), \ \text{with } c = 1, 2.$$

Figures 6 and 7 depict the distributions of the considered features in the different settings.

The Ward linkage is used in the hierarchical clustering, and the distance data matrix is computed according to the considered measures. It is worth mentioning that, for the FAMD-based case, four principal components were considered; in the computation of entropy-based distance (Silverman, 1998; Mousavi and Sehhati, 2023), a Gaussian kernel with a default bandwidth proportional to the size and standard deviation of the features is considered, as in the original paper.

Table 3 reports the resulting ARI of the considered simulation schemes. The quality of the classification is slightly impacted by the use of asymmetrical continuous distributions; additionally, the distinctions between the approaches are unaffected.

The Euclidean-based solution shows poor ARI in each case, and the highest performance is in the case of skewed Gaussian continuous and informative categorical features. Hierarchical clustering with Gower distance shows similar behaviors as the Euclidean case in the considered scenarios, although with better performances. The dendrograms based on FAMD

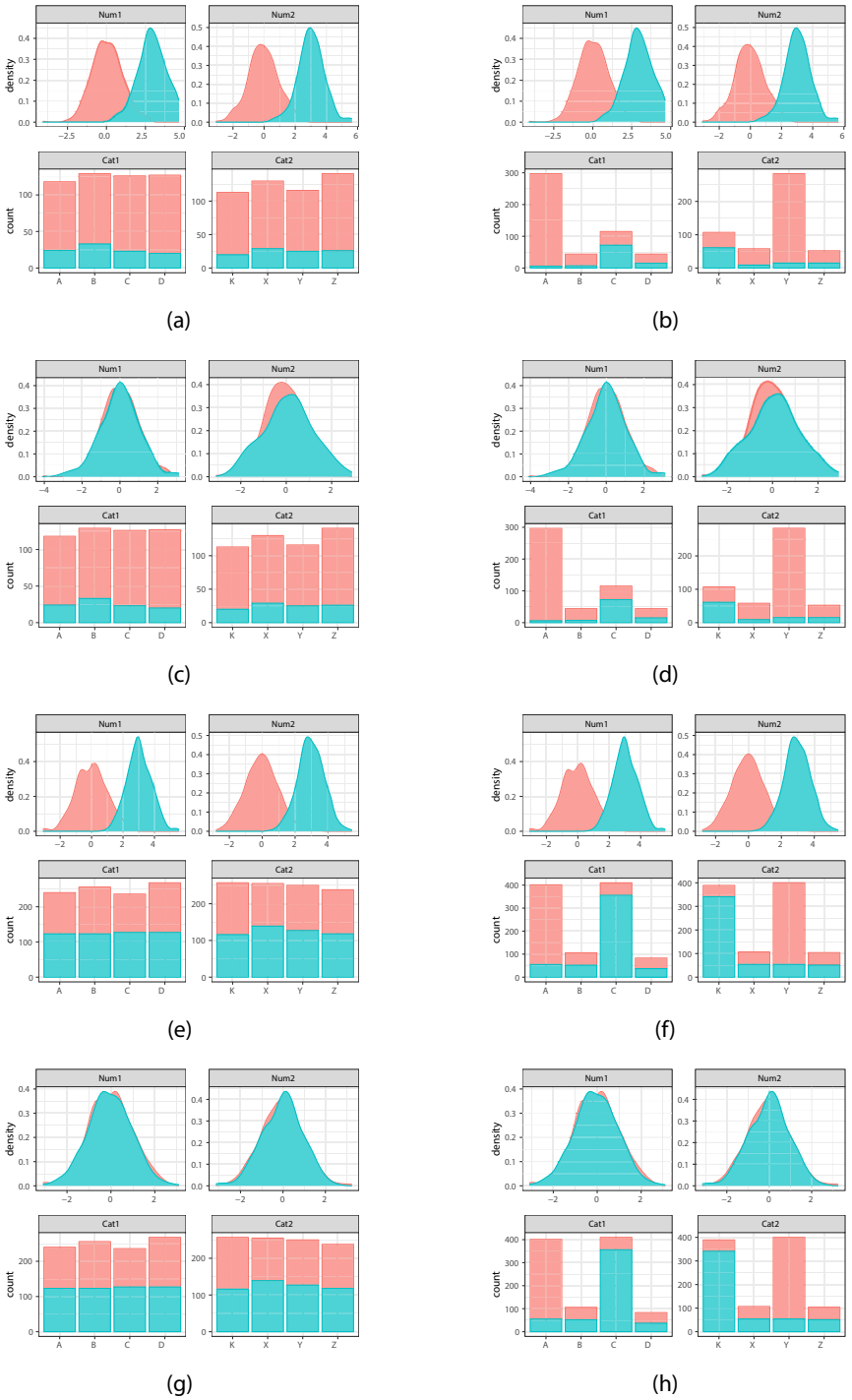

Fig. 6 Distribution of the simulated data schemes, Gaussian case

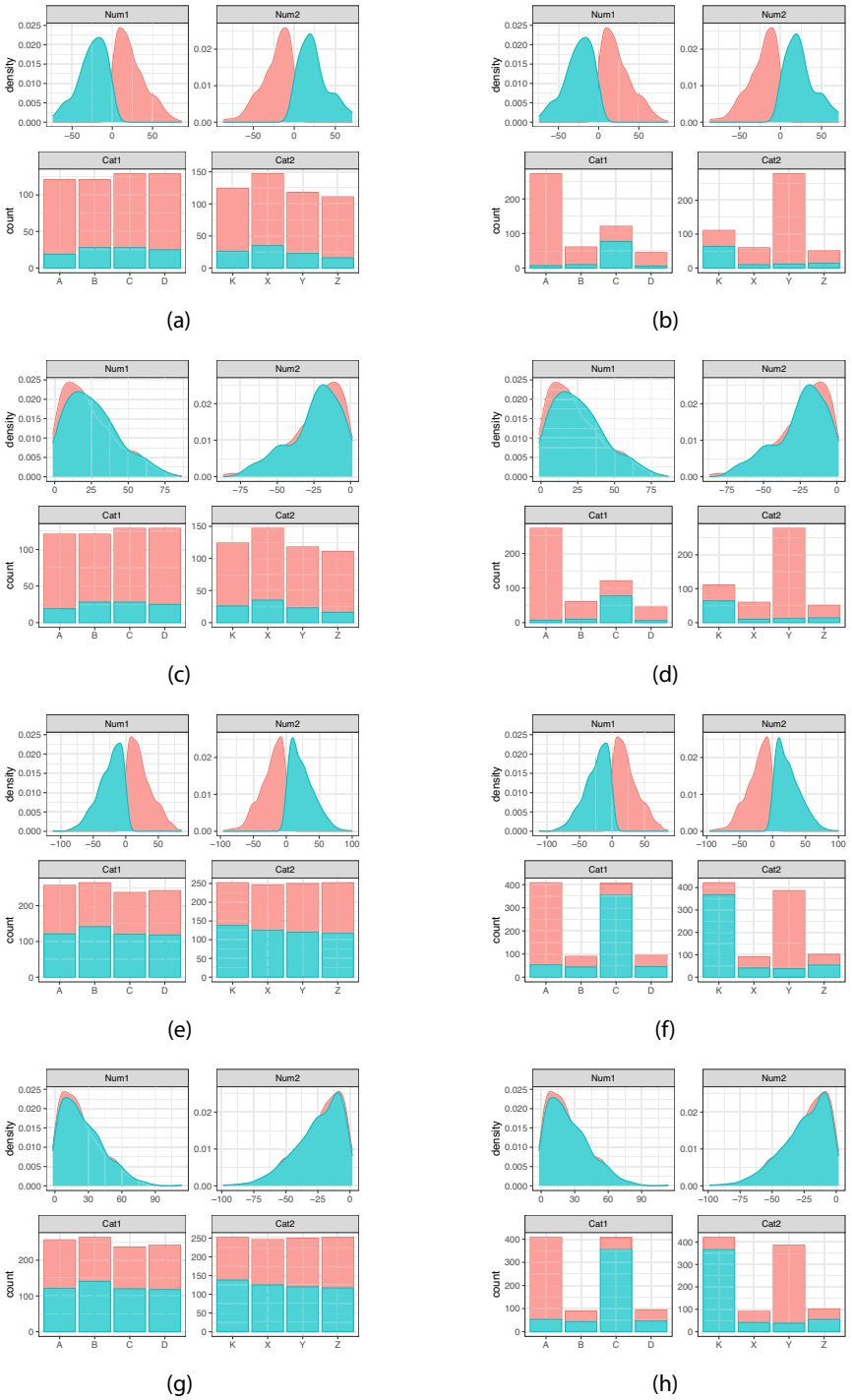

Fig. 7 Distribution of the simulated data schemes, skewed Gaussian case

Table 3 ARI of the simulated data with different feature settings

Sample setting			ARI			
			Euclidean	Gower-based	FAMD-based	Entropy-based
Gaussian Case						
• $\lvert G_1\rvert = 400, \lvert G_2\rvert = 100$						
Fig. 6a	Inf Cont	Uninf Cat	-0.01	-0.01	0.73	0.98
Fig. 6b	Inf Cont	Inf Cat	0.03	0.53	0.51	0.99
Fig. 6c	Uninf Cont	Uninf Cat	0.00	0.04	-0.01	0.01
Fig. 6d	Uninf Cont	Inf Cat	0.03	0.28	0.01	0.01
• $\lvert G_1\rvert = \lvert G_2\rvert = 500$						
Fig. 6e	Inf Cont	Uninf Cat	0.00	0.00	0.94	0.98
Fig. 6f	Inf Cont	Inf Cat	0.25	0.34	0.34	0.99
Fig. 6g	Uninf Cont	Uninf Cat	0.00	0.00	0.00	0.00
Fig. 6h	Uninf Cont	Inf Cat	0.24	0.36	0.24	0.00
Skewed Gaussian Case						
• $\lvert G_1\rvert = 400, \lvert G_2\rvert = 100$						
Fig. 7a	Inf Cont	Uninf Cat	0.00	-0.04	0.89	0.99
Fig. 7b	Inf Cont	Inf Cat	0.12	0.17	0.73	0.99
Fig. 7c	Uninf Cont	Uninf Cat	-0.04	0.01	-0.01	-0.01
Fig. 7d	Uninf Cont	Inf Cat	-0.03	0.17	0.17	0.02
• $\lvert G_1\rvert = \lvert G_2\rvert = 500$						
Fig. 7e	Inf Cont	Uninf Cat	0.00	0.00	0.93	0.98
Fig. 7f	Inf Cont	Inf Cat	0.27	0.39	0.42	1.00
Fig. 7g	Uninf Cont	Uninf Cat	0.00	0.00	0.00	0.00
Fig. 7h	Uninf Cont	Inf Cat	0.25	0.39	0.33	0.01

Note: The datasets consist of two continuous (Gaussian or skewed Gaussian) and two categorical features with different settings, i.e., Informative (Inf) or Uninformative (Uninf) and balanced or unbalanced clusters. In each scenario, four different distance matrices are computed, and for each of them, a dendrogram is grown; observations are then split into two clusters with a dendrogram cut. The feature distributions over the different settings are depicted in Figs. 6 and 7

distances generate clusters that perform well mainly in the case of informative continuous features. Overall, when the continuous features are informative, the entropy-based distance outperforms the remaining cases. On the contrary, with poor informative continuous features,

Table 4 Adjusted Rand Index of the real data

Dataset	Obs	k	Q_u	Q_c	ARI			
					Euclidean	Gower-based	FAMD-based	Entropy-based
celev_heart	303	2	5	9	0.16	0.29	0.32	0.29
credit	653	2	6	10	0.00	0.01	0.00	0.38
dermatology	366	6	1	34	0.80	0.73	0.76	0.73
german	1000	2	7	14	−0.03	0.05	0.04	0.07
adult2_dep	4000	2	6	9	−0.02	0.17	0.01	0.18
australian	690	2	6	9	−0.01	0.42	-0.01	0.24

Table 5 Rand Index of the real data

					RI			
Dataset	Obs	k	Q_u	Q_c	Euclidean	Gower-based	FAMD-based	Entropy-based
celev_heart	303	2	5	9	0.58	0.65	0.66	0.64
credit	653	2	6	10	0.50	0.50	0.50	0.69
dermatology	366	6	1	34	0.93	0.91	0.92	0.91
german	1000	2	7	14	0.56	0.57	0.56	0.57
adult2_dep	4000	2	6	9	0.58	0.59	0.62	0.59
australian	690	2	6	9	0.50	0.71	0.50	0.62

Gower and FAMD are preferred since the entropy-based approach is not capable of catching enough information from the categorical features.

The six mixed-type datasets used in Mousavi and Sehhati (2023) are the real datasets considered for performance assessment. Similar to the simulated data analysis, the default bandwidth for the Gaussian kernel is used in the computation of the entropy-based distance. For the FAMD-based distance computation, components are retained that cumulatively explain 75% of the total variability.

Tables 4 and 5 present the main results in terms of ARI and RI, together with the size of each dataset (Obs), the number of true clusters (k), and the number of continuous and categorical features (Q_u and Q_c, respectively). The Euclidean-based approach produces poor performances in most of the considered datasets but "dermatology" presents, however, 34 out of 35 categorical features. The entropy-based solutions have the best performances in terms of ARI in three of the six considered datasets ("credit," ARI = 0.38; "german," ARI = 0.07; "adult_2dep," ARI = 0.18). It is worth noting that, in terms of RI results, in all but "german" and "adult2_dep" datasets, there are discrepancies with the RIs reported in Mousavi and Sehhati (2023)[7]. While the entropy-based distance is the same, the applied clustering procedure is different: in Mousavi and Sehhati (2023) paper, spectral clustering is used, whereas in this paper, hierarchical clustering is used.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00357-025-09516-3.

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

[7] We report here the spectral clustering-based RI scores from Mousavi and Sehhati (2023) original paper: "celev_heart," 0.7157; "credit," 0.7327; "dermatology," 0.9815; and "australian," 0.7189.

References

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering, 63*(2), 503–527.

Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *Ieee Access, 7*, 31883–31902.

Anagnostou, P., Tasoulis, S., Plagianakos, V., et al. (2022). Hipart: Hierarchical divisive clustering toolbox. arXiv:2209.08680

Bhattacharjee, P., & Mitra, P. (2021). A survey of density based clustering algorithms. *Frontiers of Computer Science, 15*, 1–27.

Boley, D. (1998). Principal direction divisive partitioning. *Data Mining And Knowledge Discovery, 2*, 325–344.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis, 71*, 52–78.

Bruzzese, D., & Vistocco, D. (2015). Despota: Dendrogram slicing through a pemutation test approach. *Journal of Classification, 32*, 285–304.

Chavent, M., Kuentz-Simonet, V., Labenne, A., et al. (2014). Multivariate analysis of mixed data: The r package pcamixdata. arXiv:1411.4911

Chen, J. W., & Dhahbi, J. (2021). Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Scientific Reports, 11*(1), 13323.

Chu, K., Zhang, M., Xun, Y., et al. (2024). A hybrid similarity measure-based clustering approach for mixed attribute data. *International Journal of Machine Learning and Cybernetics, 15*(4), 1295–1311.

De Leeuw, J., & Van Rijckevorsel, J. (1980). Homals and Princals—Some generalizations of principal components analysis. *Data Analysis And Informatics, 2*, 231–242.

Ellen, J. G., Jacob, E., Nikolaou, N., et al. (2023). Autoencoder-based multimodal prediction of non-small cell lung cancer survival. *Scientific Reports, 13*(1), 15761.

Foss, A. H., Markatou, M., & Ray, B. (2019). Distance metrics and clustering methods for mixed-type data. *International Statistical Review, 87*(1), 80–109.

Gao, LL., Bien, J., Witten, D. (2022). Selective inference for hierarchical clustering. *Journal of the American Statistical Association* pp. 1–11

Good, P. (2013). Permutation tests: a practical guide to resampling methods for testing hypotheses. *Springer Science & Business Media*

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*(4), 857–871.

Greenacre, M., Groenen, P. J., Hastie, T., et al. (2022). Principal component analysis. *Nature Reviews Methods Primers, 2*(1), 100.

Hill, M., Smith, A. (1976). Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon* pp. 249–255

Horst, AM., Hill, AP., Gorman, KB. (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. https://doi.org/10.5281/zenodo.3960218, https://allisonhorst.github.io/palmerpenguins/, r package version 0.1.0

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values," proceedings of 1st pacific-asia conference on knowledge discovery and data mining

Husson, F., Lê, S., & Pagès, J. (2017). *Exploratory multivariate analysis by example using R*. CRC Press.

Jung, Y., Park, H., Du, D. Z., et al. (2003). A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization, 25*(1), 91–111.

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

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions On Information Theory, 28*(2), 129–137.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, pp. 281–297

McMurdie, PJ., Holmes, S. (2012). Phyloseq: A bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. In: Biocomputing 2012. *World Scientific*, p. 235–246

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification, 33*, 331–373.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159–179.

Mousavi, E., & Sehhati, M. (2023). A generalized multi-aspect distance metric for mixed-type data clustering. *Pattern Recognition, 138*, Article 109353.

Pagès, J. (2004). Analyse factorielle de donnees mixtes: principe et exemple d'application. *Revue de statistique appliquée, 52*(4), 93–111.
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
Ran, X., Xi, Y., Lu, Y., et al. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review, 56*(8), 8219–8264.
Ring, M., Otto, F., Becker, M., et al. (2015). Condist: A context-driven categorical distance measure. In A. Appice, P. Rodrigues, V. Santos Costa, et al. (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 251–266). Cham: Springer International Publishing.
Ritchie, M. D., Holzinger, E. R., Li, R., et al. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics, 16*(2), 85–97.
Ross, B. (2014). Mutual information between discrete and continuous data sets. *PloS one, 9*(2), Article e87357.
Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal Of Computational And Applied Mathematics, 20*, 53–65.
Roux, M. (2018). A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification, 35*, 345–366.
Sahu, S. K., Dey, D. K., & Branco, M. D. (2003). A new class of multivariate skew distributions with applications to bayesian regression models. *Canadian Journal of Statistics, 31*(2), 129–150.
Savaresi, SM., Boley, DL. (2001). On the performance of bisecting k-means and pddp. In: Proceedings of the 2001 SIAM International Conference on Data Mining, SIAM, pp. 1–14
Silverman, B. W. (1998). *Density estimation for statistics and data analysis*. Routledge.
Tasoulis, S. K., Tasoulis, D. K., & Plagianakos, V. P. (2010). Enhancing principal direction divisive clustering. *Pattern Recognition, 43*(10), 3391–3411.
Van de Velden, M., Iodice D'Enza, A., & Markos, A. (2019). Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics, 11*(3), Article e1456.
Van de Velden, M., Iodice D'Enza, A., Markos, A., et al. (2024). A general framework for implementing distances for categorical variables. *Pattern Recognition, 153*, Article 110547.
Van de Velden, M., Iodice D'Enza, A., Markos, A., et al. (2024b). Unbiased mixed variables distance. arXiv:2411.00429
Wang, Y., Xiao, X., & Li, Y. (2023). Construction and validation of a cuproptosis-related lncrna signature for the prediction of the prognosis of luad and lusc. *Scientific Reports, 13*(1), 2477.
Zeimpekis, D., Gallopoulos, E. (2008). Principal Direction Divisive Partitioning with Kernels and k-Means Steering, Springer London, London, pp. 45–64.https://doi.org/10.1007/978-1-84800-046-9_3,