Multi-view cluster aggregation and splitting, with an application to multi-omic breast cancer data

Antoine Godichon-Baggioni, Cathy Maugis-Rabusseau, Andrea Rau

Multi-view cluster aggregation and splitting, with an application to multi-omic breast cancer data

Antoine Godichon-Baggioni ,
Cathy Maugis-Rabusseau and Andrea Rau

Laboratoire de Probabilités, Statistique et Modélisation; UMR8001;
Sorbonne Université;
75005 Paris, France
e-mail: antoine.godichon_baggioni@upmc.fr

Institut de Mathématiques de Toulouse; UMR5219;
Université de Toulouse;
INSA, F-31077 Toulouse, France
e-mail: cathy.maugis@insa-toulouse.fr

GABI, INRA, AgroParisTech,
Université Paris-Saclay, Paris, France
e-mail: andrea.rau@inra.fr

Abstract: Multi-view data, which represent distinct but related groupings of variables, can be useful for identifying relevant and robust clustering structures among observations. A large number of multi-view classification algorithms have been proposed in the fields of computer science and genomics; in this work, we instead focus on the task of merging or splitting an existing hard or fuzzy cluster partition based on multi-view data. This work is specifically motivated by an application involving multi-omic breast cancer data from The Cancer Genome Atlas, where multiple molecular profiles (gene expression, miRNA expression, methylation, and copy number alterations) are used to further subdivide the five currently accepted intrinsic tumor subtypes into clinically distinct sub-groups of patients. In addition, we investigate the performance of the proposed multi-view splitting and aggregation algorithms, as compared to single- and concatenated-view alternatives, in a set of simulations. The multi-view splitting and aggregation algorithms developed in this work are implemented in the *maskmeans* R package.

MSC 2010 subject classifications: Primary 62H30, 62P10; secondary 92D10.
Keywords and phrases: Clustering, multi-view, cluster merging and splitting, multi-omics data, TCGA.

1. Introduction

Multi-view clustering refers to the problem of identifying distinct groupings of observations from data that consist of multiple related sets of features, or views; in biology, these clustering approaches have been particularly motivated by the emergence of multi-view datasets in genomics (e.g., where gene expression, copy number alterations, and methylation are measured on the same individuals,

Chao et al., 2017) and neuroinformatics (e.g., functional magnetic resonance imaging, Fratello et al., 2017). One of the underlying assumptions of such approaches is that multi-faceted and heterogeneous views of the same problem can be useful in identifying or refining relevant and robust clustering structures, as they may reflect different aspects of complex clustering structures. Multi-view learning thus falls under the broader umbrella of so-called intermediate integrative analyses (Hamid et al., 2009), in which rather than being simply concatenated together or analyzed in isolation, each view is permitted to "speak for itself" using weights, transformations, or model-based approaches to combine results across views.

Multi-view classification algorithms have been the focus of an extensive amount of research in the field of computer science in recent years; see Xu et al. (2013) and Chao et al. (2017) for reviews and discussion of the current state-of-the-art. Dimensionality reduction is a common feature of such algorithms due to the high dimensionality of data, and potentially different dimensionality among views. Existing methods make use of a variety of approaches, including spectral clustering (Kumar et al., 2011, Kumar and Daumé, 2011), bi-clustering (Koutsonikola and Vakali, 2009, Pensa et al., 2005), and density-based clustering of multi-view data (Kailing et al., 2004, Taskesen et al., 2016). Cai et al. (2013) proposed the multi-view K-means algorithm as a robust and computationally efficient method to cluster large-scale heterogeneous multi-view datasets; Chen et al. (2013) extended this idea to incorporate weights on both views and variables. Multi-view clustering techniques have also been specifically developed in the context of multiple high-throughput molecular assays; see Rappoport and Shamir (2018) for a detailed review. For example, Serra et al. (2015) proposed the MVDA approach in which membership matrices from individual omics are integrated into a single robust patient subtype, and Yang and Michailidis (2016) used non-negative matrix factorization to jointly decompose multi-view omics data. The iCluster+ approach (Shen et al., 2009, Mo et al., 2013, Shen et al., 2012) uses a joint latent-variable model to cluster multi-omics data, while SNF (Wang et al., 2014) combines omic information using a network-based approach to identify patient subtypes.

To our knowledge, these multi-view classification techniques focus on either *de novo* unsupervised clustering or supervised clustering of a multi-view dataset; here, we instead focus on the task of merging or splitting an existing hard or fuzzy cluster partition based on multi-view data. Merging/splitting can address the question of selecting the ideal number of clusters, or can be of interest when an initial overly simplistic or complex clustering is available. For instance, to address the overestimation of the number of clusters in a Gaussian mixture model as determined by the Bayesian information criterion, Baudry et al. (2010) proposed a method to hierarchically aggregate components using an entropy criterion to obtain a soft clustering for each number of clusters less than or equal to the initial number. The recently proposed *clustree* R package (Zappia and Oshlack, 2018) takes a different approach by providing a graphical approach to visualize different clustering resolutions.

In this work, we address the specific problem of aggregating or splitting an

existing initial data partition in the multi-view framework; the initial partition of data may represent a clustering of a single data view, or alternatively can represent a pre-existing grouping of individuals. This work is specifically motivated by an application involving multi-omic breast cancer data, where multiple omics profiles are used to further subdivide intrinsic tumor subtypes into clinically distinct sub-groups of patients. In particular, rather than focusing on a *de novo* clustering of patients, we instead seek to further subdivide a pre-established grouping of individuals. The remainder of this work has been organized as follows: the multi-omic breast cancer data that are the focus of our study are described in Section 2. The multi-view K-means algorithm, as well as the multi-view splitting and aggregation approaches, are described in detail in Section 3. In Sections 4 and 5, the proposed methods are benchmarked on simulated data, and results on the multi-omic breast cancer data are described in detail. Finally, a discussion and some conclusions are provided in Section 6.

2. Multi-omic breast cancer data

In women, breast cancer is the most commonly diagnosed cancer and is the leading cause of cancer death worldwide; according to the *GLOBOCAN 2018* estimates of cancer incidence and mortality, there will be about 2.1 million newly diagnosed cases worldwide in 2018 alone (Bray et al., 2018). Multiple distinct forms, or subtypes, of the disease, corresponding to both morphological and clinical heterogeneity, as well as significantly different reactions to treatment and prognosis, have been identified. In particular, molecular profiling, typically based on gene expression data, may be used to characterize breast tumors beyond classifiers such as clinical prognosis, grade, histology, and immunohistochemical analysis of estrogen and progesterone receptors (ER/PR) and human epidermal growth factor receptor-2 (HER2) over-expression (Perou et al., 2000). A robust and stable classification of intrinsic breast cancer subtypes can be inferred from gene expression profiles using the AIMS approach (Paquet and Hallett, 2000), leading to the five commonly accepted intrinsic subtypes of Luminal A and B, Basal-like, HER2-enriched, and Normal-like tumors. However, significant phenotypic heterogeneity has been observed even within these subtypes; for example, The Cancer Genome Atlas Network (2012) found that ER+ tumors (Luminal A and B) were the most heterogenous in terms of gene expression, mutation spectrum, copy number changes, and patient outcome.

The Cancer Genome Atlas (TCGA) represents a vast and valuable resource for pan-cancer genomic studies, including multi-omic molecular profiles of tumor samples, and in some cases matched normal samples, for over 30 different cancer types and over 11,000 individuals (The Cancer Genome Atlas Network et al., 2013). The public availability of the open-access tier of TCGA data has led to an explosion of research in cancer informatics and methodological developments for multi-omic data. In this paper, we focus on the multi-omic profiles (gene expression, microRNA expression, promoter methylation, and copy number alterations [CNA]) measured for 20,179 genes in 506 breast cancer patients in

the TCGA database. Details about TCGA data acquistion and pre-processing may be found in Rau et al. (2018). Briefly, gene and miRNA expression were measured in tumor samples using RNA-seq and miRNA-seq, and normalized abundance estimates were log-transformed after adding a constant of 1. Promoter methylation in tumor samples for each gene was measured using an Illumina Infinium Human Methylation450 BeadChip array, and probe values were logit-transformed. Somatic copy number gains and losses were quantified by comparing Affymetrix 6.0 probe intensities in matched normal and cancer tissue and aggregating measures to gene-level. Intrinsic breast cancer subtypes were inferred from the RNA-seq data using the *AIMS* Bioconductor package (Paquet and Hallett, 2000).

The ultimate goal in this work is to determine whether the use of multi-view cluster splitting of the inferred intrinsic breast cancer subtypes, based on RNA-seq, miRNA-seq, promoter methylation, and copy number alterations, can lead to robust and clinically meaningful sub-clusterings of patients. To this end, we focus on a subset of 226 genes that play an important role in breast carcinogenesis, corresponding to the *TP53* and *MKI67* genes (respectively a tumor suppressor and a cellular marker for proliferation), those in the estrogen signaling and ErbB signaling pathways from the KEGG database (Kanehisa et al., 2016), and those in the SAM40 DNA methylation signature (Fleischer et al., 2017). Of these, 226, 199 and 222 respectively had gene expression, methylation and CNA measurements available. In addition, we retained only the 149 miRNAs for which the average normalized expression across all 506 patients was greater than 50.

3. Multi-view clustering algorithms

To build up to our proposed multi-view aggregation and splitting procedure, the latter of which we will ultimately seek to apply to the TCGA breast cancer data, we first introduce the framework with some notation. Because the algorithm can be defined for both fuzzy and hard initial clusterings, we restrict our description in the manuscript to the former as it represents a generalization of the latter.

3.1. Framework and data scaling

In the clustering setting, we consider a data matrix $Z \in \mathbb{R}^{n\times d}$ with n individuals described by d quantitative measures, decomposed into V views:

$$Z = \left(Z^{(1)}, \ldots, Z^{(v)}, \ldots, Z^{(V)}\right),$$

where $Z^{(v)} \in \mathbb{R}^{n\times d_v}$ and $d = \sum_{v=1}^{V} d_v$. As in Cai et al. (2013), the data here are assumed to have been scaled to unit-variance. Moreover, in order to avoid problems due to the potentially different dimensionality for each view, each scaled variable is also divided by the size of its corresponding view:

$$X^{(v)} = \frac{Z^{(v)}}{d_v}.$$

We assume that an initial clustering of the n individuals, obtained with an arbitrary clustering algorithm on external data or one of the V views, is available. This initial clustering may be either a hard partition or a fuzzy clustering. In the latter case, we have an initial matrix $\Pi^{[0]} = \Pi_{K_{init}} = (\pi_{i,k}^{(0)})$ where $\pi_{i,k}^{(0)}$ is the "probability" (weight) that the i-th individual belongs to the k-th cluster, and $\sum_{k=1}^{K^{(0)}} \pi_{i,k}^{(0)} = 1$ for each individual i. In the hard clustering case, $\Pi_{K_{init}}$ is a $0-1$ matrix with a single 1 in each row.

The aggregation and splitting procedures presented hereafter are respectively based on the minimization of a criterion that is inspired by the one used in the multi-view fuzzy K-means algorithm (Wang and Chen, 2017):

$$\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{v=1}^{V}(\alpha_v)^{\gamma}(\pi_{i,k})^{\delta}\left\|X_i^{(v)}-\mu_k^{(v)}\right\|^2, \tag{3.1}$$

where $\gamma > 1$, $\delta > 1$, and $\mu = (\mu_1, \ldots, \mu_K)$ is the vector of cluster centers such that $\mu_k = (\mu_k^{(1)}, \ldots, \mu_k^{(V)})$. The vector $\underline{\alpha} = (\alpha_1, \ldots, \alpha_V)$, with $\sum_{v=1}^{V} \alpha_v = 1$, contains the weight of each view that allows more or less importance to be attributed to each view in the clustering process. The δ parameter tunes the weights on the fuzzy classifications Π_K, with larger values yielding larger weights for large probabilities of cluster membership; similarly, the γ parameter tunes the per-view weights, with larger values flattening out the view-specific contributions to the criterion value.

3.2. Multi-view splitting

In this section, the aim is, starting from an initial fuzzy clustering matrix $\Pi^{[0]}$, to successively split clusters in order to minimize the following criterion :

$$\text{Split}\,(\Pi_K, \underline{\alpha}, \mu) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{v=1}^{V}(\alpha_{k,v})^{\gamma}(\pi_{i,k})^{\delta}\left\|X_i^{(v)}-\mu_k^{(v)}\right\|^2, \tag{3.2}$$

under the constraints $\forall i,\ \sum_{k=1}^{K} \pi_{i,k} = 1$ and $\forall k,\ \sum_{v=1}^{V} \alpha_{k,v} = 1$. We remark that minimizing this criterion, given Π_K, leads to $\mu = (\mu_1, \ldots, \mu_K)$ with $\mu_k = \sum_{i=1}^{n}(\pi_{i,k})^{\delta} X_i / \sum_{i=1}^{n}(\pi_{i,k})^{\delta}$. Note that Criterion (3.2) is identical to Criterion (3.1), where the per-view weights α_v have been replaced here with per-cluster and per-view weights $\alpha_{k,v}$; this allows views to be up- or down-weighted for a specific cluster when they contain only partially relevant information about the underlying cluster structure. By default, in this work we set both γ and δ to be equal to 2.

In order to minimize Criterion (3.2), we propose an iterative algorithm described in Algorithm 1. At each step, we must identify the cluster $\mathcal{C}_{\hat{k}}$ such that

$$\hat{k} = \arg\max_{1\le k\le K}\sum_{i=1}^{n}\sum_{v=1}^{V}(\alpha_{k,v})^{\gamma}(\pi_{i,k})^{\delta}\left\|X_i^{(v)}-\mu_k^{(v)}\right\|^2.$$

Subsequently, this cluster must be split into two clusters, $\tilde{\mathcal{C}}_{k_1}$ and $\tilde{\mathcal{C}}_{k_2}$, which minimize

$$\sum_{\ell=k_1,k_2}\sum_{v=1}^{V}\sum_{i=1}^{n}\left(\alpha_{\hat{k},v}\right)^{\gamma}\left(\pi_{i,\ell}\right)^{\delta}\left\|X_i^{(v)}-\tilde{\mu}_{\ell}^{(v)}\right\|^2,$$

under the constraint $\pi_{i,k_1}+\pi_{i,k_2}=\pi_{i,\hat{k}}$ for each $i=1,\ldots,n$. This step provides a new fuzzy clustering matrix $\tilde{\Pi}_{K+1}$ and the associated vector of cluster centers $\tilde{\mu}$. It is detailed in Appendix E. Then, one can obtain the weight matrix $\underline{\tilde{\alpha}}$ associated with this split defined for all $k=1,\ldots K+1$ and for all $v=1,\ldots,V$;

$$\tilde{\alpha}_{k,v}=\frac{\left(\sum_{i=1}^{n}\left(\tilde{\pi}_{i,k}\right)^{\delta}\left\|X_i^{(v)}-\tilde{\mu}_{k}^{(v)}\right\|^2\right)^{\frac{1}{1-\gamma}}}{\sum_{l=1,\ldots,K+1}\left(\sum_{i=1}^{n}\left(\tilde{\pi}_{i,l}\right)^{\delta}\left\|X_i^{(v)}-\tilde{\mu}_{l}^{(v)}\right\|^2\right)^{\frac{1}{1-\gamma}}}.$$

Proposition 3.1. *Let K be a positive integer, and let Π_K be a fuzzy clustering matrix with K clusters. Let $\hat{k}\in\{1,\ldots,K\}$ and $\tilde{\Pi}_{K+1}$ be the fuzzy clustering matrix obtained by splitting the cluster $\mathcal{C}_{\hat{k}}$. Then, for any weight matrix $\underline{\alpha}$,*

$$\mathit{Split}\left(\Pi_K,\underline{\alpha},\mu\right)\geq\mathit{Split}\left(\tilde{\Pi}_{K+1},\underline{\tilde{\alpha}},\tilde{\mu}\right).$$

The proof is given in Appendix C.

- **Step** $t=0$**:** Let $\Pi_{K_{\text{init}}}=\left(\pi_{i,k}^{[0]}\right)_{i,k}$ be the initial fuzzy clustering matrix.
  - *Initialization of the centers*: for all $k=1,\ldots,K_{\text{init}}$,
  $$\mu_{k,[0]}=\sum_{i=1}^{n}\left(\pi_{i,k}^{[0]}\right)^{\delta}X_i\Big/\sum_{i=1}^{n}\left(\pi_{i,k}^{[0]}\right)^{\delta}$$
  - *Initialization of the weight matrix* $\underline{\alpha}^{[0]}=\left(\alpha_{k,v}^{[0]}\right)$: for all $v=1,\ldots,V$ and $k=1,\ldots,K_{\text{init}}$,
  $$\alpha_{k,v}^{[0]}=\frac{\left(\sum_{i=1}^{n}\left(\pi_{i,k}^{[0]}\right)^{\delta}\left\|X_i^{(v)}-\mu_{k,[0]}^{(v)}\right\|^2\right)^{\frac{1}{1-\gamma}}}{\sum_{v'=1}^{V}\left(\sum_{i=1}^{n}\left(\pi_{i,k}^{[0]}\right)^{\delta}\left\|X_i^{(v')}-\mu_{k,[0]}^{(v')}\right\|^2\right)^{\frac{1}{1-\gamma}}}.$$
- **Step** $t\geq 1$ **:**
  - *Update the fuzzy clustering matrix* $\Pi^{[t]}$*, centers and the weight matrix* $\underline{\alpha}^{[t]}$: split cluster $C_{\hat{k}}$ into two clusters, where
  $$\hat{k}=\arg\max_{k}\sum_{v=1}^{V}\sum_{i=1}^{n}\left(\alpha_{k,v}^{[t-1]}\right)^{\gamma}\left(\pi_{i,k}^{[t-1]}\right)^{\delta}\left\|X_i^{(v)}-\mu_{k,[t-1]}^{(v)}\right\|^2.$$

Algorithm 1: Description of the fuzzy multi-view splitting algorithm.

Remark 3.1. *Note that a version of this algorithm with common weights per cluster ($\alpha_{k,v} = \alpha_v$ for all k) as well as a version for hard clustering matrix can immediately be derived from that described here.*

3.3. Multi-view aggregation

Starting from an initial clustering matrix $\Pi^{[0]} = \Pi_{K_{\text{init}}}$, we now wish to construct a hierarchical aggregation while accounting for the information available in the different data views. At each step, the aim is to aggregate the pair of clusters that corresponds to a minimal increase of the following criterion:

$$\text{Agg}\left(\Pi_K, \underline{\alpha}, \mu\right) = \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{v=1}^{V}(\alpha_v)^{\gamma}\pi_{i,k}\left\|X_i^{(v)} - \mu_k^{(v)}\right\|^2, \tag{3.3}$$

with $\mu_k = \sum_{i=1}^{n}\pi_{i,k}X_i / \sum_{i=1}^{n}\pi_{i,k}$. Given a fuzzy clustering matrix $\Pi_K = (\pi_{i,k})_{i=1,\ldots,n,k=1,\ldots,K}$, we aggregate two clusters $\mathcal{C}_k$ and $\mathcal{C}_{k'}$ $(k \neq k')$ into a new cluster $\mathcal{C}_{k\cup k'}$ by constructing a new clustering matrix $(\tilde{\Pi}_{K-1,k\cup k'})$ with $K-1$ clusters, such that $\tilde{\pi}_\ell = \pi_\ell$ when $\ell \neq k, k'$ and $\tilde{\pi}_{k\cup k'} = \pi_k + \pi_{k'}$. The algorithm is detailed in Algorithm 2. By setting $\delta = 1$, the following proposition enables us to ensure and quantify the decrease of Criterion (3.3) when two clusters are aggregated.

Proposition 3.2. *Let K be a positive integer, $\Pi_K = (\pi_{i,k})_{i=1,\ldots,n,k=1,\ldots,K}$ be a fuzzy clustering matrix, and set $\delta = 1$. Let $k, k' \in \{1,\ldots,K\}$, such that $k \neq k'$. If two clusters $\mathcal{C}_k$ and $\mathcal{C}_{k'}$ are aggregated, then for all weight vectors $\underline{\alpha} = (\alpha_1,\ldots,\alpha_V)$,*

$$\text{Agg}\left(\Pi_K, \underline{\alpha}, \mu\right) - \text{Agg}\left(\tilde{\Pi}_{K-1,k\cup k'}, \underline{\alpha}, \tilde{\mu}\right) = -\frac{n_k n_{k'}}{n_k + n_{k'}}\sum_{v=1}^{V}(\alpha_v)^{\gamma}\left\|\mu_k^{(v)} - \mu_{k'}^{(v)}\right\|^2 \leq 0, \tag{3.4}$$

where $n_k = \sum_{i=1}^{n}\pi_{i,k}$, and $\mu = (\mu_1,\ldots,\mu_K)$ and $\tilde{\mu} = (\tilde{\mu}_1,\ldots,\tilde{\mu}_{K-1})$ are respectively associated with Π_K and $\tilde{\Pi}_{K-1}$.

The proof is given in Appendix D. Then, the multi-view aggregation algorithm consists of aggregating at each step the two clusters for which the minimal increase is obtained.

Compared to the usual aggregation algorithm using the Ward distance, the primary novelty here is that we directly account for the quality of the initial clustering in each view via the weight vector $\underline{\alpha}$. For example, in the case where $\gamma = 2$, the weights correspond to the ratio of the inverse sum of squared errors in one view to that summed across all views; as such, if the clustering pattern of one view is in complete disagreement with the others, it will tend to have a messy clustering and will thus be down-weighted with respect to the other views. Similarly, using an initial clustering constructed on one specific view typically yields larger weights for that view (and smaller weights for highly dissimilar views) in early stages of the aggregation algorithm.

- **Step** $t = 0$**:** Let $\Pi_{K_{\text{init}}} = \left(\pi_{i,k}^{[0]}\right)_{i,k}$ be the initial fuzzy clustering matrix.
  - *Initialization of the centers*: for all $k = 1, \ldots, K_{\text{init}}$,
  $$\mu_{k,[0]} = \sum_{i=1}^{n} \pi_{i,k}^{[0]} X_i \Big/ \sum_{i=1}^{n} \pi_{i,k}^{[0]}.$$
  - *Initialization of the weight vector:* $\underline{\alpha}^{[0]} = \left(\alpha_1^{[0]}, \ldots, \alpha_V^{[0]}\right)$ where for all $v = 1, \ldots, V$,
  $$\alpha_v^{[0]} = \frac{\left(\sum_{k=1}^{K_{init}} \sum_{i=1}^{n} \pi_{i,k}^{[0]} \left\| X_i^{(v)} - \mu_{k,[0]}^{(v)} \right\|^2\right)^{\frac{1}{1-\gamma}}}{\sum_{v'=1}^{V} \left(\sum_{k=1}^{K_{init}} \sum_{i=1}^{n} \pi_{i,k}^{[0]} \left\| X_i^{(v')} - \mu_{k,[0]}^{(v')} \right\|^2\right)^{\frac{1}{1-\gamma}}}.$$
- **Step** $t \geq 1$**:**
  - *Update the clustering matrix* $\Pi^{[t]}$ *and the centers* $\mu^{[t]}$:
  Determine the two clusters C_{k_1} and C_{k_2} such that
  $$(k_1, k_2) = \arg\min_{k \neq k'} \frac{n_k n_{k'}}{n_k + n_{k'}} \sum_{v=1}^{V} \left(\alpha_v^{[t-1]}\right)^{\gamma} \left\| \mu_{k,[t-1]}^{(v)} - \mu_{k',[t-1]}^{(v)} \right\|^2,$$
  and update $\Pi^{[t]}$ and $\mu_{[t]}$.
  - *Update the weight vector* $\underline{\alpha}^{[t]} = \left(\alpha_1^{[t]}, \ldots, \alpha_V^{[t]}\right)$:
  for all $v = 1, \ldots, V$,
  $$\alpha_v^{[t]} = \frac{\left(\sum_{k=1}^{K_{\text{init}}-t} \sum_{i=1}^{n} \pi_{i,k}^{[t]} \left\| X_i^{(v)} - \mu_{k,[t]}^{(v)} \right\|^2\right)^{\frac{1}{1-\gamma}}}{\sum_{v'=1}^{V} \left(\sum_{k=1}^{K_{\text{init}}-t} \sum_{i=1}^{n} \pi_{i,k}^{[t]} \left\| X_i^{(v')} - \mu_{k,[t]}^{(v')} \right\|^2\right)^{\frac{1}{1-\gamma}}}.$$

Algorithm 2: Description of the fuzzy multi-view aggregation algorithm.

Remark 3.2. *The hard clustering version of this algorithm consists of taking an initial hard clustering matrix (*$0-1$ *values with a single 1 for each observation); the remainder of the procedure is similar to the fuzzy version here.*

3.4. maskmeans ***R package***

The multi-view hard and fuzzy aggregation and splitting algorithms described above have been implemented in an open-source R software package called *maskmeans*, freely available at `https://github.com/andreamrau/maskmeans`. A package vignette provides a full worked example and description; the primary functions of this package are as follows:

- `maskmeans`, which itself calls either `mv_aggregation` or `mv_splitting`. Note that this algorithm allows either fixed multi-view weights across clusters (aggregation and splitting) or cluster-weighted multi-view weights (splitting) via `perCluster_mv_weights = FALSE` or `TRUE`, respectively;
- `mv_simulate` to simulate data as described in the following section.

- Two main plotting functions: `mv_plot`, which provides an overview visualization of multi-view data (see Figure 6 for an example), and `maskmeans_plot`, which provides several visualization of the results of the `maskmeans` function. The plotting functions notably make use of the *ggplot2* (Wickham, 2016) and *clustree* (Zappia and Oshlack, 2018) visualization packages. Several examples of output from the `maskmeans_plot` function may be seen in Figures 1-4.

4. Simulation study

In our simulation study, we wish to evaluate our proposed multi-view aggregation and splitting algorithm to the alternative naive approaches of either concatenating all views into a united view, thus effectively ignoring the multi-view structure of the data, or using only a single view, thus ignoring the additional data views. To this end, we define the following general framework to generate data arising from six views, $Z = (Z^{(1)}, \dots, Z^{(6)})$ where $Z^{(v)} \in \mathcal{M}_{n,d_v}(\mathbb{R})$. Specifically, to start the set of observations $\{1, \dots, Kn\}$ is partitioned into $K = 2\tilde{K}+1$ equally sized clusters $(\mathcal{C}_k)_k$ of n observations. The first and second views are then simulated as follows:

$$\forall i \in \mathcal{C}_k,\ Z_i^{(1)} \sim \mathcal{N}\left(\beta\left(\cos(\theta_k), \sin(\theta_k)\right)' \mathbb{1}_{k\in\{1,\dots,2\tilde{K}\}}, I_2\right) \text{ and}$$
$$\forall i \in \mathcal{C}_k,\ Z_i^{(2)} \sim \mathcal{N}\left(\beta\left(\cos(\tilde{\theta}_k), \sin(\tilde{\theta}_k)\right)' \mathbb{1}_{k\in\{1,\dots,2\tilde{K}\}}, \sigma^2 I_2\right),$$

where σ^2 represents the per-cluster variance, $\theta_k = \pi k/\tilde{K}$, $\tilde{\theta}_k = (\theta_{2p}+\theta_{2p-1})/2$ if $k = 2p$ or $k = 2p-1, p \in \{1, \dots, \tilde{K}\}$, and β is a multiplicative factor that controls the spread of clusters around the origin. The first view is thus simulated to have $2\tilde{K}$ equally spaced clusters in a circular pattern, with an additional cluster centered at the origin; in the second view, pairs of adjacent clusters from the first view have been merged, yielding $\tilde{K}$ clusters similarly evenly spaced in a circle in addition to the central cluster at the origin. For the third view, a random permutation τ of $\{1, \dots, Kn\}$ is used to permute the clustering; this intentionally creates a noisy view with no clustering coherence with respect to the other views. The fourth and fifth views are unidimensional ($d_v = 1$), where

$$Z_i^{(4)} \sim \mathcal{N}\left(\text{sign}(Z_{i1}^{(1)})\ \mu, \sigma^2\right) \text{ and } Z_i^{(5)} \sim \mathcal{N}\left(\text{sign}(Z_{i2}^{(1)})\ \mu, \sigma^2\right).$$

Finally, the clustering structure in the sixth view aggregates the clusters of the first view into four:

$$Z_i^{(6)} \sim \mathcal{N}\left(1.5\ \beta\left(\cos(\theta_k), \sin(\theta_k)\right)' \mathbb{1}_{k\in\tilde{\tau}}, \sigma^2 I_2\right),$$

where $\tilde{\tau}$ corresponds to a random selection of three elements among $\{1, \dots, 2\tilde{K}\}$. Note that by construction, there are $2\tilde{K}+1$ clusters in view 1, $\tilde{K}+1$ clusters in view 2, 3 clusters in views 4 and 5, and 4 clusters in view 6, and the spread of

clusters around the center is increased in this view by 50% with respect to views 1 and 2. As such, the simulation depends on a set of parameters including the number of observations n, the number of clusters $K = 2\tilde{K} + 1$, β (the spread of values around the origin for the first, second, and sixth views), and σ^2 (the variance of the noise added to views 2, 4, 5, and 6). In the following, we set $n = 100$, $\beta = 4$, $K = 7$, and $\sigma = 1.5$, and simulated data were generated using the `mv_simulate` function in *maskmeans*; a graphical representation of a representative simulated data set is included in Figure 6. Simulations were repeated 100 independent times.

Initial hard and fuzzy cluster partitions were respectively obtained using the K-means or the fuzzy K-means algorithms (Bezdek, 1981), where the latter was performed using the *fclust* R package (Ferraro and Giordani, 2015) with default parameters. For the splitting algorithms, the initial clustering was obtained using data from view 2, with $K_{init} = 4$; for the aggregation algorithms, the initial clustering was obtained using data from view 1, with $K_{init} = 20$. Subsequently, all aggregation and fuzzy algorithms were iterated until a total of $K = 7$ final clusters were obtained. In all multi-view splitting and aggregation algorithms, γ was set to 2. The "true" data partition used for benchmarking was that corresponding to the first data view, as partitions in all other views (with the exception of the third) were based on aggregations of the first view. All approaches were evaluated using the misclassification error rate and the adjusted Rand index (Hubert and Arabie, 1985), a corrected-for-chance measure of similarity between two data clusterings, where values close to 1 indicate close agreement.

Algorithm	Cluster type	Strategy	ARI	Misclassification
Aggregation	Hard	Multi-view	**0.770** (0.038)	**0.107** (0.020)
		Concatenated	0.754 (0.049)	0.118 (0.031)
		Single-view	0.763 (0.043)	0.111 (0.023)
	Fuzzy	Multi-view	**0.804** (0.028)	**0.089** (0.014)
		Concatenated	0.798 (0.034)	0.092 (0.017)
		Single-view	0.801 (0.030)	0.091 (0.015)
Splitting	Hard	Cluster-weighted multi-view	0.628 (0.066)	0.206 (0.061)
		Multi-view	**0.668** (0.057)	**0.179** (0.057)
		Concatenated	0.630 (0.047)	0.198 (0.045)
		Single-view	0.638 (0.039)	0.235 (0.046)
	Fuzzy	Cluster-weighted multi-view	0.553 (0.043)	0.221 (0.035)
		Multi-view	0.579 (0.034)	0.199 (0.028)
		Concatenated	0.551 (0.047)	0.220 (0.035)
		Single-view	**0.667** (0.038)	**0.185** (0.057)

TABLE 1

Benchmarking on simulated data for the aggregation and splitting algorithms using different strategies (cluster-weighted multi-view, multi-view, concatenating all views into one, and using only the first view) and different clustering types (hard and fuzzy). Average values (standard deviation) for the ARI and misclassification rate across 100 independent simulations are indicated. Bold-face font is used to indicate the best performer in each category.

The multi-view aggregation and splitting algorithms were compared to (1) concatenated view aggregation and splitting algorithms, where data from all views were combined into a single data view; and (2) a single view aggregation and splitting algorithms, where only data from the first view was used. Results are presented in Table 1. We first note that for both hard and fuzzy aggregation algorithms, the proposed multi-view approach has the best average ARI and misclassification values, closely followed by the single-view and concatenated-view strategies (note, however, that all approaches are within a standard deviation of one another). This is perhaps unsurprising, as the concatenated-view strategy is somewhat perturbed by the inclusion of the noisy view; the single-view strategy, on the other hand, benefits from the targeted use of view 1 alone, which is the view used to evaluate the clustering partition. By making use of all available views, however, the multi-view approach is able to balance the contributions of each view, successfully down-weighting views 3, 4, and 5 to accord more importance to the more informative views (Figure 1C). The dendrogram of successive cluster aggregations, as well as the evolution of Criterion (3.3), may also be visualized for a simulated dataset using the plotting capabilities of *maskmeans* (Figure 1A-B).

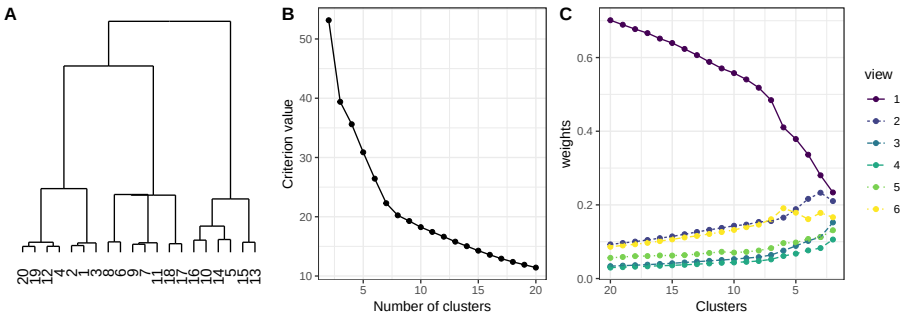

FIG 1. *Visualization of results from the multi-view aggregation algorithm (with hard clustering) applied to a single simulated data set. (A) Dendogram indicating successive cluster aggregations. (B) Plot of the value of Criterion (3.3) with respect to the total number of clusters. (C) Per-view weights at each successive step of the aggregation algorithm. Plots were produced using the* maskmeans *package.*

Regarding the splitting algorithms, we first remark a worse overall performance compared to the aggregation algorithms, particularly for fuzzy clustering; this reflects the fact that in this simulation scenario, splitting clusters appears to be more difficult than aggregating them. However, it is not of particular interest to compare the aggregation and splitting algorithms to one another, as generally in practice one strategy or the other would be more natural. For hard splitting, the multi-view strategies are more variable than the concatenated or single-view approaches (and as above, all methods are within a standard deviation of each other), but the multi-view approach has a slight advantage in uncovering the true clustering structure of view 1. However, for fuzzy splitting,

there is a very clear advantage is using only the data from view 1; this reflects the pronounced fuzziness (i.e., the overall relatively small values of maximum membership degree values) of the initial clustering used as a point of departure, and suggests that multi-view splitting approaches for fuzzy clustering are not particularly useful for very fuzzy initial clusterings.

Although the cluster-weighted multi-view approach has a lower average ARI and higher average misclassification rate than the standard multi-view approach, it does have the advantage of contributing additional information for the interpretation of cluster splits. A visualization of the cluster-weighted multi-view splitting algorithm (with hard clusters) is shown in Figure 2. The per-cluster per-view weights (Figure 2, right) represent a useful tool for identifying the views that play a determinant role in splitting clusters. We first note that views 3, 4, and 5 are attributed relatively small weights for each iteration of the algorithm; in addition, the weights of the remaining views change according to the choice of cluster that is split. To further illustrate this point, in Figure 6A a selection of the views from a simulated data set are plotted, with observations colored according to cluster membership in the initial partition (where $K_{init} = 4$; top panel) and following the initial split (where cluster 1 is split into clusters 1 and 5; bottom panel) indicated in the splitting tree in Figure 2. We note that prior to the split, the second view had the largest weight for the original cluster 1 (Figure 6B); subsequently, views 1 and 6 are up-weighted for the newly created clusters 1 and 5. As can be seen in examining the scatterplots in Figure 6A, this up-weighting of views 1 and 6 is quite logical, as the newly split clusters 1 and 5 are very clearly separated in these views; however, view 2, where the newly formed clusters largely overlap, is now down-weighted.

Based on these results, we can conclude that the multi-view aggregation and splitting procedures are able to successfully up- or down-weight views according to their informative value for clustering observations, which leads to improve clustering partitions when compared to naive single-view or concatenated-view strategies (with the exception of splitting for fuzzy clusterings). In particular, these per-view weights provide valuable information about which views contribute the most to splits or aggregations at different stages in the algorithm; although the cluster-weighted multi-view algorithm for hard clustering can slightly penalize the final cluster quality, it provides a more detailed interpretation of how each view contributes to each cluster individually.

5. Results on multi-omic breast cancer data

In this section, we apply the multi-view hard splitting algorithm with per-cluster and per-view weights described in Section 3.2 to further subdivide the five intrinisic subtypes inferred from 506 patients with breast cancer on the basis of gene expression, miRNA expression, promoter methylation, and copy number alterations in the TCGA breast cancer data.

In Figure 4, the splitting tree and corresponding per-cluster per-view weights at each split (up to $K = 10$) are provided. Strikingly, cluster splits preferentially

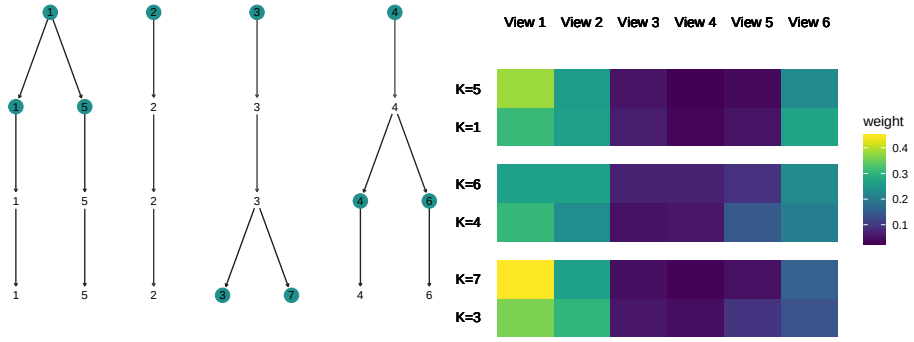

FIG 2. *Visualization of results from the cluster-weighted multi-view splitting algorithm (with hard clustering) applied to a single simulated data set. (left) Splitting tree illustrating the order of cluster splits identified by the algorithm. The initial clustering partition contained 4 clusters; in the first iteration, the first cluster was split into clusters 1 and 5, and so on. (right) Corresponding heatmap of per-cluster per-view weights at each step of the algorithm. Only clusters involved in splits are shown. Plots were produced using the* maskmeans *package.*

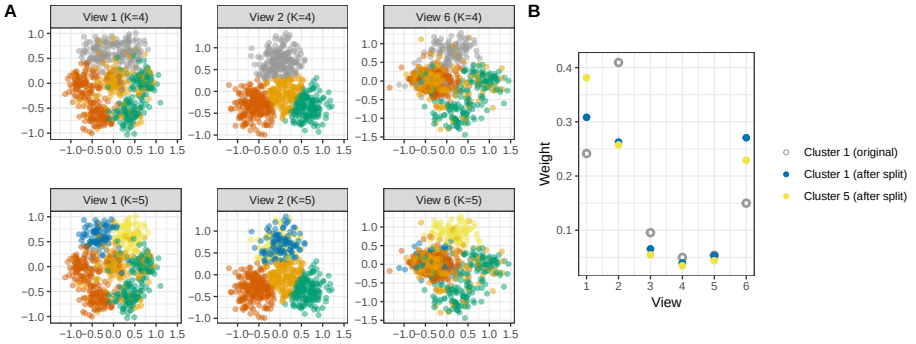

FIG 3. *Visualization of results from the cluster-weighted multi-view splitting algorithm (with hard clustering) applied to a single simulated data set. Panel A: (top) Scatterplots of simulated data for views 1 (left), 2 (middle), and 6 (right), with points colored by the initial partition into* K_{init}*=4 clusters. (bottom) Scatterplots of the same data views, with points colored by the partition after splitting cluster 1 (grey from the top panel) into two clusters (blue and yellow). Cluster colors are comparable across all graphs, and plots were produced using the maskmeans package. Panel B: Per-cluster per-view weights for cluster 1 from the original panel (grey), and for the clusters 1 and 5 (blue and yellow) after splitting.*

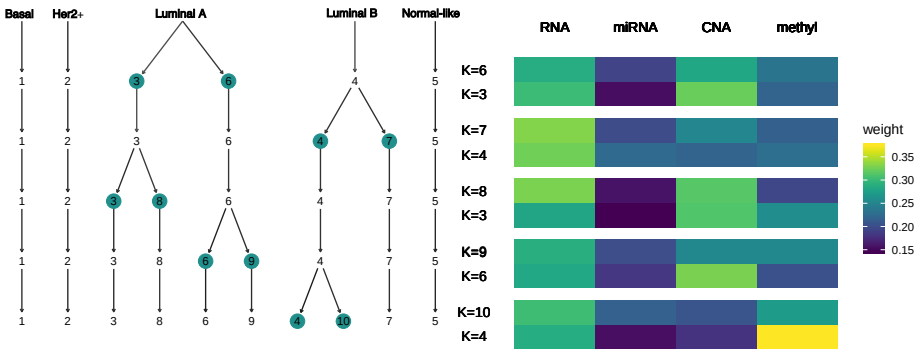

FIG 4. *Visualization of results from the cluster-weighted multi-view splitting algorithm (with hard clustering) applied to the TCGA multi-omics breast cancer data, up to $K = 10$ clusters. (left) Splitting tree illustrating the order of cluster splits identified by the algorithm. The initial clustering partition contained the five intrinsic subtypes: Basal, HER2+, Luminal A, Luminal B, and Normal-like. (right) Corresponding heatmap of per-cluster per-view weights at each step of the algorithm. Only clusters involved in splits are shown. Plots were produced using the* maskmeans *package.*

occur within the Luminal A and B subtypes, while Basal, HER2+, and Normal-like subtypes are left intact, suggesting that on a molecular level (based on the selected genes of interest), each of these groups are more homogenous than the Luminal subtypes. Basal-like breast cancer (also called triple-negative) is hormone-receptor (PR/ER) and HER2 negative and tends to be aggressive, difficult to treat, and more common among younger women and women of African descent, while HER2+ breast cancer is hormone-receptor negative but HER2 positive, grows faster than Luminal tumors, but typically responds well to treatment. Normal-like tumors, similarly to Luminal A tumors, are hormone-receptor positive and HER2 negative but typically resemble normal breast profiling and have poor outcomes. On the other hand, hormone receptor positive (Luminal A and B) tumors are the most prevalent and diverse form of breast cancer, and have been previously observed to be characterized by the most variability in survival and highest risk of late mortality (Ciriello et al., 2013); this appears to be in agreement with the fact that cluster splits occur uniquely within the Luminal tumors.

The per-cluster per-view weights in Figure 4 (right) highlight the variable contributions of each omic source to the cluster splits. For example, the first split dividing the Luminal A group in two is largely driven by gene expression and copy number alterations, while the first split of the Luminal B group is primarily due to gene expression. miRNA expression does not appear to play a major role in cluster splits, while promoter methylation only intervenes at the secondary split of the Luminal B group. By examining the multi-omic data for each of these newly identified sub-groupings (Figure 5), we can also visualize how each molecular source contributes to the cluster splits. For example (with

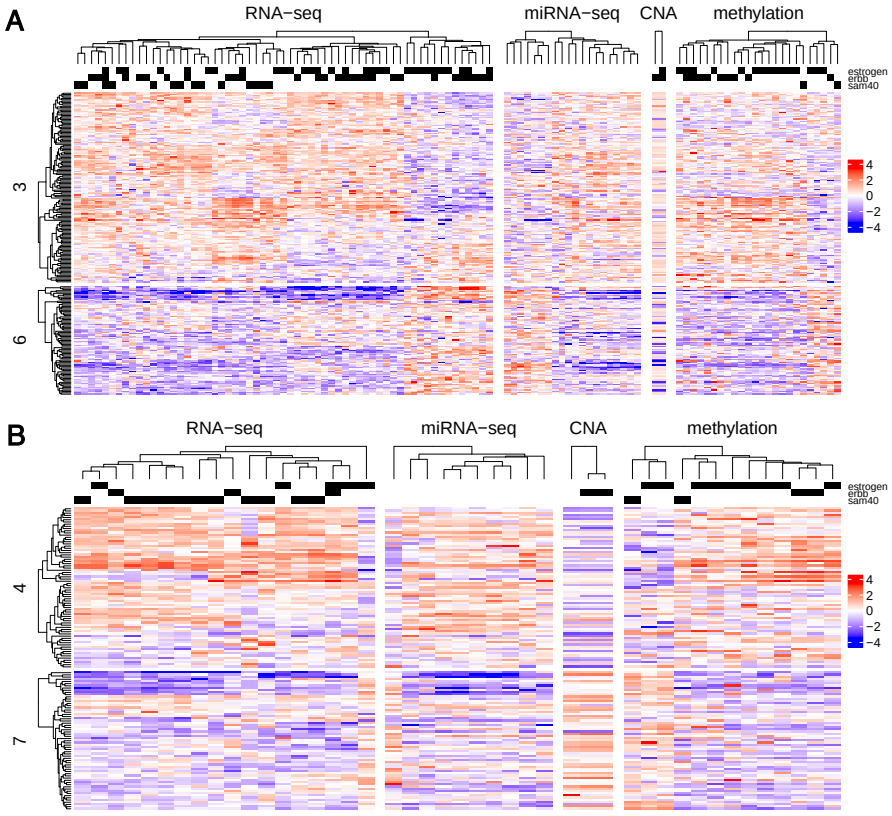

FIG 5. *Heatmaps of significantly differential variables (linear model: Bonferroni-adjusted $P < 0.01$) between pairs of subgroups identified for the Luminal A and B subtypes. (A) Z-scores for differential variables for each omic data source in the Luminal A sub-clusters 6 and 3. (B) Z-scores for differential variables for each omic data source in the Luminal B sub-clusters 4 and 7. Rows and columns are clustered with hierarchical clustering (Euclidean distance, complete linkage). For reference, genes belonging to the estrogen signaling pathway, ErbB pathway, or SAM40 list are highlighted as black annotations. Figure produced using the ComplexHeatmap package (Gu et al., 2016).*

K=7), in the first split dividing the Luminal A group in two, we note that a fairly large number of genes have striking differences in expression between clusters 3 and 6; in addition, cluster 6, which had a relatively large weight for CNAs, tends to include individuals with large copy losses in a handful of genes (Figure 5A). On the other hand, the sub-clusters of the Luminal B subtype, which were characterized by large weights on the RNA-seq view, appear to feature marked over expression of SAM40 genes in cluster 4 compared to cluster 6.

It is also of interest to identify whether the newly identified sub-clusters are

clinically meaningful; for this purpose, we consider the sub-groupings obtained for $K = 7$ clusters and analyze differences between clusters 3 and 6 (the initial split of the Luminal A subtype) and between clusters 4 and 7 (the initial split of the Luminal B subtype). We focus in particular on differences between progression-free interval survival, age at initial pathologic diagnosis, menopause status, number of lymph nodes, and pathologic tumor stage. Due to the relatively small number of deaths, no significant differences in progression-free interval (Liu et al., 2018) are detected between these two pairs of clusters (log-rank test: $P = 0.855$ for the Luminal A splits and $P = 0.165$ for the Luminal B splits, with a total of 24 out of 228 and 24 out of 136 total progression-free interval events, respectively). However, a significant difference in age at diagnosis (linear model Wald statistic: $P = 1.32\times10^{-6}$ for Luminal A; $P = 0.14$ for Luminal B) and in menopause status (χ^2 test statistic: $P = 3.65\ \times10^{-4}$ in Luminal A; $P = 0.7936$ in Luminal B) was observed between Luminal A sub-clusters 3 and 6. In addition, a significant difference in number of lymph nodes was observed for Luminal B sub-clusters 4 and 7 (Poisson GLM Wald statistic: P-value $= 0.571$ in Luminal A; P-value $= 5.41\ \times10^{-12}$ in Luminal B). No significant differences among pairs of sub-clusters were observed for pathologic tumor stage (χ^2 test statistic: $P = 0.5831$ in Luminal A; $P = 0.07632$ in Luminal B).

Taken together, these results suggest that the sub-clusters of the Luminal A subtype represent distinct groups, where cluster 6 skews towards older postmenopausal patients, while sub-clusters of the Luminal B subtype represent groups with varying severity of the disease, where individuals in cluster 7 had significantly fewer lymph nodes affected by the disease.

6. Discussion

In this work, we have presented a novel pair of algorithms to aggregate or split an existing hard or fuzzy cluster partition based on a set of multi-view data. A set of simulations demonstrated the satisfactory performance of the multi-view splitting and aggregation algorithms (with the exception of fuzzy splitting), as compared to the single- and concatenated-view strategies; in addition, we illustrated how graphical outputs from the *maskmeans* package can provide useful interpretation for the contribution provided by each view globally, or by each view per cluster, at each successive iteration. Using a set of multi-omic data (gene expression, miRNA expression, methylation, and copy number alterations) from breast cancer patients from the TCGA project, we illustrated how the cluster-weighted multi-view splitting algorithm can subdivide intrinsic cancer subtypes into more homogeneous, clinically relevant subgroups. In particular, the algorithm split the two ER+ subtypes (Luminal A and Luminal B) into groups with significant differences in age of initial diagnosis and number of affected lymph nodes, respectively.

For the cluster-weighted multi-view splitting algorithm, we observed that in cases where the initial cluster partition was in near perfect agreement with one of the data views, initial weights tend to be very large (> 0.5) for one or

more clusters in that view; this phenomenon then tends to become increasingly amplified for subsequent iterations, leading to a series of splits that are driven uniquely by that view. In such cases, if this behavior is not desired, the γ parameter can be used to moderate the multi-view influence on cluster splits at early stages of the algorithm, as larger values tend to impose a greater balance in view contributions.

In practice, the choice of the initial clustering partition to be used largely depends on the context; for example, in some cases, it may be natural to obtain the initial partition from one of the data views (this was the case for the TCGA breast cancer data presented here, as the AIMS intrinisic subtypes were inferred from the RNA-seq data), while in other cases an external dataset may be used for this purpose. Another key issue is the choice of the final number of clusters to be used following cluster aggregations or splits; currently, the multi-view aggregation and splitting algorithms allow users the flexibility to choose the ultimate number of desired clusters. One possibility to determine the "optimal" number of clusters is to examine the plot of the evolution of the criterion value (e.g., Figure 1B) and identify the so-called elbow of the curve. It is also possible that model selection approaches such as the slope heuristics (Baudry et al., 2012) could be useful for identifying the optimal number of clusters, but additional research is needed on this point.

Acknowledgements

The results shown here are in whole or part based upon data generated by the http://cancergenome.nih.gov. AR is supported by the Agreenskills fellowship program with funding from the EU's Seventh Framework Program under grant agreement FP7-609398.

References

J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–70, 2012.

J.P. Baudry, A. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combing mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2):332–353, 2010.

J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, and A. Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: Cancer Journal for Clinicians*, 2018.

X. Cai, F. Nie, and H. Huang. Multi-view k-means clustering on big data. In *Twenty-Third International Joint conference on artificial intelligence*, 2013.

G. Chao, S. Sun, and J. Bi. A survey on multi-view clustering. *arXiv preprint arXiv:1712.06246*, 2017.

X. Chen, J. Z. Xu, X. Huang, and Y. Ye. TW-k-Means: Automated two-level variable weighting clustering algorithm for multiview data. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 2013.

G. Ciriello, R. Sinha, K. A. Hoadley, A.S. Jacobsen, B. Reva, C. M. Perou, C. Sander, and N. Schultz. The molecular diversity of Luminal A breast tumors. *Breast cancer research and treatment*, 131(3):409–420, 2013.

M. B. Ferraro and P. Giordani. A toolbox for fuzzy clustering using the r programming language. *Fuzzy Sets and Systems*, 279:1–16, 2015. URL http://dx.doi.org/10.1016/j.fss.2015.05.001.

T. Fleischer, J. Klajic, M.R. Aure, R. Louhimo, A.V. Pladsen, L. Ottestad, N. Touleimat, M. Laakso, A.R. Halvorsen, G.I. Grenaker Alnæs, M.L. Riis, Å. Helland, S. Hautaniemi, P.E. Lønning, B. Naume, A.L. Børresen-Dale, J. Tost, and V.N. Kristensen. DNA methylation signature (SAM40) identifies subgroups of the Luminal A breast cancer samples with distinct survival. *Oncotarget*, 8(1):1074–1082, 2017.

M. Fratello, G. Caiazzo, F. Trojsi, A. Russo, G. Tedeschi, R. Tagliaferri, , and F. Esposito. Multi-view ensemble classification of brain connectivity images for neurodegeneration type discrimination. *Neuroinformatics*, 15(2):199–213, 2017.

Z. Gu, R. Eils, and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016.

J.S. Hamid, P. Hi, N. M. Roslin, V. Ling, C. M. T. Greenwood, and J. Beyene. Data integration in genetics and genomics: Methods and challengesr. *Human Genomics Proteomics*, page 869093, 2009.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, pages 193–218, 1985.

K. Kailing, H.-P. Kriegel, A. Pryakhin, and M. Schubert. Clustering multi-represented objects with noise. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-04)*, pages 394–403, 2004.

M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44 (Database issue):D457–D462, 2016.

V. A. Koutsonikola and A. I. Vakali. A fuzzy bi-clustering approach to correlate web users and pages. *International Journal of Knowledge and Web Intelligence*, 1(1-2):3–23, 2009.

A. Kumar and H. Daumé. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 393–400, 2011.

A. Kumar, P. Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in neural information processing systems*, pages 1413–1421, 2011.

J. Liu, T. Lichtenberg, K. A Hoadley, L. M Poisson, A. J Lazar, A. D Cherniack, A. J Kovatich, C. C Benz, D. A Levine, A. V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome

analytics. *Cell*, 173(2):400–416, 2018.

Q. Mo, S. Wang, V. E Seshan, A. B Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

E.R. Paquet and M.T. Hallett. Absolute assignment of breast cancer intrinsic molecular subtype. *Journal of the National Cancer Institute*, 107(357), 2000.

R. G Pensa, C. Robardet, and J.-F. Boulicaut. A bi-clustering framework for categorical data. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 643–650. Springer, 2005.

C.M. Perou, T. Sørlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lønning, A.L. Børresen-Dale, P.O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797): 747–52, 2000.

N. Rappoport and R. Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, gky889, 2018.

A. Rau, M. Flister, H. Rui, and P. L. Auer. Exploring drivers of gene expression in the Cancer Genome Atlas. *Bioinformatics*, bty551, 2018.

A. Serra, M. Fratello, V. Fortino, G. Raiconi, R. Tagliaferri, and D. Greco. Mvda: a multi-view genomic data integration methodology. *BMC bioinformatics*, 16(1):261, 2015.

R. Shen, A. B Olshen, and M. Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.

R. Shen, Q. Mo, N. Schultz, V. E Seshan, A. B Olshen, J. Huse, M. Ladanyi, and C. Sander. Integrative subtype discovery in glioblastoma using icluster. *PloS one*, 7(4):e35236, 2012.

E. Taskesen, S. M. H. Huisman, A. Mahfouz, J. H. Krijthe, J. de Ridder, A. van de Stolpe, E. van den Akker, W. Verheagh, and M. J. T. Reinders. Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Scientific Reports*, 6(24949), 2016.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.

The Cancer Genome Atlas Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K.R. Mills Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and Stuart J.M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013.

B. Wang, A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, 2014.

Y. Wang and L. Chen. Multi-view fuzzy clustering with minimax optimization for effective clustering of data from multiple sources. *Expert Systems with Applications*, 72:457–466, 2017.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL http://ggplot2.org.

C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

Z. Yang and G. Michailidis. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1–8, 2016.

L. Zappia and A. Oshlack. Clustering trees: a visualisation for evaluating clusterings at multiple resolutions. *Gigascience*, 7(7), 2018.

Appendix A: Supplementary Figures

In this section, we present some supplementary figures to complement those found in the main manuscript. Figure 6 can be reproduced by running the following code from the *maskmeans* R package:

```
library(maskmeans)
set.seed(12345)
sim <- mv_simulate(beta=4, n=100, K=7, sigma=1.5)
mv_plot(sim$data, labels = sim$labels[,1])
```

Appendix B: A useful lemma

Lemma B.1. *The solution of the following optimization problem*

$$\arg\min_{\underline{\beta}} \sum_{v=1}^{V} (\beta_v)^{\gamma} A_v,$$

where $\underline{\beta} = (\beta_1, \ldots, \beta_V) \in [0,1]^V$, $\sum_{v=1}^{V} \beta_v = 1$, $\gamma \in \mathbb{N}^\star$ *and* $\forall v,\ A_v \geq 0$, *is given by*

- *If* $\gamma > 1$ *: for all* v,
$$\beta_v = \frac{A_v^{\frac{1}{1-\gamma}}}{\sum_{v'=1}^{V} A_{v'}^{\frac{1}{1-\gamma}}}.$$
- *If* $\gamma = 1$ *:* $\beta_v = \mathbb{1}_{v=v_0}$ *where* $v_0 = \arg\min_{v=1,\ldots,V} A_v$.

Appendix C: Proof of Proposition 3.1

Proof. Let $\hat{k} \in \{1, \ldots, K\}$ and let $\mathcal{C}_{k_1}, C_{k_2}$ be the clusters obtained by splitting $\mathcal{C}_{\hat{k}}$. First, note that for any weight matrix $\underline{\alpha} = (\alpha_{k,v})$, we have

$$\begin{aligned} \text{Split}\left(\Pi_K, \underline{\alpha}, \mu\right) &= \sum_{i=1}^{n} \sum_{k \neq \hat{k}} \sum_{v=1}^{V} \left(\alpha_{k,v}\right)^{\gamma} \left(\pi_{i,k}\right)^{\delta} \left\| X_i^{(v)} - \mu_k^{(v)} \right\|^2 \\ &\quad + \sum_{i=1}^{n} \sum_{v=1}^{V} \left(\alpha_{\hat{k},v}\right)^{\gamma} \left(\pi_{i,\hat{k}}\right)^{\delta} \left\| X_i^{(v)} - \mu_{\hat{k}}^{(v)} \right\|^2 \end{aligned}$$

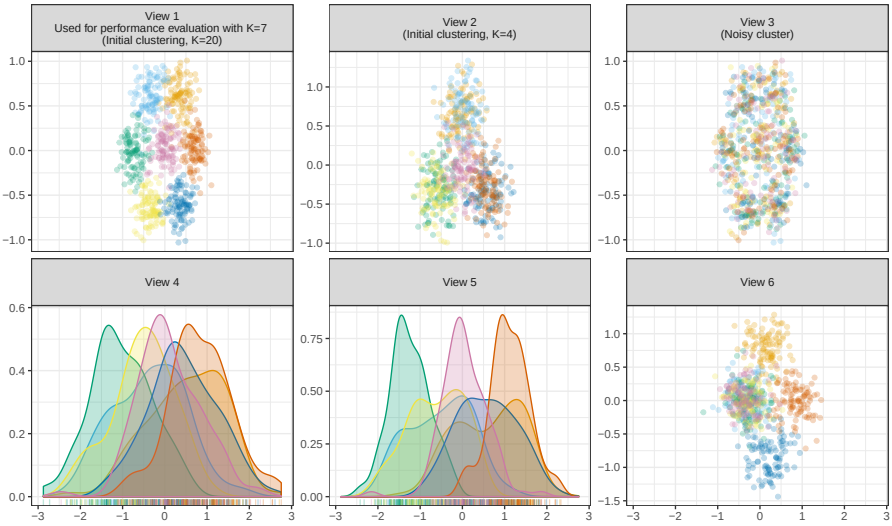

FIG 6. *Visualization of a representative set of simulated multi-view data, following the framework described in Section 4 of the main manuscript. Here* $K = 7$, $\beta = 4$, $\sigma = 1.5$, *and* $n = 100$. *Views 1, 2, 3, and 6 are bivariate, while views 4 and 5 are both univariate. Data are colored according to the true cluster labels of the first data view. The true labels of view 1 are used for performance evaluation (misclassification rate, ARI) in the simulation study. K-means and fuzzy K-means algorithms are used to create initial hard and fuzzy cluster partitions. For splitting algorithms,* $K_{init} = 4$ *and data from view 2 are used; for aggregation algorithms,* $K_{init} = 20$ *and data from view 1 are used.*

and, if we denote by $\underline{\alpha}'$ the $(K+1) \times V$ weight matrix where for all $k \notin \{k_1, k_2\}$, and for all $v = 1, \ldots, V$, $\alpha'_{k,v} = \alpha_{k,v}$ and $\alpha'_{k_1,v} = \alpha'_{k_2,v} = \alpha_{\hat{k},v}$,

$$\text{Split}\left(\tilde{\Pi}_{K+1}, \underline{\alpha}', \tilde{\mu}\right) = \sum_{i=1}^{n} \sum_{k \notin \{k_1,k_2\}} \sum_{v=1}^{V} (\alpha_{k,v})^{\gamma} (\pi_{i,k})^{\delta} \left\| X_i^{(v)} - \mu_k^{(v)} \right\|^2 + \sum_{i=1}^{n} \sum_{k \in \{k_1,k_2\}} \sum_{v=1}^{V} \left(\alpha_{\hat{k},v}\right)^{\gamma} (\tilde{\pi}_{i,k})^{\delta} \left\| X_i^{(v)} - \tilde{\mu}_k^{(v)} \right\|^2.$$

Then,

$$\text{Split}\left(\Pi_K, \underline{\alpha}, \mu\right) - \text{Split}\left(\tilde{\Pi}_{K+1}, \underline{\alpha}', \tilde{\mu}\right) = \sum_{i=1}^{n} \sum_{v=1}^{V} \left(\alpha_{\hat{k},v}\right)^{\gamma} \left(\pi_{i,\hat{k}}\right)^{\delta} \left\| X_i^{(v)} - \mu_{\hat{k}}^{(v)} \right\|^2 - \sum_{i=1}^{n} \sum_{k \in \{k_1,k_2\}} \sum_{v=1}^{V} \left(\alpha_{\hat{k},v}\right)^{\gamma} (\tilde{\pi}_{i,k})^{\delta} \left\| X_i^{(v)} - \tilde{\mu}_k^{(v)} \right\|^2.$$

Taking $k_1 = \hat{k}$ and for all i, $\pi'_{i,k_1} = \pi_{i,\hat{k}}$, under the constraint $\pi'_{i,k_1} + \pi'_{i,k_2} = \pi_{i,\hat{k}}$, it comes $\pi'_{i,k_2} = 0$ and by definition of $\tilde{C}_{k_1}, \tilde{C}_{k_2}$,

$$\sum_{i=1}^{n}\sum_{v=1}^{V}\left(\alpha_{\hat{k},v}\right)^{\gamma}\left(\pi_{i,\hat{k}}\right)^{\delta}\left\|X_i^{(v)} - \mu_{\hat{k}}^{(v)}\right\|^2 = \sum_{i=1}^{n}\sum_{k\in\{k_1,k_2\}}\sum_{v=1}^{V}\left(\alpha_{\hat{k},v}\right)^{\gamma}\left(\pi'_{i,k}\right)^{\delta}\left\|X_i^{(v)} - \left(\mu_k^{(v)}\right)'\right\|^2$$
$$\geq \sum_{i=1}^{n}\sum_{k\in\{k_1,k_2\}}\sum_{v=1}^{V}\left(\alpha_{\hat{k},v}\right)^{\gamma}\left(\tilde{\pi}_{i,k}\right)^{\delta}\left\|X_i^{(v)} - \tilde{\mu}_k^{(v)}\right\|^2.$$

Furthermore,

$$\text{Split}\left(\tilde{\Pi}_{k+1}, \underline{\alpha}', \tilde{\mu}\right) - \text{Split}\left(\tilde{\Pi}_{k+1}, \tilde{\underline{\alpha}}, \tilde{\mu}\right)$$
$$= \sum_{l\in\{k_1,k_2\}}\left(\sum_{i=1}^{n}\sum_{v=1}^{V}\left(\alpha_{\hat{k},v}\right)^{\gamma}\left(\pi_{i,l}\right)^{\delta}\left\|X_i^{(v)} - \tilde{\mu}_l^{(v)}\right\|^2 - \sum_{i=1}^{n}\sum_{v=1}^{V}\left(\tilde{\alpha}_{l,v}\right)^{\gamma}\left(\pi_{i,l}\right)^{\delta}\left\|X_i^{(v)} - \tilde{\mu}_l^{(v)}\right\|^2\right),$$

and by definition of $\tilde{\alpha}_{k_1}$ and $\tilde{\alpha}_{k_2}$ and thanks to Lemma B.1, this last term is non-negative, which leads to

$$\text{Split}\left(\Pi_K, \underline{\alpha}, \mu\right) - \text{Split}\left(\tilde{\Pi}_{K+1}, \tilde{\underline{\alpha}}, \tilde{\mu}\right) \geq 0.$$

□

Appendix D: Proof of Proposition 3.2

Proof. In Proposition 3.2, we consider the following criterion

$$\text{Agg}\left(\Pi_K, \underline{\alpha}, \mu\right) = \sum_{v=1}^{V}\sum_{k=1}^{K}\sum_{i=1}^{n}(\alpha_v)^{\gamma}\pi_{i,k}\|X_i^{(v)} - \mu_k^{(v)}\|^2$$

with $\mu_k = \frac{1}{n_k}\sum_{i=1}^{n}\pi_{i,k}X_i$ and $n_k = \sum_{i=1}^{n}\pi_{i,k}$.

When we aggregate the two clusters $\mathcal{C}_k$ et $\mathcal{C}_{k'}$ and that $\tilde{\pi}_{i,k\cup k'} = \pi_{i,k} + \pi_{i,k'}$, we obtain that

$$\begin{aligned}
&\text{Agg}(\Pi_K, \underline{\alpha}, \mu) - \text{Agg}(\tilde{\Pi}_{K-1}, \underline{\alpha}, \tilde{\mu}) \\
= &\sum_{v=1}^{V}\sum_{i=1}^{n}(\alpha_v)^{\gamma}\left(\pi_{i,k}\|X_i^{(v)} - \mu_k^{(v)}\|^2 + \pi_{i,k'}\|X_i^{(v)} - \mu_{k'}^{(v)}\|^2\right) \\
&-\sum_{v=1}^{V}\sum_{i=1}^{n}(\alpha_v)^{\gamma}\tilde{\pi}_{i,k\cup k'}\|X_i^{(v)} - \tilde{\mu}_{k\cup k'}^{(v)}\|^2 \\
= &\sum_{v=1}^{V}\sum_{i=1}^{n}(\alpha_v)^{\gamma}\Delta_{i,k,k'}
\end{aligned}$$

with

$$\Delta_{i,k,k'} = \left(\pi_{i,k}\|X_i^{(v)} - \mu_k^{(v)}\|^2 + \pi_{i,k'}\|X_i^{(v)} - \mu_{k'}^{(v)}\|^2\right) - \tilde{\pi}_{i,k\cup k'}\|X_i^{(v)} - \tilde{\mu}_{k\cup k'}^{(v)}\|^2$$

and $\tilde{\mu}_{k\cup k'} = a_k\mu_k + a_{k'}\mu_{k'}$ with $a_k = n_k/(n_k+n_{k'})$ and $a_{k'} = n_{k'}/(n_k+n_{k'})$. Thus,

$$\begin{aligned}
\Delta_{i,k,k'} &= \pi_{i,k}\|X_i^{(v)} - \mu_k^{(v)}\|^2 + \pi_{i,k'}\|X_i^{(v)} - \mu_{k'}^{(v)}\|^2 \\
&\quad - (\pi_{i,k}+\pi_{i,k'})\left\|a_k(X_i^{(v)} - \mu_k^{(v)}) + a_{k'}(X_i^{(v)} - \mu_{k'}^{(v)})\right\|^2 \\
&= \left[\pi_{i,k} - (\pi_{i,k}+\pi_{i,k'})a_k^2\right]\|X_i^{(v)} - \mu_k^{(v)}\|^2 + \left[\pi_{i,k'} - (\pi_{i,k}+\pi_{i,k'})a_{k'}^2\right]\|X_i^{(v)} - \mu_{k'}^{(v)}\|^2 \\
&\quad -2(\pi_{i,k}+\pi_{i,k'})a_k a_{k'} < X_i^{(v)} - \mu_k^{(v)}, X_i^{(v)} - \mu_{k'}^{(v)} > \\
&= \left[\pi_{i,k'} - (\pi_{i,k}+\pi_{i,k'})a_{k'}^2\right]\|\mu_k^{(v)} - \mu_{k'}^{(v)}\|^2 + 0 \times \|X_i^{(v)} - \mu_k^{(v)}\|^2 \\
&\quad +2 < X_i^{(v)} - \mu_k^{(v)}, \mu_k^{(v)} - \mu_{k'}^{(v)} > \left[\pi_{i,k'} - (\pi_{i,k}+\pi_{i,k'})a_{k'}^2 - (\pi_{i,k}+\pi_{i,k'})a_k a_{k'}\right].
\end{aligned}$$

Since

$$\begin{aligned}
&\sum_{i=1}^{n} 2 < X_i^{(v)} - \mu_k^{(v)}, \mu_k^{(v)} - \mu_{k'}^{(v)} > \left[\pi_{i,k'} - (\pi_{i,k}+\pi_{i,k'})a_{k'}^2 - (\pi_{i,k}+\pi_{i,k'})a_k a_{k'}\right] \\
= \;& -2n_{k'}(1 - a_{k'}^2 - a_k a_{k'})\|\mu_k^{(v)} - \mu_{k'}^{(v)}\|^2,
\end{aligned}$$

and

$$\begin{aligned}
\sum_{i=1}^{n}\left(\pi_{i,k'} - (\pi_{i,k}+\pi_{i,k'})\,a_k^2\right)\left\|\mu_k^{(v)} - \mu_{k'}^{(v)}\right\|^2 &= \left(n_{k'} - (n_k+n_{k'})\,a_{k'}^2\right)\left\|\mu_k^{(v)} - \mu_{k'}^{(v)}\right\|^2 \\
&= n_{k'}\left(1 - a_k a_{k'} - a_{k'}^2\right)\left\|\mu_k^{(v)} - \mu_{k'}^{(v)}\right\|^2,
\end{aligned}$$

it leads to

$$\begin{aligned}
\sum_{i=1}^{n}\Delta_{i,k,k'} &= -n_{k'}(1 - a_{k'}^2 - a_k a_{k'})\|\mu_k^{(v)} - \mu_{k'}^{(v)}\|^2 \\
&= -\frac{n_k n_{k'}}{n_k+n_{k'}}\|\mu_k^{(v)} - \mu_{k'}^{(v)}\|^2.
\end{aligned}$$

Finally,

$$\mathrm{Agg}(\Pi_K, \underline{\alpha}, \mu) - \mathrm{Agg}(\tilde{\Pi}_{K-1}, \underline{\alpha}, \tilde{\mu}) = -\frac{n_k n_{k'}}{n_k+n_{k'}}\sum_{v=1}^{V}(\alpha_v)^\gamma\|\mu_k^{(v)} - \mu_{k'}^{(v)}\|^2 < 0.$$

□

Appendix E: How to split a cluster

In this section, the aim is, starting from an initial fuzzy clustering matrix Π_K and given a weight matrix $\underline{\alpha}$ and $\hat{k} \geq 1$, to split cluster $C_{\hat{k}}$ into two clusters

$\tilde{C}_{k_1}, \tilde{C}_{k_2}$ which minimize

$$\sum_{l=k_1,k_2} \sum_{v=1}^{V} \sum_{i=1}^{n} \left(\alpha_{\hat{k},v}\right)^{\gamma} \left(\tilde{\pi}_{i,l}\right)^{\delta} \left\| X_i^{(v)} - \tilde{\mu}_l^{(v)} \right\|^2,$$

under the constraints

$$\tilde{\pi}_{i,k_1} + \tilde{\pi}_{i,k_2} = \pi_{i,\hat{k}}, \qquad \forall i = 1, \ldots, n.$$

In what follows, for the sake of simplicity, we will take $k_1 = 1$ and $k_2 = 2$, which leads to the following constrained weighted fuzzy K-means algorithm:

- **Step** $t = 0$**:**
  - *Initialization of the centers:* choose randomly two index i_1, i_2, $1 \leq i_1, i_2 \leq n$ such that $i_1 \neq i_2$, and

$$\tilde{\mu}_{1,[0]} = X_{i_1} \qquad\qquad \tilde{\mu}_{2,[0]} = X_{i_2}$$

  - *Initialization of the fuzzy clustering matrix* $\tilde{\Pi}^{[0]} = \left(\tilde{\pi}_{i,l}^{[0]}\right)$*:* for all $i = 1, \ldots, n$ and $l = 1, 2$,

$$\tilde{\pi}_{i,l}^{[0]} = \pi_{i,\hat{k}} \; \frac{\left(\sum_{v=1}^{V} \left(\alpha_{\hat{k},v}\right)^{\gamma} \left\| X_i^{(v)} - \tilde{\mu}_{l,[0]} \right\|\right)^{\frac{1}{1-\delta}}}{\sum_{l'=1,2} \left(\sum_{v=1}^{V} \left(\alpha_{\hat{k},v}\right)^{\gamma} \left\| X_i^{(v)} - \tilde{\mu}_{l',[0]} \right\|\right)^{\frac{1}{1-\delta}}}$$

- **Step** $t \geq 1$**:**
  - *Updating the centers:* for all $l = 1, 2$,

$$\tilde{\mu}_{l,[t]} = \frac{1}{\sum_{i=1}^{n} \left(\tilde{\pi}_{i,l}^{[t-1]}\right)^{\delta}} \sum_{i=1}^{n} \left(\tilde{\pi}_{i,l}^{[t-1]}\right)^{\delta} X_i.$$

  - *Updating the fuzzy clustering matrix* $\tilde{\Pi}^{[t]} = \left(\tilde{\pi}_{i,l}^{[t]}\right)$*:* for all $i = 1, \ldots, n$ and $l = 1, 2$,

$$\tilde{\pi}_{i,l}^{[t]} = \pi_{i,\hat{k}} \; \frac{\left(\sum_{v=1}^{V} \left(\alpha_{\hat{k},v}\right)^{\gamma} \left\| X_i^{(v)} - \tilde{\mu}_{l,[t]} \right\|\right)^{\frac{1}{1-\delta}}}{\sum_{l'=1,2} \left(\sum_{v=1}^{V} \left(\alpha_{\hat{k},v}\right)^{\gamma} \left\| X_i^{(v)} - \tilde{\mu}_{l',[t]} \right\|\right)^{\frac{1}{1-\delta}}}.$$