The Annals of Applied Statistics
2020, Vol. 14, No. 2, 752–767
https://doi.org/10.1214/19-AOAS1317
© Institute of Mathematical Statistics, 2020

MULTIVIEW CLUSTER AGGREGATION AND SPLITTING, WITH AN APPLICATION TO MULTIOMIC BREAST CANCER DATA

BY ANTOINE GODICHON-BAGGIONI[1], CATHY MAUGIS-RABUSSEAU[2] AND ANDREA RAU[3]

[1] *Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, antoine.godichon_baggioni@upmc.fr*

[2] *Institut de Mathématiques de Toulouse, UMR5219, Université de Toulouse, CNRS, INSA Toulouse, cathy.maugis@insa-toulouse.fr*

[3] *Université Paris-Saclay, INRAE, AgroParisTech, GABI, andrea.rau@inrae.fr*

Multiview 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 multiview classification algorithms have been proposed in the fields of computer science and genomics; here, we instead focus on the task of merging or splitting an existing hard or soft cluster partition based on multiview data. This article is specifically motivated by an application involving multiomic breast cancer data from The Cancer Genome Atlas, where multiple molecular profiles (gene expression, microRNA expression, methylation and copy number alterations) are used to further subdivide the five currently accepted intrinsic tumor subtypes into distinct subgroups of patients. In addition, we investigate the performance of the proposed multiview splitting and aggregation algorithms, as compared to single- and concatenated-view alternatives, in a set of simulations. The multiview splitting and aggregation algorithms developed here are implemented in the *maskmeans* R package.

1. Introduction. Multiview data, also called multiblock or multiway data in the literature, refer to distinct but related sets of features; multiview data have become widely available in a variety of biological applications, including genomics (e.g., where gene expression, copy number alterations and methylation are measured on the same individuals; Chao, Sun and Bi (2017)) and neuroinformatics (e.g., functional magnetic resonance imaging; Fratello et al. (2017)). One of the underlying assumptions in exploiting these data is that multifaceted and heterogeneous views of the same problem can be useful in identifying or refining relevant and robust structures, as they may reflect different aspects of complex structures. The integrative analysis of multiview data is, thus, a major challenge and represents a large area of research. Multiview learning falls under the broader umbrella of so-called intermediate integrative analyses 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 (see, e.g., Hamid et al. (2009), Acar and Yener (2008), Xu, Tao and Xu (2013), for a general survey).

Multiview classification algorithms have been the focus of an extensive amount of research in the field of computer science in recent years; see Yang and Wang (2018) and Chao, Sun and Bi (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. One notable example of such an approach is an integrative method called Joint and Individual Variation Explained (JIVE) (Lock et al. (2013)), for which several extensions are available (see, e.g., Yu et al. (2017), Feng et al. (2018)). Similarly, Dueck, Morris and Frey (2005) addressed the problem of multiview

Received November 2018; revised November 2019.
Key words and phrases. Clustering, multiview, cluster merging and splitting, multiomic data, TCGA.

clustering using a probabilistic sparse matrix factorization, and Gaynanova and Li (2019) formulated a linked component model that directly incorporates partially-shared structures. Existing clustering methods make use of a variety of approaches, including spectral clustering (Kumar, Rai and Daume (2011), Kumar and Daumé (2011), Liu et al. (2012)), biclustering (Koutsonikola and Vakali (2009), Pensa, Robardet and Boulicaut (2005)) and density-based clustering of multiview data (Kailing et al. (2004), Taskesen et al. (2016)). Cai, Nie and Huang (2013) proposed the multiview K-means algorithm as a robust and computationally efficient method to cluster large-scale heterogeneous multiview datasets; Chen et al. (2013) extended this idea to incorporate weights on both views and variables. Multiview 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; Yang and Michailidis (2016) used nonnegative matrix factorization to jointly decompose multiview omics data. The iCluster+ approach (Shen, Olshen and Ladanyi (2009), Mo et al. (2013), Shen et al. (2012)) uses a joint latent-variable model to cluster multiomic data, while SNF (Wang et al. (2014)) combines omic information using a network-based approach to identify patient subtypes.

To our knowledge, these multiview classification techniques focus on either de novo unsupervised clustering or supervised clustering of a multiview dataset; here, we instead focus on the task of merging or splitting an existing hard or soft cluster partition based on multiview 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.

This paper focus on the specific problem of aggregating or splitting an existing initial data partition in the multiview framework; the initial partition of data may represent a clustering of a single data view or, alternatively, can represent a preexisting grouping of individuals. This article is specifically motivated by an application involving multiomic breast cancer data, where multiple omics profiles are used to further subdivide intrinsic tumor subtypes into distinct subgroups of patients. In particular, rather than focusing on a de novo clustering of patients, we instead seek to further subdivide a preestablished grouping of individuals. The remainder of this article has been organized as follows: the multiomic breast cancer data that are the focus of our study are described in Section 2. The multiview K-means algorithm, as well as the multiview 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 multiomic breast cancer data are described in detail. Finally, a discussion and some conclusions are provided in Section 6.

2. Multiomic 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 were 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)). One well-known method for subtyping breast cancer is the PAM50 gene set which was developed on microarray and quantitative reverse transcriptase polymerase chain reaction data (Parker et al. (2009)). Similarly, 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 pancancer genomic studies, including multiomic 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 multiomic data. In this paper we focus on the multiomic 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 acquisition and preprocessing 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 one. 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)), corresponding to 61, 38, 228, 136 and 43 individuals for the Basal-like, Her2-enriched, Luminal A, Luminal B and Normal-like subtypes, respectively.

Our method aims to determine whether the use of multiview 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 subclusterings 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 (resp., 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. Multiview clustering algorithms.

To build up to our proposed multiview 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 soft 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,

$$Z = (Z^{(1)}, \dots, Z^{(v)}, \dots, Z^{(V)}),$$

where $Z^{(v)} \in \mathbb{R}^{n \times d_v}$ and $d = \sum_{v=1}^{V} d_v$. As in Cai, Nie and Huang (2013), the data here are assumed to have been mean centered and 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 soft clustering. In the latter case we have an initial matrix $\Pi^{[0]} = \Pi_{K_{\text{init}}} = (\pi_{i,k}^{(0)})$, where $\pi_{i,k}^{(0)}$ is the "probability" (weight) that the ith individual belongs to the kth cluster and $\sum_{k=1}^{K^{(0)}} \pi_{i,k}^{(0)} = 1$ for each individual i. In the hard-clustering case, $\Pi_{K_{\text{init}}}$ 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 multiview soft K-means algorithm (Wang and Chen (2017)):

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

where $\gamma > 1$, $\delta > 1$ and $\mu = (\mu_1, \dots, \mu_K)$ is the vector of cluster centers such that $\mu_k = (\mu_k^{(1)}, \dots, \mu_k^{(V)})$. The vector $\underline{\alpha} = (\alpha_1, \dots, \alpha_V)$, 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 soft 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. *Multiview splitting.*

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

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

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 $\mu = (\mu_1, \dots, \mu_K)$ 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, we set both γ and δ to be equal to 2 here.

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

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

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

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

Algorithm 1 Description of the soft multiview splitting algorithm

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

under the constraint $\pi_{i,k_1} + \pi_{i,k_2} = \pi_{i,\hat{k}}$ for each $i = 1, \dots, n$. This step provides a new soft clustering matrix $\tilde{\Pi}_{K+1}$ and the associated vector of cluster centers $\tilde{\mu}$. It is detailed in the Supplementary Material (Godichon-Baggioni, Maugis-Rabusseau and Rau (2020) Section E). Then, one can obtain the weight matrix $\underline{\tilde{\alpha}}$ associated with this split, defined for all $k = 1, \dots, K+1$ and for all $v = 1, \dots, V$,

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

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

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

The proof is given in the Supplementary Material (Godichon-Baggioni, Maugis-Rabusseau and Rau (2020) Section C).

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. *Multiview 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:

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

with $\mu_k = \sum_{i=1}^{n} \pi_{i,k} X_i / \sum_{i=1}^{n} \pi_{i,k}$. Given a soft clustering matrix $\Pi_K = (\pi_{i,k})_{i=1,\dots,n,k=1,\dots,K}$, we aggregate two clusters, $\mathcal{C}_k$ and $\mathcal{C}_{k'}$ $(k \neq k')$, into a new cluster,

Algorithm 2 Description of the soft multiview aggregation algorithm

```
• Step $t = 0$: Let $\Pi_{K_{\text{init}}} = (\pi_{i,k}^{[0]})_{i,k}$ be the initial soft clustering matrix.
  – Initialization of the centers: for all $k = 1, \dots, K_{\text{init}}$,
        $$\mu_{k,[0]} = \sum_{i=1}^{n} \pi_{i,k}^{[0]} X_i \Big/ \sum_{i=1}^{n} \pi_{i,k}^{[0]}.$$
  – Initialization of the weight vector: $\underline{\alpha}^{[0]} = (\alpha_1^{[0]}, \dots, \alpha_V^{[0]})$ where for all $v = 1, \dots, V$,
        $$\alpha_v^{[0]} = \frac{(\sum_{k=1}^{K_{\text{init}}} \sum_{i=1}^{n} \pi_{i,k}^{[0]} \|X_i^{(v)} - \mu_{k,[0]}^{(v)}\|^2)^{\frac{1}{1-\gamma}}}{\sum_{v'=1}^{V} (\sum_{k=1}^{K_{\text{init}}} \sum_{i=1}^{n} \pi_{i,k}^{[0]} \|X_i^{(v')} - \mu_{k,[0]}^{(v')}\|^2)^{\frac{1}{1-\gamma}}}.$$
• 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
        $$(k_1, k_2) = \arg\min_{k \neq k'} \frac{n_k n_{k'}}{n_k + n_{k'}} \sum_{v=1}^{V} (\alpha_v^{[t-1]})^{\gamma} \|\mu_{k,[t-1]}^{(v)} - \mu_{k',[t-1]}^{(v)}\|^2,$$
    and update $\Pi^{[t]}$ and $\mu_{[t]}$.
  – Update the weight vector $\underline{\alpha}^{[t]} = (\alpha_1^{[t]}, \dots, \alpha_V^{[t]})$:
        for all $v = 1, \dots, V$,
        $$\alpha_v^{[t]} = \frac{(\sum_{k=1}^{K_{\text{init}}-t} \sum_{i=1}^{n} \pi_{i,k}^{[t]} \|X_i^{(v)} - \mu_{k,[t]}^{(v)}\|^2)^{\frac{1}{1-\gamma}}}{\sum_{v'=1}^{V} (\sum_{k=1}^{K_{\text{init}}-t} \sum_{i=1}^{n} \pi_{i,k}^{[t]} \|X_i^{(v')} - \mu_{k,[t]}^{(v')}\|^2)^{\frac{1}{1-\gamma}}}.$$
```

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

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

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

The proof is given in the Supplementary Material (Godichon-Baggioni, Maugis-Rabusseau and Rau (2020) Section D). Then, the multiview 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.

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 soft version here.

3.4. *maskmeans R package*. The multiview hard and soft 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 multiview weights across clusters (aggregation and splitting) or cluster-weighted multiview 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 multiview data (see Supplementary Material Figure 1 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 multiview aggregation and splitting algorithm to the alternative naive approaches of either concatenating all views into a united view, thus effectively ignoring the multiview 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:

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

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

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

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

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

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

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, three clusters in

TABLE 1

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

Algorithm	Cluster type	Strategy	ARI	Misclassification
Aggregation	Hard	Multiview	**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)
	Soft	Multiview	**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 multiview	0.628 (0.066)	0.206 (0.061)
		Multiview	**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)
	Soft	Cluster-weighted multiview	0.553 (0.043)	0.221 (0.035)
		Multiview	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)

views 4 and 5 and four clusters in view 6; 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 Supplementary Material Figure 1. Simulations were repeated 100 independent times.

Initial hard and soft cluster partitions were respectively obtained using the K-means or the soft 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 $K_{\text{init}} = 4$; for the aggregation algorithms, the initial clustering was obtained using data from view 1, with $K_{\text{init}} = 20$. Subsequently, all aggregation and soft algorithms were iterated until a total of $K = 7$ final clusters were obtained. In all multiview 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.

The multiview 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 soft aggregation algorithms, the proposed multiview 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

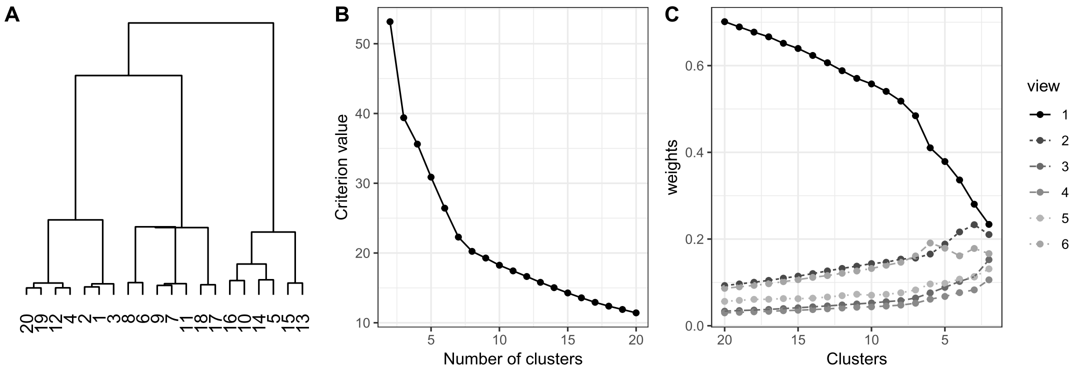

FIG. 1. *Visualization of results from the multi-view aggregation algorithm* (*with hard clustering*) *applied to a single simulated data set*. (*A*) *Dendrogram indicating successive cluster aggregations*. (*B*) *Plot of the value of Criterion* (3.3) *with respect to the total number of clusters*. (*C*)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

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 multiview 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 1(C)). 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 1(A)–(B)).

Regarding the splitting algorithms, we first remark a worse overall performance compared to the aggregation algorithms, particularly for soft 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 multiview 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 multiview approach has a slight advantage in uncovering the true clustering structure of view 1. However, for soft splitting there is a very clear advantage in 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 multiview splitting approaches for soft clustering are not particularly useful for very soft initial clusterings.

Although the cluster-weighted multiview approach has a lower average ARI and higher average misclassification rate than the standard multiview approach, it does have the advantage of contributing additional information for the interpretation of cluster splits. A visualization of the cluster-weighted multiview 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 3(A) a selection of the views from a simulated data set are plotted, with observations colored according to cluster membership in the initial partition (where $K_{\text{init}} = 4$; 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 3(B)); subsequently, views 1 and 6 are up-weighted for the newly created clusters 1 and 5. As can be

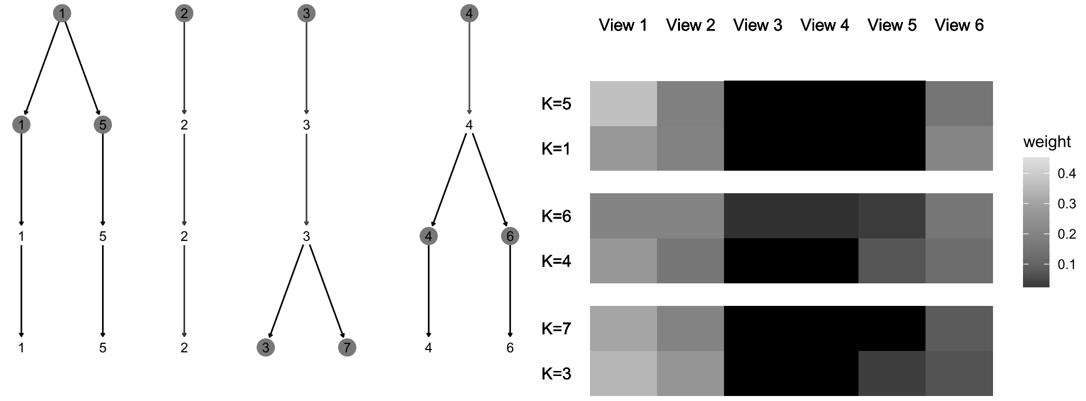

FIG. 2. *Visualization of results from the cluster-weighted multi-view splitting algorithm (with hard clustering) applied to a single simulated data set. (left) Splitting tree illustrating the order of cluster splits identified by the algorithm. The initial clustering partition contained* 4 *clusters; 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.*

seen in examining the scatter plots in Figure 3(A), 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 multiview 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 soft 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 multiview 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.

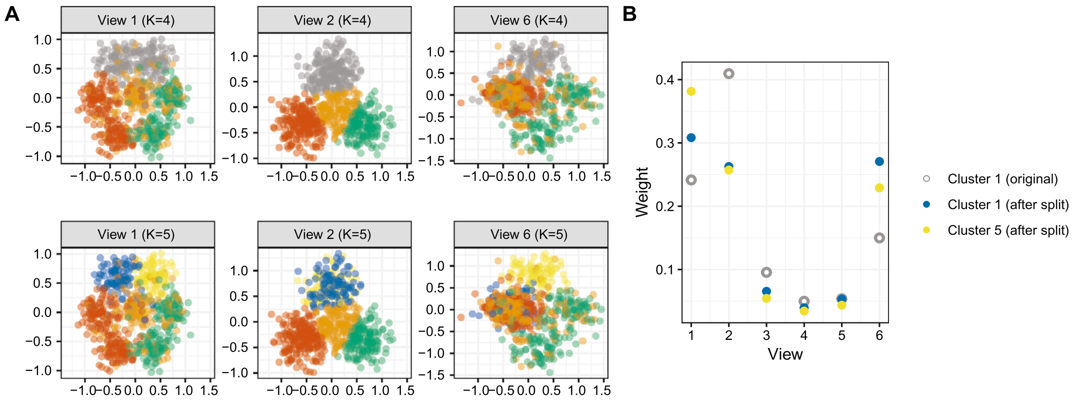

FIG. 3. *Visualization of results from the cluster-weighted multi-view splitting algorithm (with hard clustering) applied to a single simulated data set. Panel A: (top) Scatterplots of simulated data for views* 1 *(left),* 2 *(middle), and* 6 *(right), with points colored by the initial partition into* $K_{\text{init}} = 4$ *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.*

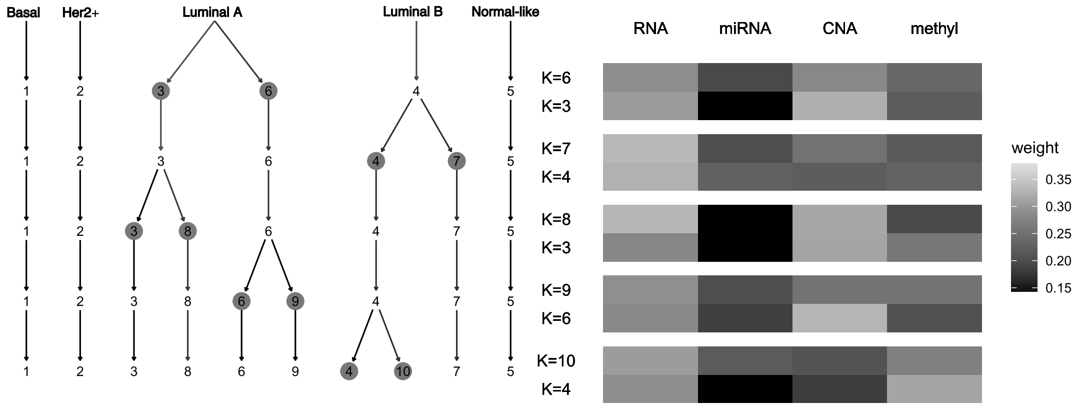

FIG. 4. *Visualization of results from the cluster-weighted multi-view splitting algorithm* (*with hard clustering*) *applied to the TCGA multiomic breast cancer data*, *up to* $K = 10$ *clusters*. (*left*) *Splitting tree illustrating the order of cluster splits identified by the algorithm. The initial 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* m*askmeans package*.

5. Results on multiomic breast cancer data. In this section we apply the multiview hard splitting algorithm with per-cluster and per-view weights described in Section 3.2 to further subdivide the five intrinsic 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 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 may be 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. We note that these preferential splits among Luminal tumors may in part be driven by a larger available sample size compared to the other subtypes. To evaluate this, we reran the *maskmeans* analysis 500 times after down-sampling the data so each subtype was equally represented ($n = 38$). After counting the number of splits per subtype for $K = 10$ clusters, we found that each subtype was split a similar number of times (2.06, 2.63, 1.91 and 2.41 average nested clusters for Basal-like, Her2-enriched, Luminal A and Luminal B, resp.) with the exception of Normal-like tumors which were never subdivided. This suggests both that unbalanced sample sizes may have some effect on the number of observed splits described above and that this is not the only determining factor driving cluster splits.

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

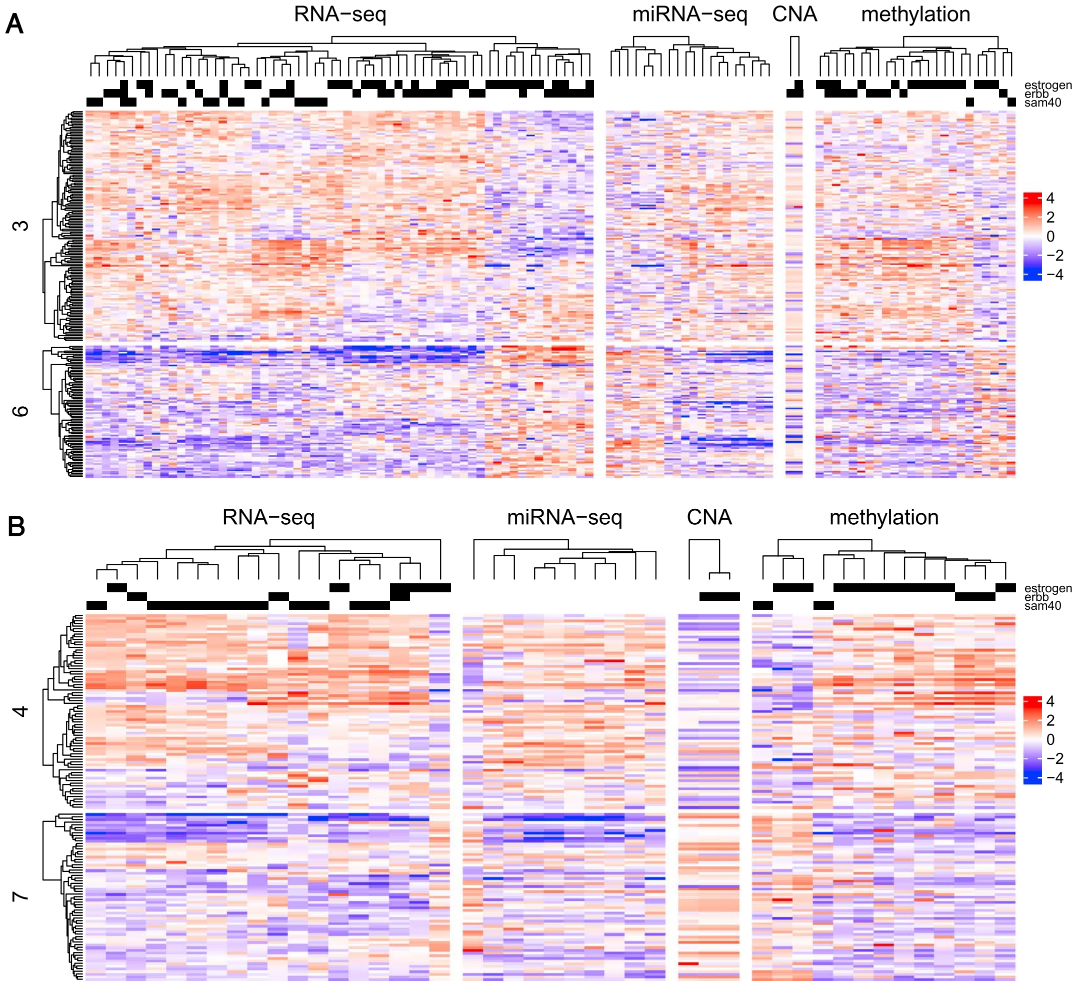

FIG. 5. *Heatmaps of significantly differential variables (linear model: Bonferroni-adjusted* $P < 0.01$*) between pairs of subgroups identified for the Luminal A and B subtypes. (A) Z-scores for differential variables for each omic data source in the Luminal A sub-clusters* 6 *and* 3. *(B) Z-scores 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 SAM*40 *list are highlighted as black annotations. Figure produced using the ComplexHeatmap package (Gu, Eils and Schlesner* (2016)).

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 multiomic data for each of these newly identified subgroupings (Figure 5), we can also visualize how each molecular source contributes to the cluster splits. For example (with $K = 7$), 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 5(A)). On the other hand, the subclusters 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 subclusters are clinically meaningful; for this purpose we consider the subgroupings 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, resp.). However, a significant difference in age at diagnosis (linear model Wald statistic: $P = 1.32 \times 10^{-6}$ for Luminal A; $P = 0.14$ for Luminal B) and in menopause status (χ^2 test statistic: $P = 3.65 \times 10^{-4}$ in Luminal A; $P = 0.7936$ in Luminal B) was observed between Luminal A subclusters 3 and 6. In addition, a significant difference in number of lymph nodes was observed for Luminal B subclusters 4 and 7 (Poisson GLM Wald statistic: P-value $= 0.571$ in Luminal A; P-value $= 5.41 \times 10^{-12}$ in Luminal B). No significant differences among pairs of subclusters 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 subclusters of the Luminal A subtype represent distinct groups, where cluster 6 skews toward older postmenopausal patients, while subclusters 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. We have presented a novel pair of algorithms to aggregate or split an existing hard or soft cluster partition based on a set of multiview data. A set of simulations demonstrated the satisfactory performance of the multiview splitting and aggregation algorithms (with the exception of soft 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 multiomic data (gene expression, miRNA expression, methylation and copy number alterations) from breast cancer patients from the TCGA project, we illustrated how the cluster-weighted multiview 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. In future work it would be of great interest to apply the multiview splitting and aggregation algorithms to other TCGA tumor types beyond breast cancer, particularly those including survival data with adequate follow-up time and sufficient sample sizes.

For the cluster-weighted multiview 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 multiview 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 multiview 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 1(B)) and identify the so-called elbow of the curve. It is also possible that model selection approaches, such as the slope heuristics (Baudry, Maugis and Michel (2012)), could be useful for identifying the optimal number of clusters, but additional research is needed on this point. Finally, an important area of future work is the extension of the methods implemented in *maskmeans* to data types beyond continuous measures, particularly binary data; this will be of great interest for genomics applications, as it would facilitate the additional use of data on somatic mutations which play a large role in cancer development and disease progression.

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

SUPPLEMENTARY MATERIAL

Multiview cluster aggregation and splitting, with an application to multiomic breast cancer data: Supplementary file (DOI: 10.1214/19-AOAS1317SUPP; .pdf). In this Supplementary Material, some additional figures are given as well as proofs of Propositions 3.1 and 3.2.

REFERENCES

ACAR, E. and YENER, B. (2008). Unsupervised multiway data analysis: A literature survey. *IEEE Trans. Knowl. Data Eng.* **21** 6–20.

BAUDRY, J.-P., MAUGIS, C. and MICHEL, B. (2012). Slope heuristics: Overview and implementation. *Stat. Comput.* **22** 455–470. MR2865029 https://doi.org/10.1007/s11222-011-9236-1

BAUDRY, J.-P., RAFTERY, A. E., CELEUX, G., LO, K. and GOTTARDO, R. (2010). Combining mixture components for clustering. *J. Comput. Graph. Statist.* **19** 332–353. MR2758307 https://doi.org/10.1198/jcgs.2010.08111

BEZDEK, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York. MR0631231

BRAY, F., FERLAY, J., SOERJOMATARAM, I., SIEGEL, R. L., TORRE, L. A. and JEMAL, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*

CAI, X., NIE, F. and HUANG, H. (2013). Multi-view k-means clustering on big data. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

CHAO, G., SUN, S. and BI, J. (2017). A survey on multi-view clustering. Preprint. Available at arXiv:1712.06246.

CHEN, X., XU, J. Z., HUANG, X. and YE, Y. (2013). TW-k-means: Automated two-level variable weighting clustering algorithm for multiview data. *IEEE Trans. Knowl. Data Eng.* **25**.

CIRIELLO, G., SINHA, R., HOADLEY, K. A., JACOBSEN, A. S., REVA, B., PEROU, C. M., SANDER, C. and SCHULTZ, N. (2013). The molecular diversity of Luminal A breast tumors. *Breast Cancer Research and Treatment* **131** 409–420.

DUECK, D., MORRIS, Q. D. and FREY, B. J. (2005). Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics* **21** i144–i151.

FENG, Q., JIANG, M., HANNIG, J. and MARRON, J. S. (2018). Angle-based joint and individual variation explained. *J. Multivariate Anal.* **166** 241–265. MR3799646 https://doi.org/10.1016/j.jmva.2018.03.008

FERRARO, M. B. and GIORDANI, P. (2015). A toolbox for fuzzy clustering using the R programming language. *Fuzzy Sets and Systems* **279** 1–16. MR3392321 https://doi.org/10.1016/j.fss.2015.05.001

FLEISCHER, T., KLAJIC, J., AURE, M. R., LOUHIMO, R., PLADSEN, A. V., OTTESTAD, L., TOULEIMAT, N., LAAKSO, M., HALVORSEN, A. R. et al. (2017). DNA methylation signature (SAM40) identifies subgroups of the Luminal A breast cancer samples with distinct survival. *Oncotarget* **8** 1074–1082.

FRATELLO, M., CAIAZZO, G., TROJSI, F., RUSSO, A., TEDESCHI, G., TAGLIAFERRI, R. and ESPOSITO, F. (2017). Multi-view ensemble classification of brain connectivity images for neurodegeneration type discrimination. *Neuroinformatics* **15** 199–213.

GAYNANOVA, I. and LI, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics* **75** 1121–1132.

GODICHON-BAGGIONI, A., MAUGIS-RABUSSEAU, C. and RAU, A. (2020). Supplement to "Multiview cluster aggregation and splitting, with an application to multiomic breast cancer data." https://doi.org/10.1214/19-AOAS1317SUPP.

GU, Z., EILS, R. and SCHLESNER, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32** 2847–2849.

HAMID, J. S., HI, P., ROSLIN, N. M., LING, V., GREENWOOD, C. M. T. and BEYENE, J. (2009). Data integration in genetics and genomics: Methods and challenges. *Human Genomics Proteomics* 869093.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* 193–218.

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

KANEHISA, M., SATO, Y., KAWASHIMA, M., FURUMICHI, M. and TANABE, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. **44** D457–D462.

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

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

KUMAR, A., RAI, P. and DAUME, H. (2011). Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems* 1413–1421.

LIU, X., JI, S., GLÄNZEL, W. and DE MOOR, B. (2012). Multiview partitioning via tensor methods. *IEEE Trans. Knowl. Data Eng*. **25** 1056–1069.

LIU, J., LICHTENBERG, T., HOADLEY, K. A., POISSON, L. M., LAZAR, A. J., CHERNIACK, A. D., KOVATICH, A. J., BENZ, C. C., LEVINE, D. A. et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173** 400–416.

LOCK, E. F., HOADLEY, K. A., MARRON, J. S. and NOBEL, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7** 523–542. MR3086429 https://doi.org/10.1214/12-AOAS597

MO, Q., WANG, S., SESHAN, V. E., OLSHEN, A. B., SCHULTZ, N., SANDER, C., POWERS, R. S., LADANYI, M. and SHEN, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. USA* **110** 4245–4250.

PAQUET, E. R. and HALLETT, M. T. (2000). Absolute assignment of breast cancer intrinsic molecular subtype. *J. Natl. Cancer Inst*. **107**.

PARKER, J. S., MULLINS, M., CHEANG, M. C. U., LEUNG, S., VODUC, D., VICKERY, T., DAVIES, S., FAURON, C., HE, X. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol*. **27** 1160–1167.

PENSA, R. G., ROBARDET, C. and BOULICAUT, J.-F. (2005). A bi-clustering framework for categorical data. In *European Conference on Principles of Data Mining and Knowledge Discovery* 643–650. Springer, Berlin.

PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H. et al. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747–752.

RAPPOPORT, N. and SHAMIR, R. (2018). Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Res*. **gky889**.

RAU, A., FLISTER, M., RUI, H. and AUER, P. L. (2018). Exploring drivers of gene expression in the Cancer Genome Atlas. *Bioinformatics* **bty551**.

SERRA, A., FRATELLO, M., FORTINO, V., RAICONI, G., TAGLIAFERRI, R. and GRECO, D. (2015). Mvda: A multi-view genomic data integration methodology. *BMC Bioinform*. **16** 261.

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

SHEN, R., MO, Q., SCHULTZ, N., SESHAN, V. E., OLSHEN, A. B., HUSE, J., LADANYI, M. and SANDER, C. (2012). Integrative subtype discovery in glioblastoma using icluster. *PLoS ONE* **7** e35236.

TASKESEN, E., HUISMAN, S. M. H., MAHFOUZ, A., KRIJTHE, J. H., DE RIDDER, J., VAN DE STOLPE, A., VAN DEN AKKER, E., VERHEAGH, W. and REINDERS, M. J. T. (2016). Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Sci. Rep*. **6**.

THE CANCER GENOME ATLAS NETWORK (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

THE CANCER GENOME ATLAS NETWORK, WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., MILLS SHAW, K. R., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C. et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45** 1113–1120.

WANG, Y. and CHEN, L. (2017). Multi-view fuzzy clustering with minimax optimization for effective clustering of data from multiple sources. *Expert Syst*. *Appl*. **72** 457–466.

WANG, B., MEZLINI, A. M., DEMIR, F., FIUME, M., TU, Z., BRUDNO, M., HAIBE-KAINS, B. and GOLDENBERG, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat*. *Methods* **11** 333–337.

WICKHAM, H. (2016). *Ggplot*2: *Elegant Graphics for Data Analysis*. Springer, New York.

XU, C., TAO, D. and XU, C. (2013). A survey on multi-view learning. Preprint. Available at arXiv:1304.5634.

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

YANG, Y. and WANG, H. (2018). Multi-view clustering: A survey. *Big Data Mining and Analytics* **1** 83–107.

YU, Q., RISK, B. B., ZHANG, K. and MARRON, J. S. (2017). Jive integration of imaging and behavioral data. *NeuroImage* **152** 38–49.

ZAPPIA, L. and OSHLACK, A. (2018). Clustering trees: A visualisation for evaluating clusterings at multiple resolutions. *GigaScience* **7**.