Approximate Graph Laplacians for Multimodal Data Clustering

Aparajita Khan and Pradipta Maji, *Senior Member, IEEE*

Abstract—One of the important approaches of handling data heterogeneity in multimodal data clustering is modeling each modality using a separate similarity graph. Information from the multiple graphs is integrated by combining them into a unified graph. A major challenge here is how to preserve cluster information while removing noise from individual graphs. In this regard, a novel algorithm, termed as CoALa, is proposed that integrates noise-free approximations of multiple similarity graphs. The proposed method first approximates a graph using the most informative eigenpairs of its Laplacian which contain cluster information. The approximate Laplacians are then integrated for the construction of a low-rank subspace that best preserves overall cluster information of multiple graphs. However, this approximate subspace differs from the full-rank subspace which integrates information from all the eigenpairs of each Laplacian. Matrix perturbation theory is used to theoretically evaluate how far approximate subspace deviates from the full-rank one for a given value of approximation rank. Finally, spectral clustering is performed on the approximate subspace to identify the clusters. Experimental results on several real-life cancer and benchmark data sets demonstrate that the proposed algorithm significantly and consistently outperforms state-of-the-art integrative clustering approaches.

Index Terms—Integrative clustering, low-rank approximation, graph Laplacian, spectral clustering, multi-view learning, matrix perturbation theory

1 Introduction

Advancement in information acquisition technologies has made multimodal data ubiquitous in numerous real-world application domains like social networking [1], image processing [2], [3], 3D modeling [4], cancer biology [5], to name a few. Whole-genome sequencing project has given rise to a wide variety of "omics" data, which include genomic, epigenomic, transcriptomic, and proteomic data. The system-level insight, provided by different omics data, has led to numerous scientific discoveries and clinical applications over the past decade [6]. Cancer subtype identification has emerged out to be a major clinical application of multi-omics study. It can provide deeper understanding of disease pathogenesis and design of targeted therapies. While each type of omic data reflects the characteristic traits of a specific molecular level, integrative analysis of multi-omics data, which considers the biological variations across multiple molecular levels, can reveal novel cancer subtypes.

Integrative clustering is the primary tool for identification of disease subtypes from multi-omics data [7], [8]. The main challenge is how to integrate information appropriately, obtained from different modalities. Naive integration of different modalities with varying scales may give inconsistent results. Another challenge is to handle efficiently the 'high dimension-low sample size' nature of the individual data sets, which degrades the signal-to-noise ratio in the data and makes clustering computationally expensive.

Separate clustering followed by manual integration is a frequently used approach to analyze multiple omics data sets for its simplicity. Cluster-of-cluster assignment [9] and Bayesian consensus clustering [10] are two such approaches, which first cluster each modality separately and the individual clustering solutions are then combined to get the final cluster assignments. However, the integration of separate clustering solutions fails to capture cross-platform correlations and shared joint structure. On the other hand, some of the direct integrative approaches, like super k-means [11], iCluster [12], iCluster+ [13], LRAcluster [14], joint and individual variance explained (JIVE) [15], and angle-based JIVE (A-JIVE) [16], proceed by concatenating the individual modalities to get the integrated data which is then used for clustering. As the naive concatenation of different modalities may degrade the signal-to-noise ratio of the data, most of the direct integrative approaches first extract a low-rank subspace representation of the high dimensional integrated data and then clustering is performed in the reduced subspace [12], [13], [14].

In multi-omics data, different modalities vary immensely in terms of unit and scale. For instance, RNA sequence based gene expression data consists of RPM (reads per million) values having six-orders of magnitude, while DNA methylation data consists of β values which lie in [0, 1]. So, concatenation of features from these heterogeneous modalities would reflect only the properties of features having high variance. In order to capture the inherent properties of different modalities, it is essential to model the variations within each modality separately and then integrate them using a common platform. One widely used approach is to

• *The authors are with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, West Bengal 700108, India. E-mail: {aparajitak_r, pmaji}@isical.ac.in.*

Manuscript received 24 May 2019; accepted 29 Sept. 2019. Date of publication 4 Oct. 2019; date of current version 3 Feb. 2021.
(Corresponding author: Pradipta Maji.)
Recommended for acceptance by S. Kaski.
Digital Object Identifier no. 10.1109/TPAMI.2019.2945574

model each individual modality using a separate similarity graph. The individual similarity graphs are constructed in such a way that their vertices represent the samples, while their edges are weighted by the pairwise affinities between the samples of the respective modalities. The challenge is then how to integrate information efficiently from multiple similarity graphs. This comes under the paradigm of multi-view learning [17], [18], [19], [20], [21], [22], [23], where the main objective is to learn a unified view that is sufficiently "close" to all the views in some sense. In most multi-view learning algorithms, spectral clustering [24], [25], [26] is performed on the similarity graph corresponding to the unified view to identify the clusters of a given data set. The spectral clustering uses spectrum of the graph Laplacian [27] to identify the clusters in a data set. It has been shown in [24] that the relaxed solution to the k cluster indicators of a data set is given by the eigenvectors corresponding to the k smallest eigenvalues of its graph Laplacian. Hence, spectral clustering algorithms perform simple k-means on the k smallest eigenvectors of the graph Laplacian. However, it also implies that only a few eigenvectors of the Laplacian contain the cluster discriminatory information of the data set. The remaining eigenvectors may not necessarily encode cluster information and may reflect background noise. As a consequence, a major drawback of these multi-view algorithms is that both similarity graphs and their Laplacians, constructed from different views, inherently contain noisy information. This unwanted noise of the individual views may get propagated into the unified view during integration. This can degrade the quality of the cluster structure inferred from the unified view. Therefore, it is essential to prevent the noise in the individual views from being propagated into the unified view.

In this regard, the paper presents a novel algorithm, termed as CoALa (Convex-combination of Approximate Laplacians), which integrates noise-free approximations of multiple similarity graphs. The proposed method models each modality using a separate similarity graph, as different modalities are highly heterogeneous in nature and are measured in different scales. The noise in each individual graph is eliminated by approximating it using the most informative eigenpairs of its Laplacian which contain cluster information. The approximate Laplacians are then integrated and a low-rank subspace is constructed that best preserves the overall cluster information of multiple graphs. The graphs are integrated using a convex combination, where they are weighted according to the quality of their inherent cluster structure. Hence, noisy graphs have lower impact on the final subspace compared to the ones with good cluster structure. However, the approximate subspace constructed by the proposed method differs from the full-rank subspace that integrates information from all the eigenpairs of each Laplacian. The matrix perturbation theory is used to theoretically upper bound the difference between the full-rank and approximate subspaces, as a function of the approximation rank. It is shown, both theoretically and experimentally, that the approximate subspace converges to the full-rank one as the rank of approximation approaches to the full-rank of the individual Laplacians. Finally, the efficacy of clustering in the approximate subspace is extensively studied and compared with different existing integrative clustering approaches, on several real-life multi-omics cancer data sets. The results on benchmark data sets from other domains like image processing and social networks are also provided to establish the generality of the proposed approach.

The rest of this paper is organized as follows: Section 2 introduces the basics of graph Laplacian and its properties, while Section 3 presents the proposed CoALa algorithm for multimodal data clustering. Section 4 upper bounds the difference between full-rank and approximate subspaces. Experimental results and comparison with existing approaches are presented in Section 5. Section 6 concludes the paper.

2 Basics of Graph Laplacian

Given a set of samples or objects $X = \{x_1, \ldots, x_i, \ldots, x_n\}$, and a similarity matrix $W = [w(i,j)]_{n\times n}$, where $x_i \in \mathbb{R}^d$ and $w(i,j) = w(j,i) \geq 0$ is the similarity between objects x_i and x_j, the intuitive goal of clustering is to partition the objects into several groups such that objects in the same group are similar to each other, while those in different groups are dissimilar. The problem of clustering can also be approached from a graph theoretic point of view, where the data set X can be represented as an undirected similarity graph $G = (V,\ E)$ having vertex set $V = \{v_1, \ldots, v_i, \ldots, v_n\}$, where each vertex v_i represents the object x_i, and the edge between vertices v_i and v_j is weighted by the similarity $w(i,j)$. The degree $\widetilde{d}_i$ of vertex v_i is given by $\widetilde{d}_i = \sum_{j=1}^{n} w(i,j)$, and the degree matrix D is given by the diagonal matrix

$$D = diag(\widetilde{d}_1, \ldots, \widetilde{d}_i, \ldots, \widetilde{d}_n). \tag{1}$$

Given the number of clusters k, clustering can be viewed as partitioning the graph G into k subgraphs such that edges between different subgraphs have lower weights, while edges within a subgraph have higher weights. For a subset of vertices $A \subset V$, let its complement $\bar{A}$ be given by $\bar{A} = V \setminus A$. A measure of size of subset A can be given by $vol(A) = \sum_{v_i \in A} \widetilde{d}_i$. For two not necessarily disjoint subsets $A, B \subset V$, let

$$\mathbb{C}(A,B) = \sum_{v_i \in A, v_j \in B} w(i,j). \tag{2}$$

For a subset A of vertices, $\mathbb{C}(A,\bar{A})$ gives the weight of the cut that separates the vertices in A from the rest of vertices in G. So, given the number of subsets k, the graph partitioning problem finds a partition $A_1, \ldots, A_k$ of V such that it minimizes the cut weight $\mathbb{C}(A_i,\bar{A}_i)$ for each A_i. However, minimizing only $\mathbb{C}(A_i,\bar{A}_i)$ can lead to singleton subsets A_i's. In clustering, it is desirable to achieve clusters with reasonably large set of points. So, minimizing $\frac{\mathbb{C}(A_i,\bar{A}_i)}{vol(A_i)}$, instead of $\mathbb{C}(A_i,\bar{A}_i)$, would constrain each subset A_i to be fairly large. The most common optimization problem in this regard is the normalized cut or $Ncut$ [28], defined as

$$\begin{aligned} \underset{A_1,\ldots,A_k}{\text{minimize}} \quad & Ncut(A_1, \ldots, A_k) = \frac{1}{2}\sum_{i=1}^{k} \frac{\mathbb{C}(A_i,\bar{A}_i)}{vol(A_i)} \\ \text{such that} \quad & A_i \cap A_j = \emptyset \text{ and } \bigcup_{i=1}^{k} A_i = V. \end{aligned} \tag{3}$$

However, the above optimization problem is NP-hard [29]. The spectral clustering [24] provides a computationally tractable solution to this Ncut problem. It analyzes the spectrum or eigenspace of graph Laplacian to find the solution [30]. The graph Laplacian and several its variants are described next.

Let $G = (V, E)$ be a graph with similarity matrix W and degree matrix D as given by (1). The matrix $(D - W)$ is called the Laplacian of graph G [30], and the normalized Laplacian of G is given by [27]

$$\mathcal{L} = D^{-1/2}(D - W)D^{-1/2} = I - D^{-1/2}WD^{-1/2}, \tag{4}$$

where I is identity matrix of appropriate order. Two important properties of normalized Laplacian are as follows [27]:

Property 1. *$\mathcal{L}$ is symmetric and positive semi-definite.*

Property 2. *The eigenvalues of $\mathcal{L}$ lie in $[0, 2]$.*

Let the k clusters in a data set X be represented by the indicator matrix

$$\mathcal{E} = [e_1 \dots e_j \dots e_k] \in \mathbb{R}^{n \times k}, \tag{5}$$

where e_j is the indicator vector in $\mathbb{R}^n$ for the jth cluster, that is, $e_j \in \{0, 1\}^n$, such that e_j has a nonzero component only for the points in the jth cluster. Let the r largest eigenvectors of a matrix correspond to its r largest eigenvalues. It is shown in [24] that if the constraint on the cluster indicators e_j's is relaxed such that $e_j \in [0, 1]$, then the real-valued solution to the indicators $e_1, \dots, e_k$ is given by the k smallest eigenvectors of the normalized Laplacian $\mathcal{L}$. The normalized spectral clustering algorithm by Ng et al. [31] is described in Algorithm 1. The spectral clustering algorithm [28], [31] first computes the graph Laplacian and then k-means clustering is performed on its k smallest eigenvectors. The main advantage of spectral clustering is that it transforms the representations of the objects $\{x_i\}$ from their original space to an indicator subspace where the cluster characteristics are more prominent. As the cluster properties are enhanced in this new subspace, even simple clustering algorithms, such as k-means, have no difficulty in distinguishing the clusters.

Algorithm 1. Normalized Spectral Clustering [31]

Input: Similarity matrix W, number of clusters k.
Output: Clusters $A_1, \dots, A_k$.

1: Construct degree matrix D and normalized Laplacian $\mathcal{L}$ as in (1) and (4), respectively.
2: Find eigenvectors $U = [u_1 \dots u_k]$ corresponding to k smallest eigenvalues of $\mathcal{L}$.
3: Normalize the rows of U, i.e. $U = diag(UU^T)^{-\frac{1}{2}}U$.
4: Perform clustering on the rows of U using k-means algorithm.
5: **Return** clusters $A_1, \dots, A_k$ from k-means clustering.

In a Laplacian matrix, the necessary cluster information is embedded in its k smallest eigenvectors. However, based on Eckart-Young theorem [32], the best low-rank approximation of a symmetric matrix can be constructed from its few largest eigenpairs. So, the best low-rank approximation of a Laplacian matrix primarily encodes noise, rather than cluster information. In the proposed work, the final subspace of a multimodal data set is constructed from low-rank approximations of individual graph Laplacians. So, in order to reflect the cluster information in the low-rank approximations, the shifted Laplacian [33] is used, which is defined as

$$L = 2I - \mathcal{L} = I + D^{-1/2}WD^{-1/2}. \tag{6}$$

The following property of shifted Laplacian makes it feasible to reflect the cluster information in its best low-rank approximation.

Property 3. *If (λ, v) is an eigenvalue-eigenvector pair of normalized Laplacian $\mathcal{L}$, then $(2 - \lambda, v)$ is an eigenpair of shifted Laplacian L* [33].

Property 3 implies that the k smallest eigenvalues and eigenvectors of normalized Laplacian $\mathcal{L}$ correspond to the k largest eigenvalues and eigenvectors of shifted Laplacian L. Therefore, the relaxed solution to the cluster indicators $e_1, \dots, e_k$ in (5) is given by the k largest eigenvectors of L. So, the best rank k approximation of L also encodes its cluster information. As the eigenvalues of $\mathcal{L}$ lie in $[0, 2]$, the eigenvalues of L also lie in $[0, 2]$. Moreover, L is symmetric and positive semi-definite [33].

3 Proposed Method

This section presents a novel algorithm to extract a low-rank joint subspace from multiple graph Laplacians. Some analytical formulations, required for subspace construction, are reported next, prior to describing the proposed algorithm.

3.1 Convex Combination of Graph Laplacians

Let a multimodal data set, consisting of M modalities, be given by $X_1, \dots, X_m, \dots, X_M$. Each modality $X_m \in \mathbb{R}^{n \times d_m}$ represents the observations for same set of n samples from the mth data source. Let X_m be encoded by the similarity graph G_m having similarity matrix W_m and degree matrix D_m. The shifted Laplacian for modality X_m is given by

$$L_m = I + D_m^{-1/2}W_mD_m^{-1/2}. \tag{7}$$

Let the eigen-decomposition of L_m be given by

$$L_m = U_m\Sigma_mU_m^T, \tag{8}$$

where $U_m = [u_1^m, \dots, u_n^m] \in \mathbb{R}^{n \times n}$ contains the eigenvectors of L_m in its columns, B^T denotes the transpose of B, and $\Sigma_m = diag(\lambda_1^m, \dots, \lambda_n^m)$, where $2 \geq \lambda_1^m \geq \dots \geq \lambda_n^m \geq 0$. For a given rank r, the eigen-decomposition of shifted Laplacian L_m in (8) can be partitioned as follows:

$$\begin{aligned} L_m &= U_m\Sigma_mU_m^T \\ &= \begin{bmatrix} U_m^r & U_m^{r\perp} \end{bmatrix} \begin{bmatrix} \Sigma_m^r & \mathbf{0} \\ \mathbf{0} & \Sigma_m^{r\perp} \end{bmatrix} \begin{bmatrix} U_m^r & U_m^{r\perp} \end{bmatrix}^T \\ &= U_m^r\Sigma_m^r(U_m^r)^T + U_m^{r\perp}\Sigma_m^{r\perp}(U_m^{r\perp})^T \\ &= L_m^r + L_m^{r\perp}, \end{aligned} \tag{9}$$

where $\mathbf{0}$ denotes a matrix of all zeros of appropriate order, $\Sigma_m^r = diag(\lambda_1^m, \dots, \lambda_r^m)$ consists of the r largest eigenvalues and U_m^r contains the corresponding r eigenvectors in its columns. Similarly, $\Sigma_m^{r\perp}$ and $U_m^{r\perp}$ contain the remaining $(n - r)$ eigenvalues $\lambda_{r+1}^m, \dots, \lambda_n^m$ and eigenvectors, respectively.

Thus, L_m^r is the rank r approximation of L_m using the r largest eigenpairs, and $L_m^{r\perp}$ is the approximation using the remaining $(n-r)$ eigenpairs. Given the number of clusters k, the properties of shifted Laplacian imply that the relaxed solution to the cluster indicators is given by the k largest eigenvectors of L_m. Therefore, for each modality X_m, a rank r eigenspace representation is constructed, where $k \leq r \ll n$, which encodes the cluster information of its shifted Laplacian L_m. Choosing the rank r to be greater than k allows extra information from each Laplacian at the initial stage.

The rank r eigenspace of shifted Laplacian L_m for modality X_m is defined by a two-tuple:

$$\Psi(L_m^r) = \langle U_m^r, \Sigma_m^r \rangle. \tag{10}$$

The individual graph Laplacians contain the cluster information of their respective modalities. Multiple modalities are integrated using a convex combination $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_m, \ldots, \alpha_M]$ of individual shifted Laplacians, defined by

$$\mathbf{L} = \sum_{m=1}^{M} \alpha_m L_m, \text{ such that } \alpha_m \geq 0 \text{ and } \sum_{m=1}^{M} \alpha_m = 1. \tag{11}$$

The matrix $\mathbf{L}$ is called the joint shifted Laplacian and it has the following properties.

Property 4. $\mathbf{L}$ *is symmetric and positive semi-definite.*

Proof. Each shifted Laplacian L_m is symmetric for $m = 1, 2, \ldots, M$. So,

$$\mathbf{L}^T = \left(\sum_{m=1}^{M} \alpha_m L_m\right)^T = \sum_{m=1}^{M} \alpha_m L_m^T = \sum_{m=1}^{M} \alpha_m L_m = \mathbf{L}.$$

Therefore, $\mathbf{L}$ is symmetric. By Property 3, each L_m is positive semi-definite, so, for any vector $a \in \mathbb{R}^n$, $a^T L_m a \geq 0$. Therefore,

$$a^T \mathbf{L} a = a^T \left(\sum_{m=1}^{M} \alpha_m L_m\right) a = \sum_{m=1}^{M} \alpha_m \left(a^T L_m a\right) \geq 0,$$

as $\alpha_m \geq 0$. Therefore, $\mathbf{L}$ is positive semi-definite. □

Property 5. $\mathbf{L}$ *has* n *eigenvalues* $\gamma_1 \geq \ldots \geq \gamma_i \geq \ldots \geq \gamma_n$, *where* $\gamma_i \in [0, 2]$.

Proof. By Property 3, the eigenvalues of each individual shifted Laplacian L_m lie in $[0, 2]$ for $m = 1, 2, \ldots, M$. So, the maximum eigenvalue of L_m and $\alpha_m L_m$ satisfy $\lambda_1^m \leq 2$ and $\alpha_m \lambda_1^m \leq 2\alpha_m$, respectively. Since each Laplacian L_m is a real symmetric matrix, it is also Hermitian as it is equal to its own conjugate transpose. Now, $\mathbf{L}$ is the sum of M Hermitian matrices. So, using Weyl's inequality [34], which bounds the eigenvalues of the sum of two Hermitian matrices, we get

$$\gamma_1 \leq \sum_{m=1}^{M} \alpha_m \lambda_1^m \leq \sum_{m=1}^{M} 2\alpha_m = 2. \tag{12}$$

$\mathbf{L}$ is positive semi-definite, so all of its eigenvalues $\gamma_i \geq 0$. Therefore, $\gamma_i \in [0, 2]$. □

Hence, the joint shifted Laplacian $\mathbf{L}$ has similar properties as individual shifted Laplacians L_m's have. In rest of the paper, the term joint Laplacian is used to refer to the joint shifted Laplacian.

3.2 Construction of Joint Eigenspace

This subsection describes the construction of eigenspace of the joint Laplacian from low-rank eigenspaces of individual shifted Laplacians. Let eigen-decomposition of $\mathbf{L}$ be given by

$$\mathbf{L} = \mathbf{Z}\Gamma\mathbf{Z}^T, \tag{13}$$

where $\mathbf{Z}$ consists of the eigenvectors of $\mathbf{L}$ in its columns and $\Gamma = diag(\gamma_1, \ldots, \gamma_n)$ is the diagonal matrix of eigenvalues arranged in descending order of magnitude. The "full-rank" eigenspace of $\mathbf{L}$ is given by the two-tuple

$$\Psi(\mathbf{L}^r) = \langle \mathbf{Z}^r, \Gamma^r \rangle, \tag{14}$$

where $\Gamma^r = diag(\gamma_1, \ldots, \gamma_r)$ and $\mathbf{Z}^r$ contains the eigenvectors corresponding to the eigenvalues in Γ^r. The term "full-rank" is used to imply that in $\mathbf{L}$, the complete information of all the eigenpairs of each Laplacian is considered during convex combination. The superscript r in $\Psi(\mathbf{L}^r)$ indicates that the eigenspace has rank r. The "approximate" joint Laplacian is defined as

$$\mathbf{L}^{r*} = \sum_{m=1}^{M} \alpha_m L_m^r. \tag{15}$$

Thus, $\mathbf{L}^{r*}$ is the convex combination of best rank r approximation of individual shifted Laplacians. For each shifted Laplacian L_m, instead of storing its complete eigen-decomposition, only the r largest eigenpairs are stored in its eigenspace $\Psi(L_m^r)$. Given these eigenspaces $\Psi(L_m^r)$s, the proposed method aims at construction of the rank r eigenspace $\Psi(\mathbf{L}^{r*})$, of the approximate joint Laplacian $\mathbf{L}^{r*}$. The main advantage of this construction is that it finds the joint eigenspace from the r largest eigenpairs of individual Laplacians. The cluster information of individual modalities is expected to embed in the k largest eigenpairs of their respective shifted Laplacians. Hence, storing $r \geq k$ eigenpairs allows for some extra information from each Laplacian as well as gets rid of the noisy information in the $(n-r)$ eigenpairs. Thus, the approximate eigenspace $\Psi(\mathbf{L}^{r*})$, constructed from the r largest eigenpairs, is expected to preserve better cluster information compared to the full-rank eigenspace $\Psi(\mathbf{L}^r)$.

One straight forward approach for the construction of eigenspace of $\mathbf{L}^{r*}$ is to first solve the eigen-decomposition of the individual L_m's, reconstruct the L_m^r's from the top r eigenpairs of respective L_m's, combine the reconstructed L_m^r's using the convex combination and then perform another eigen-decomposition on the combination $\mathbf{L}^{r*}$. This requires solving a total of $(M+1)$ eigen-decompositions of size $(n \times n)$. However, in the proposed method, the eigenspaces $\Psi(L_m^r)$'s of the individual Laplacians inorder are used to construct a smaller eigenvalue problem of size $(Mr \times Mr)$ whose solution is used to get the required eigenspace $\Psi(\mathbf{L}^{r*})$. So, it requires solving M eigen-decompositions of size $(n \times n)$ and one of size $(Mr \times Mr)$, where $Mr \ll n$. This makes the proposed approach computationally more efficient.

The block decomposition of L_m in (9) gives us that $L_m^r = U_m^r \Sigma_m^r (U_m^r)^T$. So,

$$\mathbf{L}^{r*} = \sum_{m=1}^{M} \alpha_m L_m^r = \sum_{m=1}^{M} \alpha_m U_m^r \Sigma_m^r (U_m^r)^T. \tag{16}$$

The expansion of $\mathbf{L}^{r*}$ in (16) implies that the subspace spanned by its columns is same as the one spanned by the union of the columns of U_m^r for $m = 1, \ldots, M$. Let that subspace be given by

$$\mathcal{J}^r = span\left(\bigcup_{m=1}^{M} \mathcal{C}(U_m^r) \right), \tag{17}$$

where $\mathcal{C}(B)$ denotes the column space of matrix B. To compute the eigenspace of $\mathbf{L}^{r*}$, the first step is to construct a sufficient basis that spans the subspace $\mathcal{J}^r$. Since $\mathcal{J}^r$ is the union of M subspaces, its basis is constructed iteratively in M steps. At step 1, the initial basis $\mathbf{U}_1$ is given by

$$\mathbf{U}_1 = U_1^r, \tag{18}$$

which spans the subspace $\mathcal{C}(U_1^r)$. At step m, let the union of m subspaces be given by the subspace

$$\mathcal{J}_m^r = span\left(\bigcup_{j=1}^{m} \mathcal{C}(U_j^r) \right) \tag{19}$$

and let its orthonormal basis be given by $\mathbf{U}_m \in \mathbb{R}^{n \times r}$. Given the basis $\mathbf{U}_m$ obtained at step m, and the basis U_{m+1}^r for L_{m+1}^r, the basis $\mathbf{U}_{m+1}$ at step $(m+1)$ is constructed as follows.

The basis $\mathbf{U}_{m+1}$ has to span both the subspaces $\mathcal{J}_m^r$ and $\mathcal{C}(U_{m+1}^r)$. The column vectors of $\mathbf{U}_m$ themselves form a basis for the subspace $\mathcal{J}_m^r$. Therefore, a sufficient basis for the subspace $\mathcal{J}_{m+1}^r$ can be constructed by appending a basis Υ_{m+1} that spans the subspace orthogonal to $\mathcal{J}_m^r$. The construction of basis Υ_{m+1} begins by computing the residue of each basis vector in U_{m+1}^r with respect to the basis $\mathbf{U}_m$. To compute the residues, each vector in U_{m+1}^r is projected on each of the basis vectors in $\mathbf{U}_m$. In matrix notation, this is given by

$$S_{m+1} = \mathbf{U}_m^T U_{m+1}^r. \tag{20}$$

The matrix S_{m+1} gives the magnitude of projection of the columns of U_{m+1}^r onto the orthonormal basis $\mathbf{U}_m$. The projected component P_{m+1} of U_{m+1}^r, lying in the subspace $\mathcal{J}_m^r$, is obtained by multiplying the projection magnitudes in S_{m+1} by the corresponding basis vectors in $\mathbf{U}_m$, given by

$$P_{m+1} = \mathbf{U}_m S_{m+1}. \tag{21}$$

The residual component Q_{m+1} of U_{m+1}^r is obtained by subtracting projected component P_{m+1} from itself, given by

$$Q_{m+1} = U_{m+1}^r - P_{m+1}. \tag{22}$$

An orthogonal basis Υ_{m+1} for the residual space, spanned by columns of Q_{m+1}, can be obtained by Gram-Schmidt orthogonalization of Q_{m+1}. The basis Υ_{m+1} spans the subspace orthogonal to $\mathcal{J}_m^r$. Therefore, a sufficient basis for the subspace $\mathcal{J}_{m+1}^r$ is obtained by appending Υ_{m+1} to $\mathbf{U}_m$, given by

$$\mathbf{U}_{m+1} = [\, \mathbf{U}_m \quad \Upsilon_{m+1} \,]. \tag{23}$$

Let $\Upsilon_1 = \mathbf{U}_1$. After M steps, the basis $\mathbf{U}_M$, for the subspace $\mathcal{J}^r$ in (17), is given by

$$\mathbf{U}_M = [\, \Upsilon_1 \quad \Upsilon_2 \quad \ldots \quad \Upsilon_M \,]. \tag{24}$$

Let the eigen-decomposition of $\mathbf{L}^{r*}$ be given by

$$\mathbf{L}^{r*} = \mathbf{V} \Pi \mathbf{V}^T, \tag{25}$$

where $\mathbf{V} \in \mathbb{R}^{n \times n}$ contains the eigenvectors of $\mathbf{L}^{r*}$ in its columns, and $\Pi = diag(\pi_1, \ldots, \pi_n)$ contains the eigenvalues arranged in descending order. The eigenvectors in $\mathbf{V}$ span the column space of $\mathbf{L}^{r*}$, which from (17) is the subspace $\mathcal{J}^r$. $\mathbf{U}_M$ is also a basis for $\mathcal{J}^r$. These two bases $\mathbf{V}$ and $\mathbf{U}_M$ span the same subspace $\mathcal{J}^r$ and they differ by a rotation. So,

$$\mathbf{V} = \mathbf{U}_M \mathbf{R}, \tag{26}$$

where $\mathbf{R}$ is an orthogonal rotation matrix. The eigenvalues Π in (25) and the rotation matrix $\mathbf{R}$ in (26) are obtained as follows.

$$\begin{aligned}
&\mathbf{L}^{r*} = \sum_{m=1}^{M} \alpha_m U_m^r \Sigma_m^r (U_m^r)^T, && \text{[from (16)]} \\
&\Rightarrow \mathbf{V} \Pi \mathbf{V}^T = \sum_{m=1}^{M} \alpha_m U_m^r \Sigma_m^r (U_m^r)^T, && \text{[from (25)]} \\
&\Rightarrow (\mathbf{U}_M \mathbf{R}) \Pi (\mathbf{U}_M \mathbf{R})^T = \sum_{m=1}^{M} \alpha_m U_m^r \Sigma_m^r (U_m^r)^T, && \text{[from (26)]} \\
&\Rightarrow \mathbf{R} \Pi \mathbf{R}^T = \mathbf{U}_M^T \left(\sum_{m=1}^{M} \alpha_m U_m^r \Sigma_m^r (U_m^r)^T \right) \mathbf{U}_M, \\
&\Rightarrow \mathbf{R} \Pi \mathbf{R}^T = \sum_{m=1}^{M} \alpha_m \mathbf{U}_M^T U_m^r \Sigma_m^r (U_m^r)^T \mathbf{U}_M, \\
&\Rightarrow \mathbf{R} \Pi \mathbf{R}^T = \sum_{m=1}^{M} \alpha_m \begin{bmatrix} \Upsilon_1^T \\ \vdots \\ \Upsilon_M^T \end{bmatrix} U_m^r \Sigma_m^r (U_m^r)^T [\, \Upsilon_1 \quad \ldots \quad \Upsilon_M \,], \\
&\Rightarrow \mathbf{R} \Pi \mathbf{R}^T = \sum_{m=1}^{M} \alpha_m H_m,
\end{aligned} \tag{27}$$

where $H_m \in \mathbb{R}^{(Mr \times Mr)}$ is given by

$$H_m = [\Upsilon_1 \ldots \Upsilon_M]^T U_m^r \Sigma_m^r (U_m^r)^T [\Upsilon_1 \ldots \Upsilon_M]. \tag{28}$$

While constructing the basis $\mathbf{U}_M$, the Υ_p's are appended iteratively such that whenever $p > m$, Υ_p is orthogonal to U_m^r and $\Upsilon_p^T U_m^r = 0$. Thus, the matrix H_m can be partitioned into M^2 blocks, each of size $(r \times r)$, and the (i,j)th block of H_m is given by

$$H_m(i,j) = \begin{cases} \Upsilon_i^T U_m^r \Sigma_m^r (U_m^r)^T \Upsilon_j & \text{if } i \leq m \text{ and } j \leq m, \\ 0 & \text{if } i > m \text{ or } j > m. \end{cases}$$

$$\text{Let} \quad \mathbf{H} = \sum_{m=1}^{M} \alpha_m H_m; \qquad \Rightarrow \mathbf{H} = \mathbf{R} \Pi \mathbf{R}^T. \tag{29}$$

This implies that solving the eigen-decomposition of the $(Mr \times Mr)$ matrix $\mathbf{H}$, the eigenvalues Π of $\mathbf{L}^{r*}$ and the rotation matrix $\mathbf{R}$ are obtained. Then, $\mathbf{R}$ is substituted in (26) to get the eigenvectors of $\mathbf{L}^{r*}$ in columns of $\mathbf{V}$. The rank r eigenspace of $\mathbf{L}^{r*}$ is then given by the two-tuple

$$\Psi(\mathbf{L}^{r*}) = \langle \mathbf{V}^r, \Pi^r \rangle, \tag{30}$$

where $\Pi^r = diag(\pi_1, \ldots, \pi_r)$ consists of the r largest eigenvalues of Π arranged in descending order, and $\mathbf{V}^r$ contains the corresponding r eigenvectors in its columns.

3.3 Proposed Algorithm

Given similarity matrices $W_1, \ldots, W_M$ corresponding to M modalities $X_1, \ldots, X_M$, convex combination vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]$ and rank r, the proposed algorithm, termed as CoALa, extracts a rank r eigenspace for the approximate joint Laplacian $\mathbf{L}^{r*}$. For each modality X_m, the proposed algorithm first computes the eigen-decomposition of its shifted Laplacian L_m and then stores the $r \geq k$ largest eigenpairs in its eigenspace. Next, it iteratively computes the basis $\mathbf{U}_M$ and the eigen-decomposition of the new eigenvalue problem $\mathbf{H}$. The eigenvalues of $\mathbf{L}^{r*}$ are given by the eigenvalues of $\mathbf{H}$, while the eigenvectors of $\mathbf{H}$ are used to rotate the basis $\mathbf{U}_M$ and get the eigenvectors of $\mathbf{L}^{r*}$. Finally, k-means clustering is performed on the k largest eigenvectors of $\mathbf{L}^{r*}$ to get the clusters of the multimodal data set. The proposed algorithm is described in Algorithm 2.

Algorithm 2. Proposed CoALa Algorithm

Input: Similarity matrices $W_1, \ldots, W_M$, combination vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]$, number of clusters k, and rank $r \geq k$.
Output: Clusters $A_1, \ldots, A_k$.

1: **for** $m \leftarrow 1$ **to** M **do**
2: Construct degree matrix D_m and shifted normalized Laplacian L_m as in (1) and (7), respectively.
3: Compute the eigen-decomposition of L_m.
4: Store the r largest eigenvalues in Σ_m^r and corresponding eigenvectors in U_m^r in the rank r eigenspace of X_m.
5: **end for**
6: Compute initial basis $\mathbf{U}_1 \leftarrow U_1^r$.
7: **for** $m \leftarrow 1$ **to** $M - 1$ **do**
8: Compute S_{m+1}, projected component P_{m+1}, and residual component Q_{m+1} according to (20), (21), and (22), respectively.
9: $\Upsilon_{m+1} \leftarrow$ Gram-Schmidt orthogonalization of Q_{m+1}.
10: Update basis $\mathbf{U}_{m+1} \leftarrow [\,\mathbf{U}_m \quad \Upsilon_{m+1}\,]$.
11: **end for**
12: For each modality X_m, compute H_m as in (28).
13: Compute the new eigenvalue problem $\mathbf{H}$ as in (29).
14: Solve the eigen-decomposition of $\mathbf{H}$ to get $\mathbf{R}$ and Π.
15: Compute eigenvectors $\mathbf{V} \leftarrow \mathbf{U}_M\mathbf{R}$.
16: Compute joint eigenspace $\Psi(\mathbf{L}^{r*}) \leftarrow \langle \mathbf{V}^r, \Pi^r \rangle$ as in (30).
17: Find k largest eigenvectors $\mathbf{V}^k = [v_1 \ldots v_k]$.
18: Perform clustering on the rows of $\mathbf{V}^k$ using k-means algorithm.
19: **Return** clusters $A_1, \ldots, A_k$ from k-means clustering.

In the normalized spectral clustering by Ng et al. [31], the eigenvectors are row normalized (step 1 of Algorithm 1) before clustering. The advantage of this additional normalization has been shown for the ideal case where the similarity is zero between points belonging to different clusters and strictly positive between points in the same clusters. In such a situation, the eigenvalue 0 has multiplicity k, and the eigenvectors are given by the columns of $D^{\frac{1}{2}}\mathcal{E}$, where $\mathcal{E}$ is the ideal cluster indicator matrix as in (5). By normalizing each row by its norm, the eigenvector matrix coincides with the indicator matrix $\mathcal{E}$, and the points become trivial to cluster. Ng et al. [31] have also shown that when the similarity matrix is "close" to the ideal case, properly normalized rows tend to tightly cluster around an orthonormal basis. However, in real-life data sets, the clusters are generally not well-separated due to the high dimension and heterogeneous nature of different modalities. As a result, the similarity matrices deviate far from the ideal block diagonal ones. So, additional row normalization may lead to undesirable scaling which is not advantageous for the subsequent k-means clustering step. Therefore, row normalization is not recommended in the proposed algorithm.

3.4 Computational Complexity

In the proposed algorithm, the first step is to compute the eigenspace of each modality X_m. Given the similarity matrix W_m for modality X_m, its degree matrix D_m and shifted Laplacian L_m are computed in step 2 in $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$ time, respectively. Then, the eigen-decomposition of L_m is computed in step 2 which takes $\mathcal{O}(n^3)$ time for the $(n \times n)$ matrix. Therefore, for M modalities, the total complexity of initial eigenspace construction is $\mathcal{O}(Mn^3)$. Next, the basis $\mathbf{U}_M$ is constructed in M steps. At each step of basis construction, the matrices S_{m+1}, P_{m+1}, and Q_{m+1} are computed in step 2 of the algorithm. It takes $\mathcal{O}(nr^2)$ time. The Gram-Schmidt orthogonalization in step 2 also has complexity of $\mathcal{O}(nr^2)$ for $(n \times r)$ matrix Q_{m+1}. The total complexity of basis construction in steps 7-11 is $\mathcal{O}(nr^2)$. The new eigenvalue problem $\mathbf{H}$ of size $(Mr \times Mr)$ is formulated in steps 12-13, which takes $\mathcal{O}(M^3r^3)$ time, owing to matrix multiplications. The subsequent eigen-decomposition of $\mathbf{H}$ in step 14 also takes $\mathcal{O}(M^3r^3)$ time. The rotation of $\mathbf{U}_M$ in step 2 has complexity of $\mathcal{O}(nr^2)$. Finally, after the construction of joint eigenspace $\Psi(\mathbf{L}^{r*})$, k-means clustering is performed on $(n \times k)$ matrix $\mathbf{V}^k$ which has time complexity of $\mathcal{O}(t_{max}nk^2)$, where t_{max} is the maximum number of iterations the k-means algorithm runs.

Hence, the overall computational complexity of the proposed CoALa algorithm, to extract the joint eigenspace and perform spectral clustering on a multimodal data set, is $(\mathcal{O}(Mn^3 + nr^2 + M^3r^3 + nr^2 + t_{max}nk^2) =)\mathcal{O}(Mn^3)$, assuming $M, r, k \ll n$. It implies that the overall complexity of the proposed algorithm is dominated by the individual eigenspace construction of initial stage.

3.5 Choice of Convex Combination

The convex combination vector $\boldsymbol{\alpha}$ determines the weight of the influence of each Laplacian on the final eigenspace. According to Fiedler's theory of spectral graph partitioning [35], the algebraic connectivity or the Fiedler value of a graph G is the second minimum eigenvalue of the Laplacian of G. The Fiedler value represents the weight of the minimum cut that partitions the corresponding graph into two subgraphs. Moreover, by Property 3, the lower the

eigenvalue or cut-weight of the normalized Laplacian $\mathcal{L}$, the higher is the corresponding eigenvalue of its shifted Laplacian L. The smallest eigenvalue of $\mathcal{L}$ is 0 which corresponds to largest eigenvalue, λ_1, of L which is 2, and the second largest eigenvalue, λ_2, reflects how high is the separability of graph G. The corresponding eigenvector u_2, known as the Fiedler vector, can be used to partition the vertices of G [36]. For example, if the Fiedler vector is $u_2 = (u_{21}, \ldots, u_{2j}, \ldots, u_{2n})$, spectral partitioning finds a splitting value s such that the objects with $u_{2j} \leq s$ belong to a set, while that with $u_{2j} > s$ belong to other. Several popular choices for s have been proposed, or the standard 2-means algorithm can also be applied on u_2 to obtain a 2-partition. Once a 2-partition is obtained, Silhouette index [37] can internally assess the quality of the partition. Silhouette index lies between [-1, 1] and higher value indicates a better partition. A modality with good inherent cluster information is expected to have a higher Fiedler value as well as higher Silhouette index on the Fiedler vector. Thus, a measure of "relevance" of a modality X_m is defined as

$$\boldsymbol{\chi}_m = \frac{1}{4}\lambda_2^m \left[\mathcal{S}(u_2^m) + 1\right] \tag{31}$$

where λ_2^m is the second largest eigenvalue of shifted Laplacian L_m of X_m and u_2^m is the corresponding eigenvector. The term $(\mathcal{S}(u_2^m) + 1)$ lies in $[0, 2]$, while the value of λ_2^m can be at most 2. The factor $1/4$ acts as a normalizing factor which upper bounds the value of $\boldsymbol{\chi}$ to 1. Hence, the value of relevance measure $\boldsymbol{\chi}$ lies in $[0, 1]$. Higher value of $\boldsymbol{\chi}_m$ implies higher relevance and better cluster structure. Hence, $\boldsymbol{\chi}$ can be used to obtain a linear ordering of the modalities $X_1, \ldots, X_M$. Let $X_{(1)}, \ldots, X_{(m)}, \ldots, X_{(M)}$ be the ordering of $X_1, \ldots, X_m, \ldots, X_M$ based on decreasing value of relevance $\boldsymbol{\chi}$. In the convex combination vector $\boldsymbol{\alpha}$, the component $\alpha_{(m)}$ corresponding to the weighting factor of modality $X_{(m)}$ is given by

$$\alpha_{(m)} = \boldsymbol{\chi}_{(m)}\beta^{-m}, \text{ where } \beta > 1. \tag{32}$$

This implies that based on the index of $X_{(m)}$ in the ordering $X_{(1)}, \ldots, X_{(M)}$, the relevance value of $X_{(m)}$ is damped by a factor of β^m and then used as its contribution in the convex combination $\boldsymbol{\alpha}$. Thus, in $\boldsymbol{\alpha}$, the most relevant modality has contribution of $\frac{\boldsymbol{\chi}_{(1)}}{\beta}$, while the second most relevant one contributes $\frac{\boldsymbol{\chi}_{(2)}}{\beta^2}$, and so on. This assignment of $\boldsymbol{\alpha}$ upweights modalities with better cluster structure, while dampens the effect of irrelevant ones those having poor structure.

4 Quality of Eigenspace Approximation

The proposed algorithm constructs the eigenspace $\Psi(\mathbf{L}^{r*})$ from a convex combination of rank r approximations of the individual Laplacians L_m's. This eigenspace differs from the full-rank eigenspace $\Psi(\mathbf{L}^r)$, which is the convex combination of complete or full rank information of the individual Laplacians. In real-life multimodal data sets, the individual modalities inherently contain noisy information. The approximation approach prevents propagation of noise from the individual modalities into the final approximate eigenspace $\Psi(\mathbf{L}^{r*})$. As a consequence, the approximate subspace is expected to preserve better cluster structure compared to the full-rank one. However, in the ideal case, where the clusters in the individual modalities are well-separated, the approximation approach may loose some important information. So, the difference between the two eigenspaces $\Psi(\mathbf{L}^r)$ and $\Psi(\mathbf{L}^{r*})$ is evaluated as a function of the approximation rank r, and can be quantified in terms of their eigenvalues and eigenvectors. The difference between the eigenvalues can be measured directly in terms of their magnitude, while the difference between the eigenvectors is measured in terms of difference between the subspaces spanned by the two sets of eigenvectors. Principal angles between subspaces (PABS) [38], [39] is used here to measure the difference between two subspaces. The PABS is a generalization of the concept of angle between two vectors to a set of angles between two subspaces, which is defined next.

Definition 1. *Let $\mathcal{Y}$ and $\mathcal{Z}$ be two subspaces of $\mathbb{R}^n$ of dimension p and q, respectively. Let $t = min(p, q)$. The principal angles between subspaces $\mathcal{Y}$ and $\mathcal{Z}$ are given by a sequence of t angles, $\boldsymbol{\Theta}(\mathcal{Y}, \mathcal{Z}) = [\theta_1, \ldots, \theta_j, \ldots, \theta_t]$, where $0 \leq \theta_1 \leq \ldots \leq \theta_t \leq \pi/2$. The angle θ_j is defined by*

$$\theta_j = \max_{z \in \mathcal{Z}} \max_{y \in \mathcal{Y}} \arccos\left(|y^T z|\right), \tag{33}$$

subject to $||y|| = ||z|| = 1$, $y_i^T y = 0$, $z_i^T z = 0$, for $i = 1, 2, \ldots, j-1$ [38].

The principal sines $\sin(\theta_j)'s$ of the angles can be computed using singular values as follows.

Theorem 1. *Let the columns of matrices $Y \in \mathbb{R}^{n \times p}$ and $Z \in \mathbb{R}^{n \times q}$ be orthonormal bases for subspaces $\mathcal{Y}$ and $\mathcal{Z}$, respectively. Let $\begin{bmatrix} Y & Y^\perp \end{bmatrix}$ be a unitary matrix such that the columns of $Y^\perp$ span the subspace orthogonal to $\mathcal{Y}$. Also, let the singular values of $(Y^\perp)^T Z$ be given by the elements of the diagonal matrix*

$$\Xi = diag(\sigma_1, \ldots, \sigma_t), \tag{34}$$

where $\sigma_1 \geq \ldots \geq \sigma_j \geq \ldots \geq \sigma_t$. Then, the principal sine $\sin(\theta_{t+1-j}) = \sigma_j$ [40], [41].

Thus, the principal sines between subspaces $\mathcal{Y}$ and $\mathcal{Z}$ are given by the singular values of $(Y^\perp)^T Z$. The principal sines can be used to define the difference between two subspaces as follows.

Definition 2. *Let $\mathcal{Y}$ and $\mathcal{Z}$ be two subspaces of $\mathbb{R}^n$. Let the diagonal matrix Ξ contain the singular values of $(Y^\perp)^T Z$ as in Theorem 1. The measure of difference between two subspaces $\mathcal{Y}$ and $\mathcal{Z}$ is defined by* [34]

$$\sin \boldsymbol{\Theta}(\mathcal{Y}, \mathcal{Z}) = \Xi. \tag{35}$$

Let the squared Frobenius norm of a matrix be denoted by $\| \, . \, \|_F^2$, which is given by the sum of squares of its singular values. Then, using (34) and (35), we get

$$\| \sin \boldsymbol{\Theta}(\mathcal{Y}, \mathcal{Z}) \|_F^2 = \| \Xi \|_F^2 = \sum_{j=1}^{r} \sigma_j^2 = \sum_{j=1}^{r} \sin^2\left(\theta_{t+1-j}\right). \tag{36}$$

Hence, (36) implies that the sum of squares of the principal sines between two subspaces $\mathcal{Y}$ and $\mathcal{Z}$ is given by $\| \sin \boldsymbol{\Theta}(\mathcal{Y}, \mathcal{Z}) \|_F^2$.

In order to bound the difference between the eigenvectors of two eigenspaces $\Psi(\mathbf{L}^r)$ and $\Psi(\mathbf{L}^{r*})$, the theory of *perturbation of invariant subspaces* [34] and Davis Kahan theorem [42] are used. The eigenvalues and eigenvectors of the full-rank eigenspace $\Psi(\mathbf{L}^r)$ are given by $\Gamma^r = diag\,(\gamma_1, \ldots, \gamma_r)$ and $\mathbf{Z}^r$, respectively, as in (14), where $\gamma_r \neq \gamma_{r+1}$, while those for the approximate eigenspace $\Psi(\mathbf{L}^{r*})$ are given by $\Pi^r = diag(\pi_1, \ldots, \pi_r)$ and $\mathbf{V}^r$, respectively, as in (30). The columns of $\mathbf{Z}^r$ span the full-rank subspace formed by the convex combination of full rank L_m's, while those of $\mathbf{V}^r$ span the approximate subspace formed by rank r approximation of L_m's. The difference between the subspaces spanned by the column vectors of $\mathbf{Z}^r$ and $\mathbf{V}^r$ is given by the following theorem.

Theorem 2. *For any unitarily invariant norm $\|\,.\,\|$, the following bound holds on the principal angles between the subspaces defined by $\mathcal{C}(\mathbf{Z}^r)$ and $\mathcal{C}(\mathbf{V}^r)$:*

$$\|\sin\Theta(\mathcal{C}(\mathbf{Z}^r), \mathcal{C}(\mathbf{V}^r))\| \leq \frac{\left\|\left(\sum_{m=1}^{M} \alpha_m L_m^{r\perp}\right)\mathbf{V}^r\right\|}{\left(\pi_r - \pi_{r+1} - \sum_{m=1}^{M} \alpha_m \lambda_{r+1}^m\right)}, \tag{37}$$

assuming $\pi_r > \pi_{r+1} + \sum_{m=1}^{M} \alpha_m \lambda_{r+1}^m$.

Proof. The proof is given in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2945574. □

The above theorem holds for any set of M symmetric positive semi-definite matrices and their convex combination.

Corollary 1. *Let $tr(B)$ denote the trace of matrix B. Then,*

$$\|\sin\Theta(\mathcal{C}(\mathbf{Z}^r), \mathcal{C}(\mathbf{V}^r))\|_F^2 \leq \frac{tr\left((\mathbf{V}^r)^T\left(\sum_{m=1}^{M} \alpha_m L_m^{r\perp}\right)^2\mathbf{V}^r\right)}{\left(\pi_r - \pi_{r+1} - \sum_{m=1}^{M} \alpha_m \lambda_{r+1}^m\right)}. \tag{38}$$

Proof. The proof is given in the supplementary material, available online. □

For a given value of r, $\|\sin\Theta(\mathcal{C}(\mathbf{Z}^r), \mathcal{C}(\mathbf{V}^r))\|_F^2$ measures the difference between the full-rank and approximate subspaces, in terms of the sum of squares of r principal sines between them. To make the differences comparable across different values of r, the mean squared principal sine is considered, which is given by

$$\begin{aligned}\Phi^r &= \frac{1}{r}\|\sin\Theta(\mathcal{C}(\mathbf{Z}^r), \mathcal{C}(\mathbf{V}^r))\|_F^2 \\ &\leq \frac{tr\left((\mathbf{V}^r)^T\left(\sum_{m=1}^{M} \alpha_m L_m^{r\perp}\right)^2\mathbf{V}^r\right)}{r\left(\pi_r - \pi_{r+1} - \sum_{m=1}^{M} \alpha_m \lambda_{r+1}^m\right)}.\end{aligned} \tag{39}$$

The matrix $L_m^{r\perp}$ denotes the approximation of L_m using eigenpairs $(r+1)$ to n. As r approaches the full rank n, the approximation of L_m using the remaining $(n-r)$ eigenpairs approaches to 0, that is, $L_m^{r\perp} \to 0$. Hence,

$$\lim_{r\to n} \sum_{m=1}^{M} \alpha_m L_m^{r\perp} = 0. \tag{40}$$

Taking limits in (39) and then substituting the value of (40) in the right hand side of (39), we get

$$\lim_{r\to n} \Phi^r = 0. \tag{41}$$

This implies that, as the rank r approaches to the full rank of the individual L_m, the difference between the full-rank and approximate subspace converges to 0, that is, the approximate subspace converges to the full-rank subspace.

The eigenvalues of $\mathbf{L}^r$ and $\mathbf{L}^{r*}$ are given by the elements of the diagonal matrices Γ and Π, respectively. The bound on the difference between the eigenvalues is given as follows.

Theorem 3. *The eigenvalues of $\mathbf{L}$ and $\mathbf{L}^{r*}$ satisfy the following bound:*

$$\sum_{j=1}^{n}(\gamma_j - \pi_j)^2 \leq \sum_{j=r+1}^{n}\sum_{m=1}^{M} \alpha_m (\lambda_j^m)^2. \tag{42}$$

Proof. The proof is given in the supplementary material, available online. □

Following analysis establishes that the difference between the eigenvalues of $\mathbf{L}$ and $\mathbf{L}^{r*}$ approaches to 0 as the rank r approaches to the full rank of $\mathbf{L}$. Let

$$\Delta^r = \frac{1}{n} tr\{(\Gamma - \Pi)^2\} = \frac{1}{n}\sum_{j=1}^{n}(\gamma_j - \pi_j)^2. \tag{43}$$

According to (42),

$$\Delta^r \leq \frac{1}{n}\sum_{j=r+1}^{n}\sum_{m=1}^{M} \alpha_m (\lambda_j^m)^2. \tag{44}$$

So, Δ^r bounds the squared sum of the difference between the eigenvalues of $\mathbf{L}$ and $\mathbf{L}^{r*}$. For $m = 1, \ldots, M$, each L_m is a positive semi-definite matrix with n eigenvalues $\lambda_1^m \geq \ldots \geq \lambda_n^m \geq 0$. As the value of r approaches n, the eigenvalue λ_r^m approaches the smallest eigenvalue λ_n^m. Moreover, as there are only n eigenvalues, the value of λ_j^m is 0 for any $j > r$ when r tends to n. Therefore,

$$\lim_{r\to n} \Delta^r = \lim_{r\to n} \frac{1}{n}\sum_{j=r+1}^{n}\sum_{m=1}^{M} \alpha_m (\lambda_j^m)^2 = 0. \tag{45}$$

The limits in (41) and (45) imply that as the approximation rank r approaches to the full rank, the approximate eigenspace $\Psi(\mathbf{L}^{r*})$ converges to the full-rank one $\Psi(\mathbf{L}^r)$, in terms of both eigenvectors and eigenvalues.

5 Experimental Results and Discussion

The performance of the proposed CoALa algorithm is compared with that of eight existing integrative clustering approaches, namely, cluster of cluster analysis (COCA) [9], LRAcluster [14], joint and individual variance explained (JIVE) [15], angle-based JIVE (A-JIVE) [16], iCluster [12], principal component analysis (PCA) on the concatenated data (PCA-con) [43], similarity network fusion (SNF) [44],

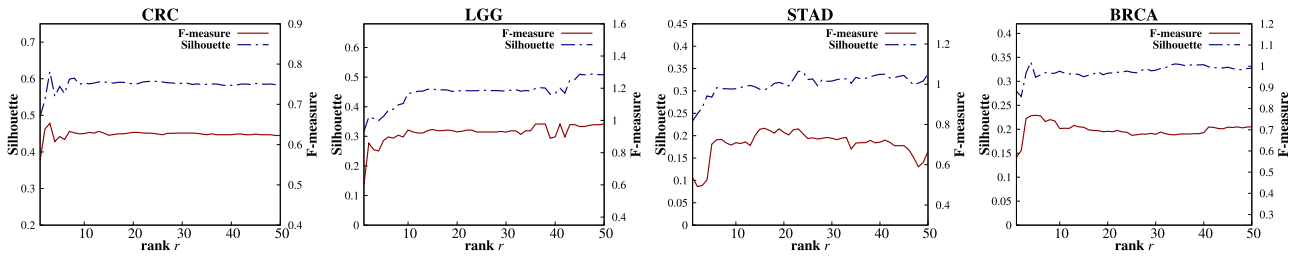


Fig. 1. Variation of Silhouette index and F-measure for different values of rank parameter r on omics data sets.

and normality based low rank subspace (termed as NormS) [45]. The performance of different algorithms is evaluated using five external cluster evaluation indices, namely, F-measure, purity, Rand index, Jaccard coefficient, and Dice coefficient, which compare the identified clusters with the clinically established cancer subtypes and the ground truth class information for the benchmark data sets. For the low-rank based approaches, where clustering is performed in a subspace, four internal cluster validity indices, namely, Silhouette, Dunn, Davies-Bouldin (DB), and Xie-Beni indices are used to evaluate the compactness and separability of the clusters in the extracted subspace.

5.1 Description of Data Sets

In this work, the clustering performance is extensively studied on four real-life cancer data sets, obtained from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/). Four types of cancers are considered here, namely, colorectal carcinoma (CRC), lower grade glioma (LGG), stomach adenocarcinoma (STAD), and breast adenocarcinoma (BRCA). The CRC has two subtypes: colon and rectum carcinoma, depending on their site of origin. For the other three cancers, TCGA research network has identified three subtypes in LGG [46], and four subtypes in STAD [47] and BRCA [48] by comprehensive integrated analysis. The CRC, LGG, STAD, and BRCA data sets have 464, 267, 242, and 398 samples, respectively. For each of these data sets, four different omic modalities are considered, namely, DNA methylation (mDNA), gene expression (RNA), microRNA expression (miRNA), and reverse phase protein array expression (RPPA). The pairwise similarity $w_m(i,j)$ between samples x_i and x_j of the modality X_m is computed using the Gaussian similarity kernel

$$w_m(i,j) = \exp\left\{-\frac{\rho_m^2(x_i,x_j)}{2\sigma_m^2}\right\}, \tag{46}$$

where $\rho_m(x_i,x_j)$ denotes the euclidean distance between samples x_i and x_j in X_m and σ_m is the standard deviation of the Gaussian kernel. The value of σ_m is empirically set to be half of the maximum pairwise distance between any two points of the modality. Choice of this similarity function results in a completely connected graph for each modality.

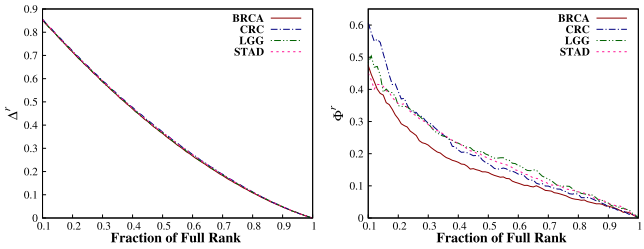
Four other data sets from different application domains like community networks and general images are also employed in this study to compare the clustering performance of the proposed and existing algorithms. Among them, Football, Politics-uk, and Rugby (http://mlg.ucd.ie/aggregation/) are three benchmark multimodal Twitter data sets, all of which consist of a heterogeneous collection of nine network and content-based modalities, namely, follows, followed-by, mentions, mentioned-by, retweets, retweeted-by, lists500, tweets500, and listmerged500. The cosine similarity is used to compute the pairwise similarities between the users in these three Twitter data set. The Digits data set (https://archive.ics.uci.edu/ml/datasets/Multiple+Features) consists of features of handwritten numerals ('0'–'9') extracted from 200 binary images per class. The Gaussian similarity kernel described above is used to construct the similarity matrices for the Digits data set. A brief description of five omics and four benchmark data sets, pre-processing steps, definitions of quantitative indices used, and some important results are described in detail in the supplementary material, available online.

5.2 Optimum Value of Rank

Similar to the existing spectral clustering algorithm [28], [31], the proposed CoALa algorithm also performs k-means clustering on k eigenvectors of the final eigenspace. Although clustering is performed in a k-dimensional subspace, the proposed algorithm stores $r \geq k$ eigenpairs from the individual Laplacians at the initial stage to allow extra information from each Laplacian. To find out the optimal value of rank r, the Silhouette index [37] is used. It lies between $[-1,1]$ and a higher value implies better clustering. In order to choose the rank parameter, the value of r is varied from 1 to 50 and for each value of r, the Silhouette index $\mathcal{S}(r)$ is evaluated for clustering on the k largest eigenvectors of the final eigenspace. The optimal value of r, that is r^*, is obtained using the following relation:

$$r^* = \arg\max_r \{\mathcal{S}(r)\}. \tag{47}$$

The variation of both Silhouette index and F-measure with respect to the rank r is shown in Fig. 1 for different omics data sets. The plots in Fig. 1 show that the values of Silhouette index and F-measure vary in a similar fashion. The Silhouette index is an internal cluster validity measure computed based on the generated clusters, while F-measure is an external index which compares the generated clusters with the ground truth class information. Since these two indices are found to vary similarly, the optimum value of Silhouette index would also produce the optimum value of F-measure for the same parameter configuration. Using this criterion, the optimal values of rank for CRC, LGG, STAD, and BRCA data sets are 3, 48, 23, and 4, respectively. It is also observed that for BRCA and CRC data sets, the F-measure corresponding to r^* coincides with the best value of F-measure obtained over different values of rank r. The similarly varying curves of Silhouette and F-measure in Fig. 1 justify the use of Silhouette index to find out the optimal rank.



(a) Difference in Eigenvalues (b) Difference in Eigenvectors

Fig. 2. Variation of difference between full-rank and approximate eigenspaces with respect to rank r.

5.3 Difference Between Eigenspaces

The proposed method constructs an eigenspace from low-rank approximations of individual graph Laplacians. This eigenspace is an approximation of the full-rank eigenspace which considers the complete or full rank information of all the Laplacians. As defined in Section 4, for a given rank r, the difference between the full-rank and approximate eigenspaces, in terms of its eigenvalues and eigenvectors, is given by Δ^r and Φ^r, respectively. Here, the variation in the difference between these two eigenspaces is observed with the increase in rank r. For each omic data set, Δ^r and Φ^r are computed for different fractions of the full rank of that data set. The variation in the values of Δ^r and Φ^r, with the increase in rank r, is shown in Figs. 2a and 2b, respectively, for different data sets. Fig. 2a shows that the difference between eigenvalues of the two eigenspaces monotonically decreases to 0 with the increase in rank, for all the data sets. Fig. 2b, on the other hand, shows that the difference between the subspaces, spanned by the eigenvectors of the two eigenspaces, also converges to 0 as the value of rank r approaches the full rank of the data set. However, the change in variation in case of eigenvectors is not monotonically decreasing as in the case of eigenvalues. For some of the smaller values of rank r, the difference also increases between two consecutive values. This is due to the fact that for a given value of r, there can be infinitely possible rank r subspaces of an n dimensional vector space. For small values of r, the rank r subspaces of individual modalities can be very different from each other due to the large number of possibilities. Consequently, the approximate subspace constructed from these subspaces tends to vary a lot from the full-rank subspace. Hence, the variation in the difference between the full-rank and approximate sets of eigenvectors fluctuates for small values of rank r. However, as r approaches the full-rank, the number of possible subspaces reduces and the difference between the eigenvectors monotonically decreases to 0.

5.4 Effectiveness of Proposed CoALa Algorithm

This subsection illustrates the significance of different aspects of the proposed algorithm such as integration of multiple modalities over individual ones, use of approximate Laplacians as opposed to full-rank ones, choice of the convex combination α, and so on, for omics data sets.

5.4.1 Importance of Data Integration

The proposed CoALa algorithm performs clustering on the k largest eigenvectors of the approximate eigenspace constructed by integrating multiple low-rank Laplacians. To establish the importance of this integration, the performance of the proposed algorithm is compared with the spectral clustering on the individual modalities in Table 1. The results in Table 1 show that the proposed algorithm performs better than all four individual modalities for CRC, LGG, and STAD data sets, in terms of all external indices, except for the purity measure on the CRC data set. The performance is equal for the purity measure on the CRC data set across all modalities. Since the highest value of the F-measure on CRC data set is obtained for the proposed algorithm, it identifies the smaller cluster better than all the individual Laplacians. For the BRCA data set, RNA outperforms the proposed algorithm, albeit by a very small margin. Among the individual modalities, mDNA gives the best performance for CRC, LGG, and STAD data sets. For LGG and STAD data sets, the performance of the proposed CoALa algorithm is significantly higher than that of their best modality, mDNA. The scatter plots of the first two dimensions for the best modality, mDNA, and the proposed CoALa algorithm are given in Figs. 3 and 4 for LGG and STAD data sets, respectively. The objects in Figs. 3 and 4 are colored according to the previously established TCGA subtypes of LGG [46] and STAD [47]. For the LGG data set, Fig. 3a shows that in the two-dimensional Laplacian subspace of mDNA, one of the subtypes is compact and well-separated while the other two intermingled amongst each other. On the other hand, Fig. 3j for LGG shows that in the proposed subspace all the three clusters are compact and separated from each other. For

TABLE 1
Comparative Performance Analysis of Individual Modalities and Proposed Approach on Omics Data

Index		mDNA	RNA	miRNA	RPPA	CoALa		mDNA	RNA	miRNA	RPPA	CoALa
F-measure	CRC	0.5849894	0.5397796	0.5673758	0.5741394	**0.6529565**	LGG	0.8269248	0.5875701	0.4717221	0.4326018	**0.9737835**
Purity	CRC	**0.7370690**	**0.7370690**	**0.7370690**	**0.7370690**	**0.7370690**	LGG	0.8352060	0.5917603	0.5318352	0.5280899	**0.9737828**
Rand	CRC	0.4989573	0.4991528	0.5022809	0.5007448	**0.5382531**	LGG	0.7861508	0.6149925	0.5593760	0.5476050	**0.9622089**
Jaccard	CRC	0.3925508	0.3789509	0.3818306	0.3853947	**0.4315561**	LGG	0.5814133	0.3235367	0.2476680	0.2328447	**0.9056723**
Dice	CRC	0.5637867	0.5496220	0.5526446	0.5563681	**0.6029189**	LGG	0.7353085	0.4888972	0.3970095	0.3777356	**0.9505016**
F-measure	STAD	0.5469686	0.4781377	0.3998266	0.4469459	**0.7778227**	BRCA	0.5982526	**0.7690661**	0.5105008	0.5630781	0.7660191
Purity	STAD	0.5867769	0.5495868	0.4917355	0.4917355	**0.7685950**	BRCA	0.6532663	**0.7688442**	0.5703518	0.5879397	0.7613065
Rand	STAD	0.6509722	0.6239155	0.5989164	0.5883543	**0.7661946**	BRCA	0.7193018	**0.7995519**	0.6455071	0.6689493	0.7922357
Jaccard	STAD	0.2869053	0.2234653	0.1994524	0.2076045	**0.4535983**	BRCA	0.3318872	**0.4857607**	0.2672039	0.3132549	0.4612885
Dice	STAD	0.4458841	0.3652989	0.3325725	0.3438286	**0.6241041**	BRCA	0.4983713	**0.6538882**	0.4217221	0.4770664	0.6313449

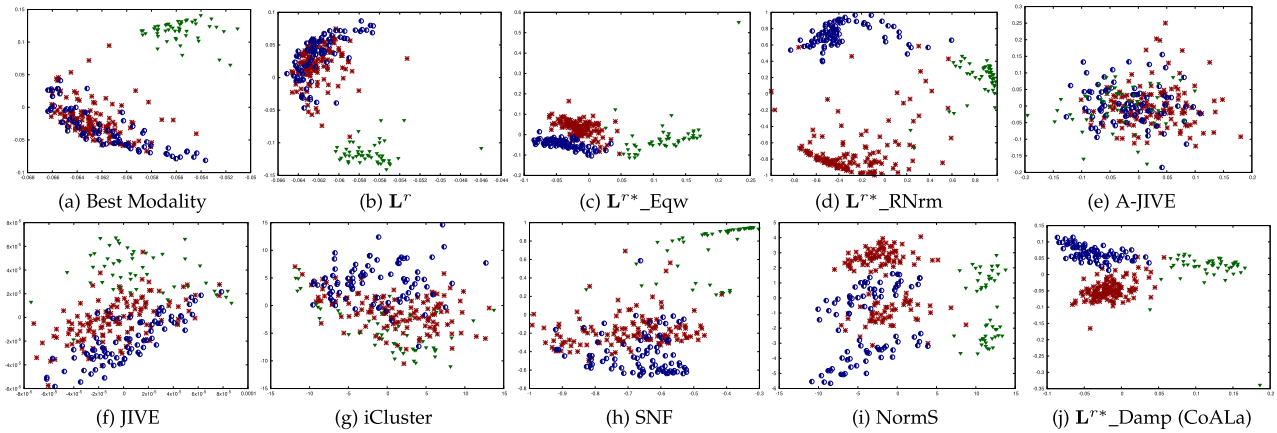


Fig. 3. Scatter plots using first two components of different low-rank based approaches on LGG data set.

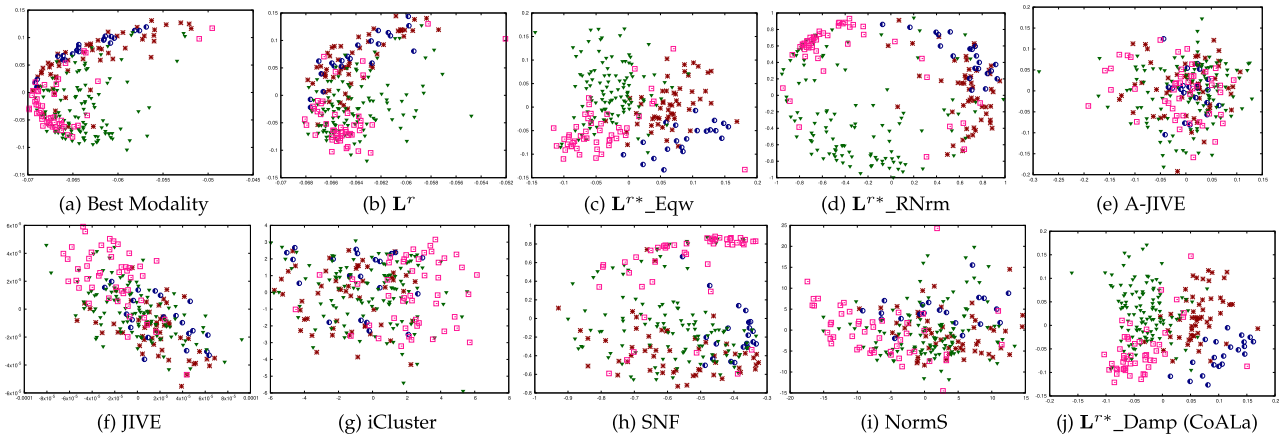


Fig. 4. Scatter plots using first two components of different low-rank based approaches on STAD data set.

STAD, Fig. 4a shows that a major part of the two-dimensional subspace consists of points randomly scattered from all the four clusters. However, Fig. 4j shows that although the clusters lack well separability, the proposed subspace can be partitioned into regions where most of the data points belong to a single cluster. The scatter plots for the remaining data sets are provided in the supplementary material, available online. The distinct omic modalities together cover a wide spectrum of biological information and the results in Table 1 show that integration of multiple modalities leads to better identification of the disease subtypes compared to unimodal analysis.

5.4.2 Importance of the Choice of Convex Combination

In order to establish the effectiveness of the proposed weighting factor (termed as $\mathbf{L}^{r*}$_Damp) described in Section 3.5, the clustering performance of the resulting subspace obtained using $\mathbf{L}^{r*}$_Damp is compared with that of the one where all the modalities are equally weighted (termed as $\mathbf{L}^{r*}$_Eqw). The damping factor β in (32) is empirically set to 1.25 for all data sets. The scatter plots for the first two components of $\mathbf{L}^{r*}$_Eqw and $\mathbf{L}^{r*}$_Damp (CoALa) subspaces are given in Figs. 3 and 4, for LGG and STAD data sets, respectively. For LGG, Fig. 3c for $\mathbf{L}^{r*}$_Eqw shows that two of the three clusters are highly compact, however, they also lack inter-cluster separability. In case of the proposed $\mathbf{L}^{r*}$_Damp subspace, in Fig. 3j, these two clusters have lower compactness but are well-separated from each other. For STAD, scatter plots for $\mathbf{L}^{r*}$_Eqw and $\mathbf{L}^{r*}$_Damp (CoALa) in Figs. 4c and 4j, respectively, are of similar nature, although $\mathbf{L}^{r*}$_Eqw shows slightly better inter-cluster separability compared to $\mathbf{L}^{r*}$_Damp. The quantitative results for this comparison are reported in Table 2, which show that for CRC, LGG, and BRCA data sets, the damping strategy $\mathbf{L}^{r*}$_Damp performs better than $\mathbf{L}^{r*}$_Eqw, in terms of all external indices. Only for the STAD data set, weighting all the modalities equally gives slightly better performance. This is also evident from the increased inter-cluster separability in Fig. 4c compared to Fig. 4j. However, the results in Table 2 show that assigning maximum weightage to the most relevant modality and gradually damping it by a factor β, based on its relevance, preserves better cluster information in majority of the cases.

5.4.3 Importance of Noise-Free Approximation

The proposed eigenspace is an approximate one, as it is constructed from de-noised approximations of the individual eigenspaces. This approximate eigenspace is expected to preserve better cluster structure compared to the full-rank eigenspace constructed from the complete set of eigenpairs of the individual Laplacians. In order to establish this, the performance of clustering on the k largest eigenvectors of the full-rank eigenspace $\mathbf{L}^r$ is compared with that of the approximate

TABLE 2
Comparative Performance Analysis of Equally and Damped Weighted Combination on Omics Data

Index		$\mathbf{L}^{r*}$_Eqw	$\mathbf{L}^{r*}$_Damp		$\mathbf{L}^{r*}$_Eqw	$\mathbf{L}^{r*}$_Damp		$\mathbf{L}^{r*}$_Eqw	$\mathbf{L}^{r*}$_Damp		$\mathbf{L}^{r*}$_Eqw	$\mathbf{L}^{r*}$_Damp
F-measure	CRC	0.6309431	**0.6529565**	LGG	0.9625844	**0.9737835**	STAD	0.7788198	**0.7778227**	BRCA	0.6834253	**0.7660191**
Purity		**0.7370690**	**0.7370690**		0.9625468	**0.9737828**		**0.7727273**	0.7685950		0.6783920	**0.7613065**
Rand		0.5260669	**0.5382531**		0.9437921	**0.9622089**		**0.7703782**	0.7661946		0.7523132	**0.7922357**
Jaccard		0.4194417	**0.4315561**		0.8619640	**0.9056723**		**0.4579454**	0.4535983		0.3986848	**0.4612885**
Dice		0.5909953	**0.6029189**		0.9258654	**0.9505016**		**0.6282066**	0.6241041		0.5700852	**0.6313449**

TABLE 3
Comparative Performance Analysis of Full-Rank and Approximate Subspaces of Omics Data

Index		$\mathbf{L}^{r}$	CoALa ($\mathbf{L}^{r*}$)		$\mathbf{L}^{r}$	CoALa ($\mathbf{L}^{r*}$)		$\mathbf{L}^{r}$	CoALa ($\mathbf{L}^{r*}$)		$\mathbf{L}^{r}$	CoALa ($\mathbf{L}^{r*}$)
F-measure	CRC	0.6052757	**0.6529565**	LGG	0.6577440	**0.9737835**	STAD	0.6158419	**0.7778227**	BRCA	0.6197007	**0.7660191**
Purity		**0.7370690**	**0.7370690**		0.6441948	**0.9737828**		0.6157025	**0.768595**		0.7185930	**0.7613065**
Rand		0.5007448	**0.5382531**		0.6524739	**0.9622089**		0.6706560	**0.7661946**		0.7403390	**0.7922357**
Jaccard		0.4018471	**0.4315561**		0.4053390	**0.9056723**		0.2966164	**0.4535983**		0.3586770	**0.4612885**
Dice		0.5733108	**0.6029189**		0.5768558	**0.9505016**		0.4575237	**0.6241041**		0.5279798	**0.6313449**

TABLE 4
Effect of Row-normalization on Different Subspaces on Omics Data

Index		$\mathbf{L}^{r*}$_RNrm	CoALa		$\mathbf{L}^{r*}$_RNrm	CoALa		$\mathbf{L}^{r*}$_RNrm	CoALa		$\mathbf{L}^{r*}$_RNrm	CoALa
F-measure	CRC	0.6169586	**0.6529565**	LGG	0.9010565	**0.9737835**	STAD	0.7389739	**0.7778227**	BRCA	0.6946324	**0.7660191**
Purity		**0.7370690**	**0.7370690**		0.8951311	**0.9737828**		0.7355372	**0.7685950**		0.6859296	**0.7613065**
Rand		0.5186192	**0.5382531**		0.8771367	**0.9622089**		0.7474024	**0.7661946**		0.7588193	**0.7922357**
Jaccard		0.4084971	**0.4315561**		0.7134883	**0.9056723**		0.4060156	**0.4535983**		0.3958974	**0.4612885**
Dice		0.5800468	**0.6029189**		0.8327904	**0.9505016**		0.5775407	**0.6241041**		0.5672299	**0.6313449**

eigenspace $\mathbf{L}^{r*}$ (CoALa) in Table 3. From the results of Table 3, it can be observed that the proposed CoALa algorithm outperforms the full-rank subspace $\mathbf{L}^{r}$ for all the data sets. The performance is significantly better for BRCA, LGG, and STAD data sets. The full-rank information of individual Laplacians in $\mathbf{L}^{r}$ inherently contains the noisy information of the $(n-r)$ smallest eigenvectors of each Laplacian. However, in the proposed algorithm, each individual Laplacian is truncated at rank r, to contain mostly the cluster discriminatory information, where $r \ll n$. So, the approximate eigenspace automatically eliminates the noise present in the $(n-r)$ remaining eigenvectors. The results of Table 3 show that this truncated de-noised Laplacians preserve better cluster structure in the resulting eigenspace compared to the full-rank one. The scatter plots for the full-rank subspaces of LGG and STAD data sets are given in Figs. 3b and 4b, respectively. For LGG, Fig. 3b shows that only one cluster is well-separated. On the other hand, data points from the other two clusters of LGG and all the four clusters of STAD in Fig. 4b are cluttered amongst each other exhibiting poor separability. The optimal rank, r^{*}, for LGG and STAD data sets are 48, and 39, respectively, while their full-ranks are 267 and 242, respectively. The scatter plots in full-rank approximation in Figs. 3j and 4j show that filtering out the noise in the remaining 219 and 203 eigen-pairs of the individual Laplacians preserves significantly better cluster structure for these data sets.

5.4.4 Advantage of Averting Row-normalization

In normalized spectral clustering (Algorithm 1), row-normalization tends to shift the objects in the projected subspace in such a way that they cluster tightly around an orthogonal basis. This is primarily justified when the objects lie close to the ideal case where the clusters are infinitely apart [31]. However, row-normalization may not necessarily give better performance on real-life data sets. The two-dimensional scatter plots for the row-normalized subspaces of LGG and STAD data sets are given in Figs. 3d and 4d, respectively. For both data sets, as expected, row-normalization pushes objects from different clusters further away from the origin in different directions of the subspace, which increases the inter-cluster separability. However, points lying in the boundaries of different clusters are not necessarily pushed away and are projected around the origin of the subspaces, which in turn reduces the compactness of the clusters. When the number of boundary points is relatively large, row-normalization tends to give degraded performance. To study this quantitatively, the clustering performance of the row-normalized subspace (termed as $\mathbf{L}^{r*}$_RNrm) is compared with that of not normalized one in Table 4. The results reported in Table 4 show that for all four data sets, the proposed subspace performs better than its row-normalized counterpart $\mathbf{L}^{r*}$_RNrm.

5.5 Performance Analysis on Omics Data

The performance of the proposed algorithm is compared with that of the existing ones, in Table 5, in terms of several external and internal cluster evaluation indices, and execution efficiency. The COCA is a consensus clustering based approach, while the seven other existing algorithms are subspace based approaches for which the optimal rank of the

TABLE 5
Comparative Performance Analysis of Proposed and Existing Approaches on Omics Data

Data Set		Measure	COCA	LRAcluster	JIVE (PERM)	A-JIVE	iCluster	PCA-con	SNF	NormS	CoALa
CRC $[n = 464; k = 2; M = 4]$		Subspace Rank	-	3	16	32	1	2	2	16	2
	External	F-measure	0.5586055	0.5410661	0.6210774	0.6206032	0.6298050	0.5641984	0.6178576	0.6345375	**0.6529565**
		Purity	**0.7370690**	**0.7370690**	**0.7370690**	**0.7370690**	**0.7370690**	**0.7370690**	**0.7370690**	**0.7370690**	**0.7370690**
		Rand	0.5010706	0.4992552	0.5203694	0.5203694	0.5260669	0.5016106	0.5186192	0.5281150	**0.5382531**
		Jaccard	0.3815861	0.3791352	0.4100133	0.4108634	0.4213948	0.3804537	0.4068687	0.4221878	**0.4315561**
		Dice	0.5523884	0.5498159	0.5815736	0.5824283	0.5929314	0.5512009	0.5784032	0.5937159	**0.6029189**
	Internal	Silhouette	-	0.4337712	0.4199826	0.5016133	0.6229586	0.3417350	**0.7834208**	0.3640602	0.3483722
		Dunn	-	0.0160840	0.0120740	0.0043986	**0.3317529**	0.0190144	0.0549104	0.0185685	0.0179209
		DB	-	0.8751325	0.8821177	0.6872426	0.5770987	1.1650270	**0.2980235**	1.0995640	1.1021510
		Xie-Beni	-	202.80470	348.50660	2314.1590	**0.3629792**	155.15390	17.069770	116.69010	115.98920
		Time (in sec)	21.02	104.12	3098.75	946.18	337.51	**2.62**	9.66	1.45	32.77
LGG $[n = 267; k = 3; M = 4]$		Subspace Rank	-	2	8	48	2	3	3	14	3
	External	F-measure	0.6619834	0.5137382	0.5757978	0.7326381	0.5187438	0.6574834	0.8720595	0.7916535	**0.9737835**
		Purity	0.6602995	0.5280899	0.5730337	0.7314606	0.5355805	0.6666667	0.8689139	0.7940075	**0.9737828**
		Rand	0.6462251	0.5831714	0.6056715	0.6978063	0.5821858	0.6616823	0.8268142	0.7465292	**0.9622089**
		Jaccard	0.3738023	0.2746607	0.3132418	0.4338314	0.2718037	0.4161442	0.6303420	0.5049772	**0.9056723**
		Dice	0.5441832	0.4309549	0.4770512	0.6048921	0.4274302	0.5877145	0.7732635	0.6710762	**0.9505016**
	Internal	Silhouette	-	0.3921144	0.4138221	0.3375023	0.3952103	0.4624043	0.4441981	0.4305583	**0.6273401**
		Dunn	-	0.0344110	**0.0355064**	0.0241153	0.0252834	0.0322859	0.0149314	0.0218683	0.0287595
		DB	-	0.8593495	0.8684623	0.9444459	0.9330074	0.7439401	0.7388554	0.8441603	**0.4905286**
		Xie-Beni	-	43.233820	51.054660	87.842080	93.144060	58.96720	318.54730	175.06670	**12.563470**
		Time (in sec)	11.33	37.71	665.82	364.43	3230.52	1.08	1.33	**0.96**	17.02
STAD $[n = 242; k = 4; M = 4]$		Subspace Rank	-	1	8	196	3	4	4	9	4
	External	F-measure	0.4440130	0.4746753	0.4487487	0.4618001	0.3832114	0.6959782	0.6333622	0.5770884	**0.7778227**
		Purity	0.5132231	0.5619835	0.5165289	0.5330578	0.4917355	0.6900826	0.6363636	0.5950413	**0.7685950**
		Rand	0.5959055	0.6122218	0.5981619	0.6159288	0.5855423	0.7110524	0.6945235	0.6435993	**0.7661946**
		Jaccard	0.2186385	0.2247892	0.2158726	0.2240769	0.1930832	0.3557119	0.3150853	0.2617559	**0.4535983**
		Dice	0.3587118	0.3670659	0.3550908	0.3651238	0.3236710	0.5247603	0.4791862	0.4149074	**0.6241041**
	Internal	Silhouette	-	0.4015128	0.3618677	0.3365825	0.3790058	0.3862858	**0.4477905**	0.3395181	0.4102003
		Dunn	-	0.0304117	0.0257650	0.0203049	0.0357959	0.0182291	**0.0596324**	0.0181344	0.0325467
		DB	-	0.7928001	0.9526717	0.9617136	0.9584001	0.8355266	**0.7872797**	0.9157146	0.8490579
		Xie-Beni	-	40.097030	84.992880	101.3060	54.2869300	227.84070	**19.297210**	181.9440	58.722830
		Time (in sec)	25.92	49.36	734.70	302.98	1138.88	1.02	1.14	**0.80**	13.79
BRCA $[n = 398; k = 4; M = 4]$		Subspace Rank	-	2	12	64	3	4	4	11	4
	External	F-measure	0.7466189	0.7101385	0.6889363	0.6522419	0.7658865	0.7601317	0.7006154	**0.7699789**	0.7660191
		Purity	0.7449748	0.7110553	0.6859296	0.6756281	0.7638191	0.7587940	0.6834171	**0.7688442**	0.7613065
		Rand	0.7913825	0.7521740	0.7464906	0.7257420	0.7842867	0.7984380	0.7475919	**0.7999063**	0.7922357
		Jaccard	0.4654364	0.4075049	0.3827282	0.3666530	0.4554054	0.4746982	0.3775440	**0.4796577**	0.4612885
		Dice	0.6351523	0.5790458	0.5535842	0.5355660	0.6258124	0.6437903	0.5481408	**0.6483360**	0.6313449
	Internal	Silhouette	-	0.4300455	0.4429883	0.3148863	0.4400869	0.4232505	**0.5005988**	0.4218991	0.4478377
		Dunn	-	**0.0369472**	0.0134063	0.0142913	0.0258263	0.0241363	0.0189055	0.0090550	0.0253506
		DB	-	0.8211325	0.7463430	0.9765342	0.7819524	0.8269517	**0.6814998**	0.8069696	0.7873740
		Xie-Beni	-	**43.223840**	277.15980	187.56970	77.708790	86.81890	112.0742	504.69590	81.048340
		Time (in sec)	35.36	88.32	866.00	686.85	511.87	**0.93**	1.91	1.47	14.36

clustering subspace is reported in Table 5. The optimal ranks are selected using the selection criteria suggested by the authors for the respective approaches. The results in Table 5 show that the proposed algorithm performs better than all the existing approaches for CRC, LGG, and STAD data sets in terms of the external indices, except for the purity measure on the CRC data set. However, F-measure and other external indices indicate that the proposed algorithm identifies the smaller sized cluster better than the existing ones. For BRCA data set, NormS has the highest clustering performance in terms of external indices, while the proposed algorithm achieves second best performance. Among the existing algorithms, the second best performance for CRC, LGG, and STAD data sets is obtained by NormS, SNF, and PCA-con, respectively. The iCluster algorithm has comparable performance for BRCA and CRC data sets, however, its degraded performance in the remaining data sets is due to the poor selection of its optimal lasso penalty parameter from the high-dimensional parameter space.

Due to the heterogeneous nature of the individual modalities, LRAcluster models each modality using a separate probability distribution having its own set of parameters. The proposed algorithm handles data heterogeneity by considering separate similarity matrices for separate modalities. Moreover, the modalities are integrated using their shifted Laplacians whose elements always lie in $[0, 2]$ as opposed to the raw data format. So, the difference in unit and scale of the individual modalities does not affect the final eigenspace. Similar to the proposed algorithm, the SNF approach also uses spectral clustering on a unified similarity graph to identify the clusters. However, in terms of the external indices, the proposed algorithm outperforms SNF on all data sets. In SNF, the unified graph is iteratively made similar to the individual graphs. This can often lead to propagation of unwanted information from noisy graphs into the final unified one. On the other hand, the proposed algorithm amplifies the effect of the most relevant graph, as well as dampens the effect of the irrelevant ones in the convex combination. Moreover, truncation of individual Laplacians at rank $r \ll n$ helps in propagating mostly cluster discriminatory information into the final subspace and automatically filters out the noise. These two aspects of the proposed CoALa algorithm are primarily responsible for its significantly better performance, especially for the LGG and STAD data sets.

Different low-rank based approaches extract subspaces of different ranks. Table 5 shows that the ranks vary from 1 to as high as 64. The comparison of cluster compactness and

TABLE 6
Comparative Performance Analysis of Proposed and Existing Approaches on Benchmark Data Sets

Measure			Best View	$\mathbf{L}^r$	SNF	CoALa		Best View	$\mathbf{L}^r$	SNF	CoALa
Subspace Rank		Football $[n = 248; k = 20; M = 9]$	20	20	20	20	Politics-uk $[n = 419; k = 5; M = 9]$	5	5	5	5
External	F-measure		0.7747023	0.6616297	0.8431825	**0.8683491**		0.9175316	0.8192186	0.9701235	**0.9736129**
	Purity		0.7282258	0.6572580	0.8266129	**0.8584677**		0.9713604	0.8591885	0.9761337	**0.9785203**
	Rand		0.9472965	0.8843737	0.9735862	**0.9739682**		0.9196880	0.8603076	0.9814665	**0.9826084**
	Jaccard		0.3963918	0.2612106	**0.6125478**	0.6005824		0.8019766	0.7257236	0.9529074	**0.9559279**
	Dice		0.5667814	0.4136485	**0.7597267**	0.7504383		0.8901077	0.8410659	0.9758859	**0.9774674**
Internal	Silhouette		**0.5565601**	0.4392812	0.4750064	0.5170209		**0.7877163**	0.5531584	0.7599383	0.6165161
	Dunn		0.0122200	0.0304905	0.0496361	**0.05060948**		0.0691656	0.0082616	0.0121941	**0.02166768**
	DB		**0.4087806**	0.5388078	0.6463104	0.5318746		0.5042173	**0.4124179**	0.4971371	0.6299340
	Xie-Beni		181.35320	36.629720	16.878340	**15.47080**		**4.1253610**	544.13860	66.892230	68.551380
Time (in sec)			**0.68**	1.13	1.05	1.34		**0.95**	1.86	3.83	3.68
Subspace Rank		Rugby $[n = 854; k = 15; M = 9]$	15	15	15	15	Digits $[n = 2000; k = 10; M = 6]$	17	17	17	17
External	F-measure		0.7426962	0.6845209	0.7778990	**0.8349647**		0.7209662	0.8481826	**0.8932872**	0.8839913
	Purity		0.7796253	0.6803279	0.8454333	**0.8606557**		0.7100000	0.8500000	**0.8835000**	**0.8835000**
	Rand		0.8672685	0.8578210	0.8818113	**0.9067597**		0.9173923	0.9503602	**0.9715983**	0.9576618
	Jaccard		0.4447761	0.4883211	0.4446208	**0.5982183**		0.4163257	0.6055477	**0.7534116**	0.6502019
	Dice		0.6155136	0.6562039	0.6155536	**0.7486065**		0.5878948	0.7543192	**0.8593665**	0.7880271
Internal	Silhouette		**0.5444214**	0.5195532	0.4713082	0.4123312		0.4860050	**0.5265748**	0.4452352	0.4269673
	Dunn		0.0012972	0.0085216	0.0051843	**0.0086649**		0.0050409	0.0064673	0.0031041	**0.0071841**
	DB		**0.4727219**	0.4603219	0.5856659	0.7474256		0.5722576	**0.5331665**	0.8063785	0.7470644
	Xie-Beni		780.66640	**212.29020**	827.27610	328.6280		1275.5800	830.76950	1166.0330	**659.67560**
Time (in sec)			**4.77**	7.21	22.94	27.42		**80.71**	135.65	189.03	154.57

separability in these subspaces of varying dimensions is not reasonable. So, the goodness of clustering is evaluated using internal cluster validity indices by performing k-means clustering on the first two dimensions of each subspace. This makes the internal evaluation results comparable and also easy to visualize. Four internal cluster evaluation measures, namely, Silhouette and Dunn, which are maximization based indices, and Davies-Bouldin (DB) and Xie-Beni, which are minimization based, are used. The internal cluster evaluation results in Table 5 show that the proposed algorithm has best performance for Silhouette, DB, and Xie-Beni indices for LGG data set and the second best for Silhouette and Dunn indices for BRCA data set. The SNF has best performance for two or more internal indices for CRC, STAD, and, BRCA data sets. This implies that on these three data sets, the cluster structure reflected in the first two dimensions of SNF more are compact and well-separated compared to the proposed and other existing algorithms. The scatter plots for the first two dimensions of some low-rank based approaches are given in Figs. 3 and 4, respectively, for LGG and STAD data sets. The data points are labeled in different colors based on the previously established TCGA subtypes. Although SNF has the best performance for all the internal indices for STAD data set, the scatter plot of SNF for LGG, in Fig. 3h, shows that the compact and well-separated clusters do not necessarily conform with the clinically established TCGA labellings. In brief, out of 20 cases, the proposed CoALa algorithm ranks among the top three in 10 cases. On the other hand, the results of external evaluation indices in both Tables 1 and 5 show that the clusters identified by the proposed algorithm have the closest resemblance with the clinically established TCGA subtypes of each cancer data set.

The execution times reported in Table 5 show that the proposed CoALa algorithm is computationally much faster than the consensus based COCA approach and other low-rank approaches like LRAcluster, JIVE, A-JIVE, and iCluster. However, PCA-con, SNF, and NormS have lower execution time compared to the proposed algorithm across all the data sets. For model fitting, iCluster uses expectation maximization algorithm, while JIVE uses alternate optimization. These iterative algorithms have slow convergence on the high-dimensional multimodal data sets. This leads to huge execution time and poor scalability of these algorithms as seen in Table 5. PCA-con achieves the lowest execution time on CRC and STAD data sets, as it performs SVD on the concatenated data only once. On the other hand, NormS achieves the same on LGG and STAD data sets. NormS achieves this computational advantage by simply concatenating relevant principal components from different modalities, at the cost of constructing a relatively much higher dimensional subspace. However, the external evaluation indices show that such naive concatenation in PCA-con and NormS often fails to capture the true cluster structure of the multimodal data.

5.6 Performance Analysis on Benchmark Data

Finally, the performance of different algorithms is studied on four benchmark multimodal data sets, namely, Football, Politics-uk, Rugby, and Digits. Among them, Football, Politics-uk, and Rugby are Twitter data sets whose most of the component modalities have graph based representation. However, apart from SNF, all other existing algorithms require feature based representations of the component modalities, so their performance could not be evaluated on Twitter data. The comparative performance of the best modality (in terms of external indices), the full-rank subspace $\mathbf{L}^r$, SNF, and the proposed CoALa algorithm are reported in Table 6. The convex combination $\boldsymbol{\alpha}$ and the optimal rank r^* are assigned as described previously in Sections 3.5 and 5.2, respectively. Supportive results on the benchmark data sets are provided in the supplementary material, available online.

The comparative results of Table 6 show that the proposed algorithm has the best performance in terms of all five external indices for all three Twitter data sets, namely, Football, Politics-uk, and Rugby. The SNF algorithm has the second best performance on these data sets and the best modality always outperforms the full-rank subspace $\mathbf{L}^r$. For the Digits data set, SNF outperforms the proposed algorithm in four external indices. The proposed algorithm

has the second best performance and is followed by the full-rank subspace $\mathbf{L}^r$. The Football data set has been recently been used for the performance evaluation of latent multi-view subspace clustering (LMSC) [49] algorithm. LMSC has two formulations, namely, linear (lLMSC) and generalized (gLMSC). For the Football data set, the aggregate F-measure values for lLMSC and gLMSC are 0.7082 and 0.7940, respectively, while aggregate Rand index are 0.9714 and 0.9797, respectively, while F-measure and Rand index for CoALa are 0.8852 and 0.9780, respectively, which show that CoALa outperforms both lLMSC and gLMSC in terms of F-measure. In terms of Rand index, performance of LMSC and CoALa are competitive. Also, the Digits data set has been used for the evaluation of multiple kernel learning based late fusion incomplete multi-view clustering (LF-IMVC) [50] algorithm and spectral clustering based Wang et al.'s algorithm [44]. The aggregate purity and normalized mutual information (NMI) values for Digits data set for LF-IMVC are 0.7980 and 0.6899, respectively, while for Wang et al.'s algorithm NMI achieved is 0.785. For CoALa, aggregate purity and NMI obtained are 0.8835 and 0.797659, respectively. The results imply that CoALa outperforms both these algorithms on Digits data set.

In terms of internal cluster evaluation indices, Table 6 shows that out of 16 cases, the proposed algorithm achieves best performance in 6 cases, while the second best in three cases. For the Twitter data sets, the best modality achieves superior performance for majority of the internal indices. The execution times reported in Table 6 indicate that the proposed method is computationally more efficient compared to SNF for three out of four data sets. Although for omics data sets in Table 5, SNF needs lower execution time compared to CoALa, CoALa demonstrates higher computational efficiency compared to SNF for the benchmark data sets with larger number of component modalities.

6 Conclusion

This paper presents a novel algorithm, for integration of multiple similarity graphs, that prevents the noise of the individual graphs from being propagated into the unified one. The proposed method first approximates each graph using the most informative eigenpairs of its Laplacian which contains its cluster information. Thus, the noise in the individual graphs is not reflected in their approximations. These de-noised approximations are then integrated for the construction of a low-rank subspace that best preserves the overall cluster structure of multiple graphs. However, this approximate subspace differs from the full-rank one which integrates information of all the eigenpairs of each Laplacian. Using the concept of matrix perturbation, theoretical bounds are derived as a function of the approximation rank, inorder to precisely evaluate how far the approximate subspace deviates from the full-rank one. The clusters in the data set are identified by performing k-means clustering on the approximate de-noised subspace. The effectiveness of the proposed approximation based approach is established by showing that the approximate subspace encodes better cluster structure compared to the full-rank one. The clustering performance of the approximate subspace is compared with that of existing integrative clustering approaches on four real-life cancer data sets as well as on four benchmark data sets from varying application domains. Experimental results show that the clusters identified by the proposed approach have closest resemblance with the clinically established cancer subtypes and also with the ground-truth class information, when compared with individual modalities as well as existing algorithms.

Acknowledgments

The publication is an outcome of the R&D work undertaken in the project under the Visvesvaraya PhD Scheme of Ministry of Electronics and Information Technology, Government of India, being implemented by Digital India Corporation.

References

[1] D. Greene and P. Cunningham, "Producing a unified graph representation from multiple social network views," in *Proc. 5th Annu. ACM Web Sci. Conf.*, 2013, pp. 118–121. [Online]. Available: http://doi.acm.org/10.1145/2464464.2464471

[2] A. Djelouah, J. Franco, E. Boyer, F. Le Clerc, and P. Prez,"Sparse multi-view consistency for object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1890–1903, Sep. 2015.

[3] J. Li, C. Xu, W. Yang, C. Sun, and D. Tao, "Discriminative multi-view interactive image re-ranking," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3113–3127, Jul. 2017.

[4] B. Pepik, M. Stark, P. Gehler, and B. Schiele, "Multi-view and 3d deformable part models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2232–2245, Nov. 2015.

[5] N. K. Speicher and N. Pfeifer, "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery," *Bioinf.*, vol. 31, no. 12, pp. i268–275, Jun. 2015.

[6] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biol.*, vol. 18, no. 1, p. 83, May 2017.

[7] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in multi-omics data integration methods," *Frontiers genetics*, vol. 8, 2017, Art. no. 84.

[8] P. Chalise, D. C. Koestler, M. Bimali, Q. Yu, and B. L. Fridley, "Integrative clustering methods for high-dimensional molecular data," *Translational Cancer Res.*, vol. 3, no. 3, 2014, Art. no. 202.

[9] K. A. Hoadley, C. Yau, et al., "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," *Cell*, vol. 158, pp. 929–944, 2014.

[10] E. F. Lock and D. B. Dunson, "Bayesian consensus clustering," *Bioinf.*, vol. 29, no. 20, pp. 2610–2616, Oct. 2013.

[11] W. Zhang, Y. Liu, et al., "Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer," *Cell Reports*, vol. 4, no. 3, pp. 542–553, 2013.

[12] 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," *Bioinf.*, vol. 25, no. 22, pp. 2906–2912, 2009.

[13] 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," *Proc. Nat. Academy Sci. United States America*, vol. 110, no. 11, pp. 4245–4250, 2013.

[14] D. Wu, D. Wang, M. Q. Zhang, and J. Gu, "Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification," *BMC Genomics*, vol. 16, no. 1, 2015, Art. no. 1022.

[15] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, "Joint and individual variation explained (jive) for integrated analysis of multiple data types," *Ann. Appl. Statistics*, vol. 7, no. 1, pp. 523–542, 2013.

[16] Q. Feng, M. Jiang, J. Hannig, and J. Marron, "Angle-based joint and individual variation explained," *J. Multivariate Anal.*, vol. 166, pp. 241–265, 2018.

[17] Z. Zhang, Z. Zhai, and L. Li, "Uniform projection for multi-view learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1675–1689, Aug. 2017.

[18] Y. Lin, T. Liu, and C. Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1147–1160, Jun. 2011.
[19] B. Long, P. S. Yu, and Z. Zhang, "A general model for multiple view unsupervised learning," in *Proc. SIAM Int. Conf. Data Mining*, 2008, pp. 822–833.
[20] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Trans. Syst. Man Cybern., Part B (Cybern.)*, vol. 40, no. 6, pp. 1438–1446, Dec. 2010.
[21] D. Zhou and C. J. Burges, "Spectral clustering and transductive learning with multiple views," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 1159–1166.
[22] A. Kumar, P. Rai, and H. Daumé, III,"Co-regularized multi-view spectral clustering," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1413–1421.
[23] C. Xu, D. Tao, and C. Xu, "Large-margin multi-viewinformation bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.
[24] U. Von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.
[25] M. Meila and J. Shi, "Learning segmentation by random walks," in *Proc. Advances Neural Inf. Process. Syst.*, 2001, pp. 873–879.
[26] M. Maila and J. Shi, "A random walks view of spectral segmentation," in *Proc. AI Statistica (AISTATS)*, Jan. 2001.
[27] F. R. Chung, *Spectral Graph Theory*. Providence, RI, USA: American Mathematical Society, 1997.
[28] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[29] D. Wagner and F. Wagner, "Between min cut and graph bisection," in *Proc. Int. Symp. Math. Foundations Comput. Sci.*, 1993, pp. 744–750.
[30] B. Mohar, Y. Alavi, G. Chartrand, and O. Oellermann, "The laplacian spectrum of graphs," *Graph Theory Combinatorics Appl.*, vol. 2, no. 871–898, 1991, Art. no. 12.
[31] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Advances Neural Inf. Process. Syst.*, 2002, pp. 849–856.
[32] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, Sep. 1936.
[33] C. Dhanjal, R. Gaudel, and S. Clémençon, "Efficient eigen-updating for spectral graph clustering," *Neurocomputing*, vol. 131, pp. 440–452, 2014.
[34] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*. New York, NY, USA: Academic Press, 1990.
[35] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory," *Czechoslovak Math. J.*, vol. 25, no. 4, pp. 619–633, 1975. [Online]. Available: http://dml.cz/dmlcz/101357
[36] D. A. Spielman and S.-H. Teng, "Spectral partitioning works: Planar graphs and finite element meshes," *Linear Algebra Appl.*, vol. 421, no. 2, pp. 284–305, 2007, special issue in honor of Miroslav Fiedler. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0024379506003454
[37] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
[38] A. Bjorck and G. Golub, "Numerical methods for computing the angles between linear subspaces," *Mathematics Comput.*, vol. 27, pp. 579–594, 1973.
[39] K. Fukui and A. Maki, "Difference subspace and its generalization for subspace-based methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2164–2177, Nov. 2015.
[40] S. Ji-guang, "Perturbation of angles between linear subspaces," *J. Comput. Math.*, vol. 5, no. 1, pp. 58–61, 1987.
[41] A. V. Knyazev and P. Zhu, "Principal angles between subspaces and their tangents," Mitsubishi Electric Research Laboratories, Cambridge, MA, Tech. Rep. TR2012–058, Sep. 2012.
[42] C. Davis and W. Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM J. Numerical Anal.*, vol. 7, no. 1, pp. 1–46, 1970.
[43] I. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer, 2002.
[44] B. Wang, et al., "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014.
[45] A. Khan and P. Maji, "Low-rank joint subspace construction for cancer subtype discovery," *IEEE/ACM Trans. Comput. Biology Bioinf.*, early access, Jan. 23, 2019, doi: 10.1109/TCBB.2019.2894635.
[46] TCGA ResearchNetwork, "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas," *New England J. Med.*, vol. 372, no. 26, pp. 2481–2498, 2015.
[47] TCGA Research Network, "Comprehensive molecular characterization of gastric adenocarcinoma," *Nature*, vol. 513, no. 7517, pp. 202–209, 2014.
[48] TCGANetwork, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.
[49] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, and D. Xu, "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 23, 2018, doi: 10.1109/TPAMI.2018.2877660.
[50] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, and W. Gao, "Late fusion incomplete multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2410–2423, Oct. 2019.

Aparajita Khan received the BE degree in computer science and engineering from Burdwan University, India, in 2012, and the MTech degree in computer technology from Jadavpur University, India, in 2015. Currently, she is a research scholar in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata. Her research interests include pattern recognition, machine learning, computational biology and bioinformatics and so forth. She has published a few papers in international journals and conferences. She was the recipient of the University Gold Medal from Jadavpur University, India, for standing First in Master of Technology in Computer Technology, 2015.

Pradipta Maji received the BSc degree in physics, the MSc degree in electronics science, and the PhD degree in the area of computer science from Jadavpur University, India, in 1998, 2000, and 2005, respectively. Currently, he is a professor with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. His research interests include pattern recognition, machine learning, computational biology and bioinformatics, medical image processing, and so forth. He has published more than 125 papers in international journals and conferences. He is author of a book published by Wiley-IEEE Computer Society Press and another book published by Springer-Verlag, London. He has received the 2008 Microsoft Young Faculty Award from Microsoft Research Laboratory India Pvt., the 2009 Young Scientist Award from the National Academy of Sciences, India, the 2011 Young Scientist Award from the Indian National Science Academy, India, and the 2015 Young Faculty research fellowship from the Department of Electronics and Information Technology, Government of India. He has been selected as the 2009 Young Associate of the Indian Academy of Sciences, India, and elevated to the 2016 Senior Member of the IEEE, USA. He is a Fellow of the National Academy of Sciences, India.