Selective Update of Relevant Eigenspaces for Integrative Clustering of Multimodal Data

Aparajita Khan and Pradipta Maji*, Senior Member, IEEE

Abstract—One of the major problems in cancer subtype discovery from multimodal omic data is that all the available modalities may not encode relevant and homogeneous information about the subtypes. Moreover, the high dimensional nature of the modalities makes sample clustering computationally expensive. In this regard, a novel algorithm is proposed to extract a low-rank joint subspace of the integrated data matrix. The proposed algorithm first evaluates the quality of subtype information provided by each of the modalities, and then judiciously selects only relevant ones to construct the joint subspace. The problem of incrementally updating the singular value decomposition of a data matrix is formulated for the multimodal data framework. The analytical formulation enables efficient construction of the joint subspace of integrated data from low-rank subspaces of the individual modalities. Construction of joint subspace by the proposed method is shown to be computationally more efficient as compared to performing principal component analysis (PCA) on the integrated data matrix. Some new quantitative indices are introduced to measure theoretically the accuracy of subspace construction by the proposed approach with respect to the principal subspace extracted by the PCA. The efficacy of clustering on the joint subspace constructed by the proposed algorithm is established over existing integrative clustering approaches on several real-life multimodal cancer data sets.

Index Terms—Integrative clustering, low-rank approximation, singular value decomposition, modality selection.

I. INTRODUCTION

Integrative genomic data analysis refers to the design of algorithms to combine, infer, and analyze data from multiple genomic modalities like gene expression, DNA methylation, copy number variation, etc. Data from a single modality reflects biological patterns and variations within a specific molecular level. Integrative analysis allows modeling of intrinsic patterns of the individual modalities, and also captures correlated patterns across multiple modalities. One major objective of integrative genomic data analysis is cancer subtype discovery. Cancer subtyping provides insight into disease pathogenesis and design of personalized therapies. Data driven subtype discovery is most popularly achieved by clustering data from one or more genomic modalities [1]. Several integrative clustering algorithms exist for cancer subtyping [2]–10. Clustering multimodal genomic data has mainly three major challenges. The main challenge is the appropriate selection of modalities those provide relevant and shared subtype information, over modalities that provide noisy and inconsistent information. Another challenge is handling the highly heterogeneous nature, in terms of scale, unit, and variance, of different genomic modalities. Moreover, due to the high dimensional nature of the genomic modalities, the feature space becomes geometrically sparse; and most of the clustering methods become computationally expensive and prone to degrade their performance [11].

The existing integrative clustering approaches do not address all these challenges together. They can be divided into two categories, namely, two-stage consensus clustering approaches and direct integrative clustering approaches. Two-stage approaches such as Bayesian consensus clustering (BCC) [4] and cluster-of-cluster assignments (Coca) [5], [6] first cluster each modality separately, and the clustering solutions are then grouped using consensus clustering [12], to get the final cluster assignments. As individual modalities are clustered separately, the difference in scale, unit, and variance between modalities are automatically taken care of by the two-stage approaches. However, the second-stage integration of separate clustering results fails to capture the correlation and joint structure shared among different modalities. On the other hand, direct integrative clustering approaches concatenate data matrices obtained from multiple modalities into a single matrix, which is used to get the joint clusters. However, the curse of dimensionality gets amplified due to the concatenation of several modalities. This problem is addressed by first embedding the data into a low-rank subspace followed by clustering in that subspace. Several algorithms such as iCluster [2], iCluster+ [3], joint and individual variation explained (JIVE) [7], and LRAcluster [8] use subspace clustering approaches.

Most of the existing integrative clustering approaches assume that all the available modalities provide homogeneous and consistent subtype information; and thus consider all of them for integrative clustering. However, some omic modalities may provide disparate or even worse information [10]. Due to the presence of such noisy modalities, naive integration of information from all the available modalities can degrade the final cluster structure. During data integration, relevant modalities with shared cluster information should be chosen, instead of considering noisy and inconsistent ones. Therefore, one of the important problems in multimodal data clustering is how to select a subset of relevant modalities.

Another major challenge in clustering high-dimensional omics data is how to extract a lower dimensional subspace that best preserves the underlying cluster structure. Principal component analysis (PCA) is an extensively used dimensionality reduction method for large-scale genomic data sets.
It extracts the principal subspace that maximizes the variance along the projected axes and also minimizes the reconstruction error for any given rank. The principal subspace can be effectively represented using eigenspaces, which are widely used in various pattern recognition and image processing applications [15]–[18]. Eigenspaces can be computed using singular value decomposition (SVD) which has high computational complexity. This motivates the use of eigenspace update algorithms to prevent re-computation of eigenspace from scratch every time new observations are added to the data set. Such strategies include incremental update [15], [19] where eigenspace is updated on addition of every new observation, and block update [20]–[22] where update occurs on addition of new sets of observations. However, these algorithms have been proposed for a framework where data sets are incrementally updated with new observations. Hence, the SVD eigenspace model is used in this work.

Let $X \in \mathbb{R}^{n \times d}$ be a data matrix of $n$ observations or samples, each having $d$ features, and $\text{rank}(X) = r$. The SVD of the mean-centered data matrix $X$ is given by

$$X - \mathbf{1}_n \mu(X)^T = U(X) \Sigma(X) V(X)^T,$$

where $\mu(X) \in \mathbb{R}^d$ is the mean of the data, $A^T$ denotes the transpose of a matrix $A$, and $\mathbf{1}$ denotes a column vector of length $n$ of all ones. The matrix $U(X)$ contains the $r$ left singular vectors of $X$ in its columns, which gives the $r$-dimensional principal subspace projection of the $n$ samples of $X$. $\Sigma(X)$ is a diagonal matrix with entries $\sigma_1 \geq \cdots \geq \sigma_r$, where $\sigma_1 \geq \cdots \geq \sigma_r > 0$. The $\sigma_i$’s are the singular values of $X$, which give the spread of the projections along singular vectors in $U(X)$. The matrix $V(X)$ contains the $r$ right singular vectors of $X$ in its columns, which are the loadings of the $d$ variables of $X$ corresponding to the projections in $U(X)$. The principal components of $X$ are obtained by multiplying the projections in $U(X)$ with the corresponding spread values in $\Sigma(X)$, given by $Y = U(X) \Sigma(X)$. The eigenspace of $X$ is given by a four-tuple as follows [22]:

$$\Psi(X) = (\mu(X), U(X), \Sigma(X), V(X)).$$

In [23], Zha et al. showed that the continuous relaxation of the discrete cluster membership indicators in $k$-means clustering problem is given by the top $k$ principal components. So, the rank $k$ truncated eigenspace, containing only top $k$ singular vectors and corresponding singular values, sufficiently represents the cluster information of $X$. The noisy information, embedded in remaining $(n - k)$ singular triplets, gets eliminated from the truncated eigenspace. So, the rank $r$ of the eigenspace is considered to be $k$ in the current work.

### II. Eigenspace Model

The basic assumption of the eigenspace model is that the data follows a multivariate Gaussian distribution. Under this assumption, the eigenspace model of the data set refers to the statistical description of a set of $n$ observations in $d$-dimensional space in the form of a hyper-ellipsoid [19]. The hyper-ellipsoid is centered at the mean of the observations, and its axes point in directions where spread of the observations is maximized, subject to orthogonality. The hyper-ellipsoid is flat in the directions where the spread is negligible. This indicates a lower dimensional embedding of the hyper-ellipsoid considering only the top few axes along which the spread is significantly high. Eigenspace models can be computed either by eigenvalue decomposition of the covariance matrix of the data or by SVD of the mean centered data matrix itself. As $n \ll d$ for omics data, computation of large $d \times d$ covariance matrix needs intensive space and time. Also, multicollinearity of omic features often leads to a singular covariance matrix. Hence, the SVD eigenspace model is used in this work.

This section presents the proposed SURE algorithm to construct a low-rank joint subspace of the integrated data. Prior to describing the proposed algorithm, theoretical formulation for the eigenspace update problem is described next.

#### A. Eigenspace Update

Let $X_1, \ldots, X_m, \ldots, X_M$, where $X_m \in \mathbb{R}^{n \times d_m}$, be $M$ different modalities of a multimodal data set, all measured on the same set of $n$ samples. The $M$ data matrices can be concatenated to form an integrated data matrix as follows:

$$X = \begin{bmatrix} X_1 & \cdots & X_m & \cdots & X_M \end{bmatrix}.$$
of the eigenspace of $X$ from the low-rank eigenspaces of the individual modalities.

Let the rank $k$ eigenspace for modality $X_m$ be given by

$$
\Psi(X_m) = \langle \mu(X_m), U(X_m), \Sigma(X_m), V(X_m) \rangle.
$$

(4)

Let the data matrix, formed by column-wise concatenation of $m$ modalities, be given by

$$
\tilde{X}_m = \begin{bmatrix} X_1 & X_2 & \ldots & X_m \end{bmatrix}.
$$

The eigenspace of $X$ is constructed sequentially in $M$ steps by constructing the eigenspace of $\tilde{X}_m$ at each step for $m = 1, \ldots, M$. Let the eigenspace of $\tilde{X}_m$ obtained at $m$-th step be given by

$$
\Psi(\tilde{X}_m) = \langle \mu(\tilde{X}_m), U(\tilde{X}_m), \Sigma(\tilde{X}_m), V(\tilde{X}_m) \rangle.
$$

(5)

At $(m+1)$-th step, the matrix $\tilde{X}_{m+1}$ is formed by concatenation of the matrix $\tilde{X}_m$ of $m$-th step and the $(m+1)$-th modality $X_{m+1}$, given by

$$
\tilde{X}_{m+1} = \begin{bmatrix} \tilde{X}_m & X_{m+1} \end{bmatrix}.
$$

(6)

Let the eigenspace of $\tilde{X}_{m+1}$ be given by

$$
\Psi(\tilde{X}_{m+1}) = \langle \mu(\tilde{X}_{m+1}), U(\tilde{X}_{m+1}), \Sigma(\tilde{X}_{m+1}), V(\tilde{X}_{m+1}) \rangle.
$$

(7)

The relation between a data matrix $X$ and the components of its eigenspace $\Psi(X)$ is obtained by SVD using (1). Applying (1) to data matrices $X_m$ and $X_{m+1}$, the following relations are obtained:

$$
\begin{align*}
\tilde{X}_m - \mu(\tilde{X}_m) & = U(\tilde{X}_m) \Sigma(\tilde{X}_m) V(\tilde{X}_m)^T; \\
X_{m+1} - \mu(X_{m+1}) & = U(X_{m+1}) \Sigma(X_{m+1}) V(X_{m+1})^T.
\end{align*}
$$

(9)

The matrix $\tilde{X}_{m+1}$ is constructed by column-wise concatenation of $X_{m+1}$ to $\tilde{X}_m$. So, the mean component $\mu(\tilde{X}_{m+1})$ of $\Psi(\tilde{X}_{m+1})$ is also obtained directly by column-wise concatenation of mean vectors $\mu(X_m)$ and $\mu(X_{m+1})$, that is,

$$
\mu(\tilde{X}_{m+1}) = \left[ \mu(X_m), \mu(X_{m+1}) \right].
$$

(10)

The left singular subspace $U(\tilde{X}_{m+1})$ consists of unit vectors corresponding to the principal subspace projection of the data in $\tilde{X}_{m+1}$. The matrix $X_m$ has $X_m$ and $X_{m+1}$ as its constituent block matrices. Therefore, the left subspace of $\tilde{X}_{m+1}$ must be constructed in such a way that it contains the information of projection of data from the $m$ modalities in $X_m$ and also the projection information from $(m+1)$-th modality $X_{m+1}$. So, both $U(X_m)$ and $U(X_{m+1})$ must be subspaces of $U(\tilde{X}_{m+1})$. The subspace $U(\tilde{X}_{m+1})$ can be obtained by constructing a basis sufficient to span both the subspaces $U(X_m)$ and $U(X_{m+1})$. $U(\tilde{X}_m)$ is itself a basis for the left subspace of $\tilde{X}_m$. Let $\Gamma$ be the basis for the subspace lying orthogonal to left subspace spanned by $U(\tilde{X}_m)$. Therefore, a sufficient basis for $U(\tilde{X}_{m+1})$ can be formed by augmenting the basis $U(\tilde{X}_m)$ with basis $\Gamma$ of the orthogonal space.

In Fig. 1(a), the gray plane represents the subspace $U(X_{m+1})$ and the dotted plane represents the subspace $U(X_m)$. The basis $\Gamma$ has to span the subspace orthogonal to the dotted plane $U(X_m)$. It is constructed by projecting $U(X_{m+1})$ to $U(X_m)$ and then obtaining orthogonal bases for the residual matrix. The projection is given by

$$
I = U(\tilde{X}_m)^T U(X_{m+1}).
$$

(11)

The component of $U(X_{m+1})$, lying in the subspace spanned by $U(X_m)$, is obtained by multiplying the projection $I$ with the corresponding basis $U(X_m)$. The projected component $\mathcal{P}$ is given by

$$
\mathcal{P} = U(\tilde{X}_m) I.
$$

(12)

Finally, the residual component $\mathcal{Q}$ is obtained by subtracting the projected component $\mathcal{P}$ from $U(X_{m+1})$ itself, given by

$$
\mathcal{Q} = U(X_{m+1}) - \mathcal{P}.
$$

(13)

The residual $\mathcal{Q}$ lies in the subspace orthogonal to the one spanned by $U(\tilde{X}_m)$. In Fig. 1(a), $\mathcal{P}$ denotes the projection of the gray plane $U(X_{m+1})$ onto the dotted plane $U(X_m)$ and the stripped plane denotes the residual component $\mathcal{Q}$, which is orthogonal to the dotted plane $U(X_m)$. An orthonormal basis $\Gamma$ for the residual component can be obtained by Gram-Schmidt orthogonalization of $\mathcal{Q}$. However, if the intersection between
the subspaces $U(\tilde{X}_m)$ and $U(X_{m+1})$ is non-empty, the rank of the residual space reduces. So, the rank of the intersection space is evaluated in order to choose the right number of basis vectors required for the residual space. This can be evaluated using the following theorem.

**Theorem 1.** Let $A$ and $B$ be two subspaces of $\mathbb{R}^n$. Let columns of matrices $A \in \mathbb{R}^{n \times r_1}$ and $B \in \mathbb{R}^{n \times r_2}$ be orthonormal bases for the subspaces $A$ and $B$, respectively. Let SVD of $A^TB$ be $U\Sigma V^T$, where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$. Then, the dimension of the intersection subspace $A \cap B$ is $\omega$ iff $\sigma_1 = \sigma_2 = \ldots = \sigma_\omega = 1 \geq \sigma_{\omega+1}$ [24].

The above theorem states that the number of singular values of $A^TB$, which are equal to 1, gives the dimension of the intersection subspace of $A$ and $B$. For subspaces $U(\tilde{X}_m)$ and $U(X_{m+1})$, the matrices themselves form orthonormal bases. Therefore, according to Theorem 1, the number of singular values of the matrix $I$ of (11), those are equal to 1, gives the dimension of the intersection space. Let $t$ be the number of such singular values of $I$ that are equal to 1. Then, the dimension of the residual space is $(r-t)$, where $r$ is the dimension of the subspace $U(X_{m+1})$. Let $G$ be the orthonormal basis obtained from Gram-Schmidt orthogonalization of $Q$. If the rank of residual space is $(r-t)$, then exactly $t$ column vectors of $G$ would have norm zero. Finding the rank $t$ of the intersection space through SVD of $I$ has complexity of $O(r^3)$. Alternatively, $t$ can be computed by finding the number of vector numbers of $G$ having norm zero. For $t > 0$, $(r-t)$ non-zero vectors of $G$ are used to form $\Gamma$, which spans the residual space. Following two special cases arise while considering the intersection between the subspaces.

- **Case 1 - Intersection between two subspaces $U(\tilde{X}_m)$ and $U(X_{m+1})$ is empty:** This case arises when $U(X_{m+1})$ lies entirely in the subspace orthogonal to $U(\tilde{X}_m)$ as shown in Fig. 1(b). Therefore, when $U(X_{m+1})$ is projected onto $U(\tilde{X}_m)$, the projection magnitudes are all zeros, that is, $I = 0$, where $0$ denotes a zero matrix of appropriate dimension. Hence, the projected component $P$ in (12) is also $0$. So, the residual $Q$ in (13) is the whole subspace $U(X_{m+1})$, that is, $Q = U(X_{m+1})$, which is itself an orthonormal basis. Therefore, the basis for the residual space is $\Gamma = U(X_{m+1})$.

- **Case 2 - Subspace $U(X_{m+1})$ is a subspace of $U(\tilde{X}_m)$:** This case arises when $U(X_{m+1})$ is itself a subspace of $U(\tilde{X}_m)$, as shown in Fig. 1(c) where the subspaces are parallel to each other. This implies that all the column vectors of $U(X_{m+1})$ can be expressed as a linear combination of those in $U(\tilde{X}_m)$. So, the projected component $P$ in (12) is $U(X_{m+1})$ itself, and the residual $Q$ in (13) is $0$. Since the residual space is empty, the basis $U(\tilde{X}_m)$ is sufficient to span both the subspaces $U(\tilde{X}_m)$ and $U(X_{m+1})$. Therefore, $\Gamma = 0$.

After constructing the appropriate basis $\Gamma$ for residual space, it is appended to the basis $U(\tilde{X}_m)$. Thus, $[U(\tilde{X}_m) \quad \Gamma]$ spans both the subspaces $U(\tilde{X}_m)$ and $U(X_{m+1})$. This basis differs from the required basis $U(X_{m+1})$ by a rotation $R(\tilde{X}_{m+1})$. Hence, $U(X_{m+1})$ is obtained as follows:

$$U(X_{m+1}) = [U(\tilde{X}_m) \quad \Gamma] R(\tilde{X}_{m+1}), \quad (14)$$

where $R(\tilde{X}_{m+1})$ is an orthonormal rotation matrix. The $\Sigma(X_{m+1})$ and $V(X_{m+1})$ components of the eigenspace of $X_{m+1}$ and the rotation matrix $R(\tilde{X}_{m+1})$ are computed as follows. The SVD of $\tilde{X}_{m+1}$ gives the following relation:

$$\tilde{X}_{m+1} - I \mu(\tilde{X}_{m+1})^T = U(\tilde{X}_{m+1}) \Sigma(X_{m+1}) V(\tilde{X}_{m+1})^T. \quad (15)$$

Substituting $U(X_{m+1})$ from (14) in (15), we get

$$\tilde{X}_{m+1} - I \mu(\tilde{X}_{m+1})^T = [U(\tilde{X}_m) \quad \Gamma] R(\tilde{X}_{m+1}) \Sigma(X_{m+1}) V(\tilde{X}_{m+1})^T; \quad (16)$$

$$\Rightarrow \quad R(\tilde{X}_{m+1}) \Sigma(X_{m+1}) V(\tilde{X}_{m+1})^T = [U(\tilde{X}_m) \quad \Gamma]^T \left( \tilde{X}_{m+1} - I \mu(\tilde{X}_{m+1})^T \right). \quad (17)$$

as $U(\tilde{X}_m)$ and $\Gamma$ are orthonormal matrices, and

$$\left[ U(\tilde{X}_m) \quad \Gamma \right]^T \left[ U(\tilde{X}_m) \quad \Gamma \right] = I_s,$$

where $I_s$ is the identity matrix of order $s$.

Substituting the values of $\tilde{X}_{m+1}$ and $\mu(\tilde{X}_{m+1})$ from (6) and (10), respectively, in (17), we get

$$R(\tilde{X}_{m+1}) \Sigma(X_{m+1}) V(\tilde{X}_{m+1})^T = \left[ M_{11} \Sigma(\tilde{X}_m) V(\tilde{X}_m)^T \quad M_{12} \Sigma(X_{m+1}) V(X_{m+1})^T \right] \left[ M_{21} \Sigma(\tilde{X}_m) V(\tilde{X}_m)^T \quad M_{22} \Sigma(X_{m+1}) V(X_{m+1})^T \right]^T; \quad (19)$$

$$M_{11} = U(\tilde{X}_m)^T U(\tilde{X}_m) = I_s; \quad M_{21} = \Gamma^T U(X_{m+1}) = 0; \quad M_{12} = U(\tilde{X}_m)^T U(X_{m+1}) = I_s; \quad M_{22} = \Gamma^T U(X_{m+1}).$$

Substituting the values of $M_{ij}, \forall i, j = 1, 2$ in (19), we get

$$R(\tilde{X}_{m+1}) \Sigma(X_{m+1}) V(\tilde{X}_{m+1})^T = \left[ I_s \Sigma(\tilde{X}_m) V(\tilde{X}_m)^T \quad 0 \quad I_s \Sigma(X_{m+1}) V(X_{m+1})^T \right] \left[ 0 \quad 0 \quad M_{22} \Sigma(X_{m+1}) V(X_{m+1})^T \right]^T. \quad (20)$$

Solving the SVD problem for the matrix of (20), the components $R(\tilde{X}_{m+1}), \Sigma(X_{m+1}),$ and $V(X_{m+1})$ are obtained. The left subspace $U(X_{m+1})$ is obtained by substituting the value of $R(\tilde{X}_{m+1})$ in (14). Finally, the matrices $U(\tilde{X}_{m+1})$ and $V(X_{m+1})$ are truncated to store only the top $k$ singular vectors and $\Sigma(X_{m+1})$ is truncated to store the corresponding $k$ largest singular values in the eigenspace of $X_{m+1}$. For Case 1, where intersection between two left subspaces is empty, substituting the values $I = 0$ and $\Gamma = U(X_{m+1})$ in the SVD of (20), we get

$$R(\tilde{X}_{m+1}) \Sigma(X_{m+1}) V(X_{m+1})^T = \left[ I_k \Sigma(\tilde{X}_m) V(\tilde{X}_m)^T \quad 0 \quad I_k \Sigma(X_{m+1}) V(X_{m+1})^T \right]. \quad (21)$$
The SVD of (21) is a block-diagonal SVD problem whose solution is given by
\[
\mathcal{R}([\tilde{X}_{m+1}]) = I_{2k}; \Sigma([\tilde{X}_{m+1}]) = \begin{bmatrix} \Sigma(\tilde{X}_m) & 0 \\ 0 & \Sigma(X_{m+1}) \end{bmatrix}; \\
V(\tilde{X}_{m+1})^T = \begin{bmatrix} V(\tilde{X}_m)^T & 0 \\ 0 & V(X_{m+1})^T \end{bmatrix}.
\]
Substituting \(\mathcal{R}([\tilde{X}_{m+1}])\) in (14),
\[
U(\tilde{X}_{m+1}) = [U(\tilde{X}_m) U(X_{m+1})].
\]
This signifies that for non-intersecting subspaces \(U(\tilde{X}_m)\) and \(U(X_{m+1})\), the bases for the joint left and right singular subspaces are formed by the union of the individual bases and has rank \(2k\). In the context of integrative clustering, this implies that the cluster structure reflected in modality \(X_{m+1}\) is completely disparate with respect to cluster structure embedded in joint modality \(X_m\). So, incorporation of modality \(X_{m+1}\) can introduce totally inconsistent cluster information into the joint cluster structure embedded in the eigenspace of \(X_m\). Therefore, careful evaluation of a modality is necessary before updating it into the joint eigenspace.

B. Evaluation of Individual Modality

This section introduces two modality evaluation measures, namely, relevance and concordance. While relevance assesses the quality of cluster information provided by each modality, the concordance measures the amount of cluster information shared between two modalities.

1) Relevance: The relevance of a modality is defined in terms of the compactness of the cluster structure embedded in its eigenspace. The compactness is evaluated in the left subspace, which contains principal subspace projection of the samples. The relevance measure is independent of the difference in scale, unit, and variance of the modalities, as the left subspace of each modality contains \(k\) unit vectors. The compactness of cluster structure of a modality \(X_i\) is given by the percentage of variance explained (PVE) by a partition of its left subspace \(U(X_i)\). Let \(C^i = \{C_{i1}, . . . , C_{i, m}\}\) be a partition of the left subspace \(U(X_i)\) into \(k\) clusters. The PVE in \(U(X_i)\) by partition \(C^i\) is given by the ratio of between-cluster variance in \(C^i\) to the total variance of \(U(X_i)\). The relevance of a modality \(X_i\) is given by
\[
\text{Rel}(X_i) = \text{PVE}(U(X_i)) = \frac{\text{SSB}}{\text{SST}}.
\]
The relevance measure gives a value in between 0 and 1 with higher value indicating better cluster structure. So, the modality \(X_i\) has higher relevance than modality \(X_j\) if \(\text{PVE}(U(X_i)) > \text{PVE}(U(X_j))\). The relevance measure gives an ordering of the modalities, based on the quality of their cluster structures.

2) Concordance: The construction of joint eigenspace begins with the most relevant modality, having the best inherent cluster structure. Updating this eigenspace with a modality having very discordant cluster structure may degrade the final cluster solution. Therefore, a concordance measure, based on normalized mutual information (NMI) [25] between the cluster assignments of two modalities, is used to capture the joint cluster information shared between two modalities. Let \(C^i\) and \(C^j\) be \(k\)-partitions of the subspaces \(U(X_i)\) and \(U(X_j)\), respectively. The concordance \(\mathcal{C}\) between \(X_i\) and \(X_j\) is given by the NMI between the cluster solutions \(C^i\) and \(C^j\)
\[
\mathcal{C}(X_i, X_j) = \text{NMI}(C^i, C^j).
\]
The value of concordance \(\mathcal{C}\) lies in the range \([0, 1]\), with larger value being indicative of more shared information between two modalities. While selecting a modality, the average concordance between a candidate modality and all the previously integrated ones is computed. A candidate modality is selected for update only if its average concordance is beyond some threshold \(\tau\).

C. SURE: Proposed Algorithm

The relevance and concordance measures together help to select relevant modalities during data integration. The main steps of the proposed SURE algorithm are reported next. Let \(X_1, . . . , X_m, . . . , X_M\), where \(X_m \in \mathbb{R}^{n \times d_m}\), be \(M\) modalities, \(S\) is the set of selected modalities and initially \(S = \emptyset\). The SURE algorithm constructs the joint eigenspace \(\Psi(X_M)\) as follows:

1: for \(m \leftarrow 1\) to \(M\) do in parallel
2: Compute eigenspace: \(\Psi(X_m)\) using SVD of \(X_m\).
3: Perform \(k\)-means on the left subspace \(U(X_i)\) of \(X_i\).
4: Compute relevance: \(\text{Rel}(X_m)\) using (22).
5: end for
6: Compute pairwise concordance \(\mathcal{C}(X_i, X_j)\), \(\forall i \neq j\).
7: \(X_s\) ← modality with maximum relevance.
8: \(m \leftarrow 1; S = \{X_s\}\).
9: \(X_\pi = X_s;\) Initial eigenspace: \(\tilde{\Psi}(X_\pi) = \Psi(X_s)\).
10: for \(m \leftarrow 1\) to \((M - 1)\) do
11: for each \(X_j\) not added to joint eigenspace \(\tilde{\Psi}(X_m)\) do
12: Compute average concordance of \(X_j\) with previously integrated modalities: \(\tilde{\mathcal{C}}(X_j) = 1/|S| \sum_{X_i \in S} \mathcal{C}(X_i, X_j)\).
13: end for
14: \(X_j\) ← \(X_j\) with maximum average concordance.
15: if \(\tilde{\mathcal{C}}(X_j) \geq \tau\), then
16: Update \(\tilde{\Psi}(X_{m+1}) = \tilde{\Psi}(X_m) \oplus \Psi(X_i)\) as follows:
17: \(X_{m+1} = \{X_{m+1}, X_i\}\).
18: \(m \leftarrow m + 1; S = S \cup \{X_i\}\).
19: Compute \(\mu(X_{m+1})\) using (10).
20: Compute \(I, P, \text{ and } Q\) using (11), (12), and (13), respectively.
21: \(G\) ← Gram-Schmidt orthogonalization of \(Q\).
22: \(t\) ← number of columns of \(G\) having norm zero.
23: \(L\) ← first \((k - t)\) basis vectors of \(G\).
24: Compute \(\mathcal{R}(\tilde{X}_{m+1}), \Sigma(\tilde{X}_{m+1}), \text{ and } V(\tilde{X}_{m+1})\) using SVD of (20).
25: Compute \(U(\tilde{X}_{m+1})\) from (14).
26: Truncate the matrices \(U(\tilde{X}_{m+1}), \Sigma(\tilde{X}_{m+1}), \text{ and } V(\tilde{X}_{m+1})\) at rank \(k\).
27: \(\tilde{\Psi}(X_{m+1}) = \langle \mu(\tilde{X}_{m+1}), U(\tilde{X}_{m+1}), \Sigma(\tilde{X}_{m+1}), V(\tilde{X}_{m+1})\rangle\).
28: else break
29: end for
30: \(\Psi(X_M) = \tilde{\Psi}(X_m)\).

After the construction of \(\tilde{\Psi}(X_M)\), the principal components of the integrated data are obtained by
\[
Y = U(\tilde{X}_M) \Sigma(\tilde{X}_M).
\]
Finally, the rows of \((n \times k)\) matrix \(Y\) are clustered using \(k\)-means algorithm to obtain the cancer subtypes.
IV. ACCURACY OF EIGENSPACE CONSTRUCTION

This section introduces some quantitative indices to measure the gap between “full-rank” eigenspace of the integrated data and its approximate eigenspace constructed by the proposed SURE algorithm. Let $\mathbf{X}$ be the integrated data given by (3). The full-rank eigenspace of $\mathbf{X}$ contains the full-rank information of all its component modalities and constructed by the SVD of $\mathbf{X}$ using (1). Let its rank $r$ eigenspace of $\mathbf{X}$ be given by

$$
\Psi(\mathbf{X}) = \langle \mu(\mathbf{X}), U(\mathbf{X}^T), \Sigma(\mathbf{X}^T), V(\mathbf{X}^T) \rangle.
$$

The superscript $r$ denotes that $r$ largest singular values and corresponding singular vectors are considered in the eigenspace. This full-rank eigenspace representation is also same as the principal subspace extracted by PCA on the integrated data using (1). Let $\Psi(\mathbf{X})$ be the approximate rank $r$ eigenspace of $\mathbf{X}$ obtained by the proposed SURE algorithm, that is,

$$
\Psi(\mathbf{X}) = \bigoplus_{m=1}^M \Psi(X_m),
$$

where $\Psi(X_m)$ is the rank $r$ eigenspace for modality $X_m$. It is further assumed that all the $M$ modalities are used during the eigenspace update. Let $\Psi(\mathbf{X})$ be given by

$$
\Psi(\mathbf{X}) = \langle \mu(\mathbf{X}), U(\mathbf{X}^T), \Sigma(\mathbf{X}^T), V(\mathbf{X}^T) \rangle.
$$

Here, $\Psi(\mathbf{X})$ is an approximate eigenspace of $\mathbf{X}$ as it is constructed from truncated rank $r$ individual eigenspaces. The truncation errors, inherent in individual eigenspaces, get propagated onto joint eigenspace during the updating process. This results in a gap between full-rank eigenspace $\Psi(\mathbf{X})$ and approximate eigenspace $\Psi(\mathbf{X})$. However, as $r$ increases, the truncation errors in the individual eigenspaces reduce and the gap decreases. So, the gap between two eigenspaces can be computed for different values of rank $r$. For any $r' > r$, an eigenspace of rank $r'$ has more singular values and vectors in its $\Sigma$, $U$, and $V$ components than an eigenspace of rank $r$. So, for different values of $r$, the gap is always measured between fixed number of singular values and vectors of two eigenspaces.

A. Error Bound on Principal Sines

The gap between left and right subspaces can be measured using the principal angles between subspaces (PABS) [24]. PABS generalizes the concept of angle between two lines to a set of angles between two subspaces, defined next.

**Definition 1.** Let $\mathcal{A}$ and $\mathcal{B}$ be two subspaces of $\mathbb{R}^n$ of dimension $r_1$ and $r_2$, respectively. Let $t = \min(r_1, r_2)$. The principal angles between subspaces $\mathcal{A}$ and $\mathcal{B}$ are given by a sequence of $t$ angles, $\Theta(\mathcal{A}, \mathcal{B}) = [\theta_1, \theta_2, \ldots, \theta_t]$, where $0 \leq \theta_1 \leq \ldots \leq \theta_t \leq \pi/2$. The angle $\theta_0$ is defined by

$$
\theta_0 = \max_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \arccos (|a^T b|);
$$

subject to $||a|| = ||b|| = 1$, $a_i^T a = 0$, $b_i^T b = 0$, for $i = 1, 2, \ldots, j - 1$ [24].

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

**Theorem 2.** Let the columns of matrices $A \in \mathbb{R}^{n \times r_1}$ and $B \in \mathbb{R}^{n \times r_2}$ be orthonormal bases for subspaces $\mathcal{A}$ and $\mathcal{B}$, respectively. Let $[A \ A^\perp]$ be a unitary matrix such that the columns of $A^\perp$ span the subspace orthogonal to $\mathcal{A}$. Also, let the singular values of $(A^\perp)^T B$ be given by the elements of the diagonal matrix $\Xi = \text{diag}(\nu_1, \ldots, \nu_t)$, where $\nu_1 \geq \ldots \geq \nu_t \geq \ldots \geq \nu_t$. The principal sine $\sin(\theta_{t+1-j}) = \nu_j$ [26], [27].

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

**Definition 2.** Let $\mathcal{A}$ and $\mathcal{B}$ be two subspaces of $\mathbb{R}^n$. Let the diagonal matrix $\Xi$ contains the singular values of $(A^\perp)^T B$ as in Theorem 2. The measure of difference between two subspaces $\mathcal{A}$ and $\mathcal{B}$ is defined by $\sin \Theta(\mathcal{A}, \mathcal{B}) = \Xi$ [28].

The squared Frobenius norm of a matrix, denoted by $|| \cdot ||_F^2$, is the sum of squares of its singular values. So, using Theorem 2 and Definition 2, we get

$$
|| \sin \Theta(\mathcal{A}, \mathcal{B}) ||_F^2 = || \Xi ||_F^2 = \sum_{j=1}^t \nu_j^2 = \sum_{j=1}^t \sin^2 (\theta_{t+1-j}).
$$

Hence, (26) implies that the sum of squares of the principal sines between two subspaces $\mathcal{A}$ and $\mathcal{B}$ is given by $|| \sin \Theta(\mathcal{A}, \mathcal{B}) ||_F^2$.

The gaps between two left subspaces $U(\mathbf{X}^T)$ and $U(\tilde{\mathbf{X}}^T)$ and two right subspaces $V(\mathbf{X}^T)$ and $V(\tilde{\mathbf{X}}^T)$ are computed using the sum of squares of the principal sines between the two sets of subspaces. The matrices $U(\mathbf{X}^T)$ and $U(\tilde{\mathbf{X}}^T)$ are themselves orthonormal bases of rank $r$ for the corresponding left subspaces. Let the principal angles between subspaces $U(\mathbf{X}^T)$ and $U(\tilde{\mathbf{X}}^T)$ be given by $\theta_1, \ldots, \theta_r$ and the singular values of $U(\mathbf{X}^T)^T U(\tilde{\mathbf{X}}^T)$ be given by $\gamma_1, \ldots, \gamma_r$, arranged in decreasing order, where columns of $U(\mathbf{X}^T)^T$ span the subspace orthogonal to one spanned by $U(\mathbf{X}^T)$. Then, following Theorem 2 and Definition 1, the sum of squared principal sines between two left subspaces $U(\mathbf{X}^T)$ and $U(\tilde{\mathbf{X}}^T)$ is given by

$$
|| \sin \Theta(U(\mathbf{X}^T), U(\tilde{\mathbf{X}}^T)) ||_F^2 = \sum_{i=1}^r \gamma_i^2 = \sum_{i=1}^r \sin^2 (\theta_{r+1-i}).
$$

Similarly, for two right subspaces $V(\mathbf{X}^T)$ and $V(\tilde{\mathbf{X}}^T)$, let the principal angles between them be given by $\phi_1, \ldots, \phi_r$ and the singular values of $V(\mathbf{X}^T)^T V(\tilde{\mathbf{X}}^T)$ be given by $\omega_1, \ldots, \omega_r$, arranged in decreasing order, where columns of $V(\mathbf{X}^T)^T$ span the subspace orthogonal to $V(\mathbf{X}^T)$. Then, sum of squared principal sines between two right subspaces is given by

$$
|| \sin \Theta(V(\mathbf{X}^T), V(\tilde{\mathbf{X}}^T)) ||_F^2 = \sum_{j=1}^r \omega_j^2 = \sum_{j=1}^r \sin^2 (\phi_{r+1-j}).
$$

The cumulative gap between full-rank and approximate pairs of left and right subspaces is given by the root mean squared
principal sines between them, which is given by

\[
\text{Gap}(X^r, \bar{X}^r) = \left[ \frac{1}{2r} \left\{ \left\| \sin(\Theta(U(X^r)), U(\bar{X}^r)) \right\|_F^2 + \left\| \sin(\Theta(V(X^r)), V(\bar{X}^r)) \right\|_F^2 \right\}^{\frac{1}{2}} \right.
\]

(27)

Since the principal angles \( \theta_i^\prime_s \) and \( \phi_j^\prime_s \) lie in \([0, \pi/2]\), \( \sin^2 \theta_i^\prime_s \) and \( \sin^2 \phi_j^\prime_s \) lie in \([0, 1]\) and \text{Gap}(\cdot) \) also lies in \([0, 1]\). If the approximate left and right subspaces \( U(X^r) \) and \( V(X^r) \) are close approximations of the full-rank ones, then \( \theta_i^\prime_s \) and \( \phi_j^\prime_s \) are close to 0. This implies that a value of \text{Gap}(\cdot) \) close to 0 indicates a better approximation.

Next, upper bound on the value of \text{Gap}(X^r, \bar{X}^r) \) is derived as a function of rank \( r \) of the singular subspaces. Without loss of generality, let us assume that the individual modalities \( X_m^r \)'s are mean-centered and have dimension \((n \times d_m)\), where \( n \leq d_m \). The SVD of a modality \( X_m^r \) can be partitioned as:

\[
X_m^r = U(X_m^r)\Sigma(X_m^r)V(X_m^r)^T
\]

(28)

where \( \Sigma(X_m^r) = \text{diag}(\lambda_{0_m}^r, \ldots, \lambda_{r_m}^r) \) consists of \( r \) largest singular values of \( X_m^r \), and \( U(X_m^r) \) and \( V(X_m^r) \) contain the corresponding left and right singular vectors in their columns, respectively. Similarly, \( \Sigma(X_m^{r+1}) \) contains the remaining \((n-r)\) singular values \( \lambda_{r_{m+1}}^r, \ldots, \lambda_{n-1_m}^r \), while \( U(X_m^{r+1}) \) and \( V(X_m^{r+1}) \) contain the corresponding singular vectors. Thus, \( X_m^r \) is the rank \( r \) approximation of \( X_m^r \) using the \( r \) largest singular triplets, and \( X_m^{r+1} \) is the approximation using the remaining \((n-r)\) singular triplets. Using (28), the integrated data matrix \( X \) in (3) can be decomposed as

\[
X = [X_1 \ldots X_m \ldots X_M]
\]

\[
= \left([X_1^r + X_1^{r+1}] \ldots [X_m^r + X_m^{r+1}] \ldots [X_M^r + X_M^{r+1}]\right)
\]

\[
= X^r + X^{r+1}.
\]

(29)

Thus, \( X \) is the full-rank integrated data and \( X^r \) is its approximation using rank \( r \) approximations of the individual modalities. The SVD of \( X \) is used to obtain the full-rank eigenvalues \( \Psi(X) \) in (24). On the other hand, the proposed algorithm constructs the approximate eigenvalues \( \Psi(X) \) for data matrix \( X^r \) by iteratively updating the rank \( r \) eigenspaces of the individual modalities \( \Psi(X_m^r) \)'s. Let the SVD of \( X \) be partitioned as

\[
X = U(X)\Sigma(X)V(X)^T
\]

(30)

and the SVD of \( X^r \) obtained by eigenspace update be partitioned as

\[
X^r = U(\bar{X})\Sigma(\bar{X})V(\bar{X})^T
\]

(31)

where \( U(X^r), U(\bar{X}^r) \in \mathbb{R}^{n \times r}, V(X^r), V(\bar{X}^r) \in \mathbb{R}^{d \times r}, \) and \( \Sigma(X^r) = \text{diag}(\sigma_1^r, \ldots, \sigma_r^r), \Sigma(\bar{X}^r) = \text{diag}(\tilde{\sigma}_1^r, \ldots, \tilde{\sigma}_r^r), \Sigma(X^{r+1}) = \text{diag}(\sigma_{r+1}, \ldots, \sigma_n), \Sigma(\bar{X}^{r+1}) = \text{diag}(\tilde{\sigma}_{r+1}, \ldots, \tilde{\sigma}_n) \).

According to (29), \( X = X^r + X^{r+1} \), therefore, using matrix perturbation theory [28], the integrated data matrix \( X \) can be viewed as a perturbation of its rank \( r \) approximation \( X^r \) due to the presence of error component \( X^{r+1} \). Next, Wedin’s sin \( \Theta \) theorem [29] can be used to bound the principal angles between the rank \( r \) left and right singular subspaces of a matrix and its perturbation. Let the residuals of left and right subspaces be

\[
\mathbb{R}_L = \mathbb{R}_L^r V(X^r) - U(X^r)\Sigma(X^r)
\]

and \( \mathbb{R}_R = (X^r)^T U(X^r) - V(X^r)\Sigma(X^r) \).

Let \( \delta \) be defined as

\[
\delta = \inf\left\{ \min_{1 \leq i \leq r, 1 \leq j \leq (n-r)} |\sigma_i - \tilde{\sigma}_{r+j}|, \min_{1 \leq i \leq r} |\sigma_i| \right\}
\]

Welin’s sin \( \Theta \) theorem states that if \( \delta > 0 \), then

\[
\leq \frac{\sqrt{\|\mathbb{R}_L\|_F^2 + \|\mathbb{R}_R\|_F^2}}{\delta}.
\]

(32)

So, \( \text{Gap}(X^r, \bar{X}^r) \) \leq \( \frac{\sqrt{\|\mathbb{R}_L\|_F^2 + \|\mathbb{R}_R\|_F^2}}{\sqrt{2r\delta}} \).

The above relation states that the cumulative sum of squares of the principal sines between the full-rank and approximate left and right subspaces is bounded in terms of the Frobenius norm of the residual matrices \( \mathbb{R}_L \) and \( \mathbb{R}_R \), and the minimum difference between full-rank and approximate sets of singular values, \( \delta \).

As the value of rank \( r \) approaches the full rank \( n \), the residual component \( X^{r+1} \rightarrow 0 \) and \( X^r \rightarrow X \). Similarly, the components \( U(X^r) \), \( \Sigma(X^r) \), and \( V(X^r) \) also tend towards \( U(X), \Sigma(X), \) and \( V(X) \), respectively. So, \( \lim_{r \rightarrow n} \mathbb{R}_L = \mathbb{R}_L^r \) and \( \lim_{r \rightarrow n} \mathbb{R}_R = \mathbb{R}_R^r \), and substituting the limiting values of \( \mathbb{R}_L \) and \( \mathbb{R}_R \) in (32), we get

\[
\lim_{r \rightarrow n} \text{Gap}(X^r, \bar{X}^r) = 0.
\]

This implies that as the approximation rank \( r \) approaches the full rank \( n \), the principal angles between full-rank and approximate pairs of left and right subspaces reduce to 0.

**B. Accuracy of Singular Triplets**

This subsection introduces two more quantitative indices to evaluate the difference between \( U, \Sigma, \) and \( V \) components of full-rank and approximate eigenspaces.
1) Mean Relative Difference of Singular Values: For any rank \( r \), both \( \Sigma(X^r) \) and \( \Sigma(\tilde{X}^r) \) consist of \( r \) largest singular values. The relative difference between the singular values in \( \Sigma(X^r) \) and \( \Sigma(\tilde{X}^r) \), with respect to the singular values of \( \Sigma(X^r) \), is given by a sequence \( \mathcal{H} = [\lambda_1, \ldots, \lambda_i, \ldots, \lambda_r] \), where \[
\lambda_i = \frac{\Sigma(X^r)_i - \Sigma(\tilde{X}^r)_i}{\Sigma(X^r)_i}; \tag{33}
\]

\( \Sigma(.)_i \) is the \( i \)-th largest singular value of the respective eigenspace. The singular values capture the spread of the data along the principal axes. The maximum spread, captured by the singular values in \( \Sigma(X^r) \), is bound by spread captured by the top \( r \) components of the individual eigenspaces. This is much less than the actual spread of samples in \( X \), which is reflected in \( \Sigma(X^r) \). Hence, \( \Sigma(X^r)_i \geq \Sigma(\tilde{X}^r)_i \), so the value of \( \lambda_i \) lies in \([0, 1]\). A value of \( \lambda_i \) close to 0 indicates less difference between the \( i \)-th component of the two eigenspaces. A cumulative measure of the gap between \( \Sigma(X^r) \) and \( \Sigma(\tilde{X}^r) \) is given by the mean of first \( h \) values of \( \mathcal{H} \) as follows:

\[
\text{DiffSV}(X^r, \tilde{X}^r) = \frac{1}{h} \sum_{i=1}^{h} \lambda_i. \tag{34}
\]

The value of \( \text{DiffSV} \) also lies in \([0, 1]\), with a value closer to 0 indicating a better approximation.

2) Relative Dimension of Intersection Space: Let us assume that \( r' \) is the dimension of the space lying in the intersection of two left subspaces \( U(X^r) \) and \( U(\tilde{X}^r) \). According to Theorem 1, reported in Section III-A, \( r' \) is the number of singular values of \( U(X^r)^T U(\tilde{X}^r) \) having value 1. The relative dimension of intersection space between two left subspaces is defined as the ratio of the dimension of intersection space and that of the left subspace \( U(X^r) \), which is as follows:

\[
\text{DimIS}(X^r, \tilde{X}^r) = \frac{r'}{r}; \tag{35}
\]

where \( r' \leq r \). So, the value of \( \text{DimIS} \) lies in \([0, 1]\). If the overlap between two left subspaces is high, the dimension of the intersection subspace \( r' \) is close to \( r \). Thus, the value of \( \text{DimIS} \) close to 1 indicates lower gap between the two left subspaces. Similarly, \( \text{DimIS} \) between two right subspaces \( V(X^r) \) and \( V(\tilde{X}^r) \) can be calculated using the number of singular values of \( V(X^r)^T V(\tilde{X}^r) \) having value 1.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed SURE algorithm is used to extract a low-rank joint subspace of the integrated data. The clustering performance of the extracted subspace is studied and compared with several existing integrative clustering approaches. The approaches compared are (i) two stage clustering approaches, namely, BCC [4] and COCA [6]; and (ii) direct integrative clustering approaches, namely, PCA on naively concatenated data (PCA-Con) [13], JIVE [7], iCluster [2], iCluster+ [3], and LRACluster [8]. The performance of JIVE is reported considering both permutation test (PERM) and Bayesian information criteria (BIC) for rank selection. The source code of the proposed SURE algorithm, written in R language, is available at www.isical.ac.in/~bibl/results/sure/sure.html.

To evaluate the performance of different clustering algorithms, four external cluster evaluation indices, namely, F-measure, adjusted rand index (ARI), purity, and NMI are used, which compare the identified subtypes with the established subtypes. For all four indices, a value close to one indicates that the identified subtypes have close resemblance with the previously established ones. Two other performance measures, namely, \( p \)-value of Cox log-rank test [30] and \( p \)-value of Peto & Peto’s modification of the Gehan-Wilcoxon test [31], are also considered to evaluate the significance of the differences in survival profiles of the identified subtypes.

Multimodal omics data for five types of cancers, namely, cervical carcinoma (CESC), glioblastoma multiforme (GBM), lower grade glioma (LGG), lung carcinoma (LUNG) and kidney carcinoma (KIDNEY), are obtained from The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov/), having 124, 168, 267, 671, and 737 samples, respectively. By comprehensive integrated analysis, TCGA Research Network has identified three molecular subtypes of both CESC [32] and LGG [33], and four subtypes of GBM were identified by Veerhak et al. [34]. The samples of LUNG and KIDNEY data sets are divided into two and three subtypes, respectively, based on the tissue of origin. The CESC, LGG, KIDNEY, and LUNG data sets have four different modalities, namely, gene expression (RNA), DNA methylation (mDNA), miRNA expression (miRNA), and reverse phase protein array expression (RPPA), while the GBM data set has three modalities, namely, RNA, miRNA, and copy number variation (CNV). A brief descriptions of these data sets, along with the computational efficiency of the proposed algorithm and the method for finding optimal concordance threshold \( \tau \), are reported in the supplementary material.

A. Accuracy of Subspace Representation

The proposed SURE algorithm constructs the joint subspace of the integrated data from individual principal subspaces, using an eigenspace update approach. The extracted joint subspace is an approximation of the principal subspace extracted by PCA on the integrated data matrix. Three quantitative indices, namely, \( Gap(\Theta) \), \( \text{DiffSV} \), and \( \text{DimIS} \), are proposed in Section IV to evaluate the gap between full-rank and approximate eigenspaces. To observe the variation in gap between these two eigenspaces with the increase in rank parameter \( r \), the three proposed indices are evaluated for different values of rank \( r \). Due to the high dimension and low sample size nature of the data sets, the full rank of the integrated data matrix is always bounded by the number of samples. So, for each data set, the indices are evaluated for different fractions of the full rank of the integrated data. The value of the \( h \) parameter for the \( \text{DiffSV} \) is set to be 10, which implies that the gap between singular values is measured between the top 10 components of the two eigenspaces. The variation of these quantitative indices, with increase in rank, is shown in Fig. 2 for different data sets.

While Fig. 2(a) shows the root mean squared principal sines between the left and right subspaces of full-rank and approxi-
mate eigenspaces, Fig. 2(b) shows the difference between their singular values. Fig. 2(b) shows that the difference between singular values monotonically decreases to 0 with the increase in rank, for all data sets. Fig. 2(a) shows that the difference between the singular subspaces, in terms of their principal sines, also converges to 0. However, the change in variation in case of singular subspaces is not monotonically decreasing as of singular values in Fig. 2(b). 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 many rank \( r \) subspaces of an \( n \)-dimensional vector space. For smaller values of \( r \), the rank \( r \) singular subspaces of individual modalities can be very different from each other due to the large number of possibilities. Consequently, the approximate singular subspace, constructed from these individual subspaces, tends to vary a lot from the full-rank subspace. However, as \( r \) approaches the full rank \( n \), the number of possible subspaces reduces and the difference between them converges to 0.

Fig. 2(c) shows that the intersection between two left subspaces increases gradually and uniformly with the increase in rank \( r \). But, for the right subspaces, as seen in Fig. 2(d), intersection continues to remain almost 0 for all data sets, until the rank considered for eigenspace is more than 70% of the full rank. This implies that there is more gap between the pair of right subspaces compared to that of left ones. This difference in gap arises because the right subspaces consist of loadings from different sets of variables in different modalities, while the left subspaces consist of the projections of same set of samples across all the modalities. The disjointness of variables in the right subspaces leads to larger gap between the pair of right subspaces.

B. Importance of Data Integration and Modality Selection

To establish the importance of data integration, the clustering performance on top \( k \) principal components of individual modalities is compared with that of the rank \( k \) joint subspace extracted by the proposed algorithm. There can be a total of \((2^M - M - 1)\) possible combinations of two or more modalities from \( M \) modalities. Each multimodal combination gives a different clustering solution. Therefore, the clustering performance of the top \( k \) principal components of each

---

**Table I**

Comparative Performance Analysis of Individual Modalities, PCA Combinations, and SURE

<table>
<thead>
<tr>
<th>Different Data Sets</th>
<th>Modalities/Algorithms</th>
<th>Best Subset by PCA</th>
<th>Selected by SURE</th>
<th>F-Measure</th>
<th>External Evaluation Index</th>
<th>Survival Analysis (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CESC ((k = 3; M = 4))</td>
<td>mRNA, miRNA, RPPA</td>
<td>RNA, miRNA</td>
<td>mRNA</td>
<td>0.5453798</td>
<td>0.1175544</td>
<td>0.2968495</td>
</tr>
<tr>
<td>GBM ((k = 4; M = 3))</td>
<td>mRNA, miRNA, RPPA, SURE</td>
<td>RNA, miRNA</td>
<td>RNA</td>
<td>0.8310850</td>
<td>0.1512301</td>
<td>0.5887097</td>
</tr>
<tr>
<td>LGG ((k = 3; M = 4))</td>
<td>mRNA, miRNA, RPPA</td>
<td>mRNA, miRNA</td>
<td>mRNA</td>
<td>0.7882956</td>
<td>0.5579264</td>
<td>0.8145161</td>
</tr>
<tr>
<td>LUNG ((k = 2; M = 4))</td>
<td>mRNA, miRNA, RPPA, SURE</td>
<td>mRNA, miRNA</td>
<td>mRNA</td>
<td>0.8521028</td>
<td>0.6507274</td>
<td>0.8629352</td>
</tr>
<tr>
<td>KIDNEY ((k = 3; M = 4))</td>
<td>mRNA, miRNA, RPPA, SURE</td>
<td>mRNA, miRNA</td>
<td>mRNA</td>
<td>0.8521028</td>
<td>0.6507274</td>
<td>0.8629352</td>
</tr>
</tbody>
</table>

---

Fig. 2. Different quantitative indices for the evaluation of gap between full-rank and approximate eigenspaces.
multimodal combination is evaluated using Silhouette index [35]. The best combination is chosen to be the one with maximum value of Silhouette index. To evaluate the strength of the proposed SURE algorithm in selecting appropriate subset of modalities, its performance is compared with that of PCA on the best combination of modalities, henceforth termed as PCA_Combine. The comparative performance of the individual modalities, the best multimodal combination, and the proposed SURE approach is reported in Table I.

Table I shows that the joint subspace extracted by the SURE algorithm gives better performance compared to all the unimodal solutions for four data sets, namely, CESC, GBM, LGG, and KIDNEY, in terms of four external evaluation indices. This establishes the significance of integrative analysis over unimodal analysis. For LGG data set, the mDNA gives the best performance among all possible unimodal and multimodal combinations. The SURE algorithm also efficiently chooses only mDNA to construct the final eigenspace. For GBM and LUNG data sets, the modalities selected by SURE algorithm are same as the best combination of modalities obtained for PCA. The combination differs for CESC and KIDNEY data sets, however, the performance of SURE is always better as compared to PCA_Combine. This is due to the fact that the individual eigenspaces in the proposed algorithm are truncated at rank $k$, thus filtering out the noisy information present in them. The joint subspace constructed from these informative truncated eigenspaces preserves better cluster structure compared to PCA_Combine that considers the complete information of each eigenspace. The results in Table I also show that the performance of SURE is at least as good as that of PCA on best combination of modalities for all data sets. This establishes that the proposed SURE approach is able to select the best subset of modalities among all possible $(2^M - 1)$ combinations.

The results corresponding to survival analysis show that the subtypes identified by SURE algorithm have statistically significant difference in survival profiles, considering 5% significance level of both log rank and generalized Wilcoxon tests, for the LGG, and KIDNEY data sets. For different data

<table>
<thead>
<tr>
<th>Different Data Sets</th>
<th>Integration Starts with</th>
<th>Starting Modality</th>
<th>Relevance Measure</th>
<th>Selected Modalities (in order)</th>
<th>External Evaluation Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>CESC</td>
<td>2nd Most Relevant</td>
<td>mDNA</td>
<td>0.4170079</td>
<td>RNA, miRNA, mDNA</td>
<td>0.8120413 0.5588514 0.7976190 0.5815764</td>
</tr>
<tr>
<td></td>
<td>3rd Most Relevant</td>
<td>RPPA</td>
<td>0.4555006</td>
<td>RNA, miRNA, RNA, mDNA</td>
<td>0.8390298 0.633073 0.8548387 0.6759978</td>
</tr>
<tr>
<td></td>
<td>4th Most Relevant</td>
<td>miRNA</td>
<td>0.4495150</td>
<td>RNA, miRNA, mDNA</td>
<td>0.8512028 0.6570274 0.8629032 0.6461964</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>RNA</td>
<td>0.475322</td>
<td>RNA, mDNA, miRNA</td>
<td>0.8512028 0.6570274 0.8629032 0.6461964</td>
</tr>
<tr>
<td>GBM</td>
<td>2nd Most Relevant</td>
<td>CNV</td>
<td>0.4918464</td>
<td>RNA, miRNA, mDNA, CNV</td>
<td>0.4678243 0.1061846 0.4444286 0.1207564</td>
</tr>
<tr>
<td></td>
<td>3rd Most Relevant</td>
<td>miRNA</td>
<td>0.433253</td>
<td>RNA, miRNA, RNA, CNV</td>
<td>0.8120413 0.5588514 0.7976190 0.5815764</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>RNA</td>
<td>0.5085964</td>
<td>RNA, miRNA, mDNA, CNV</td>
<td>0.8120413 0.5588514 0.7976190 0.5815764</td>
</tr>
<tr>
<td>LGG</td>
<td>2nd Most Relevant</td>
<td>RNA</td>
<td>0.447985</td>
<td>RNA, RPPA, miRNA, mDNA</td>
<td>0.6574834 0.3031112 0.6666667 0.3438738</td>
</tr>
<tr>
<td></td>
<td>3rd Most Relevant</td>
<td>miRNA</td>
<td>0.4359114</td>
<td>RNA, RPPA, miRNA, mDNA</td>
<td>0.6574834 0.3031112 0.6666667 0.3438738</td>
</tr>
<tr>
<td></td>
<td>4th Most Relevant</td>
<td>miRNA</td>
<td>0.4287162</td>
<td>RNA, mDNA, miRNA</td>
<td>0.4425295 0.0251035 0.5018727 0.0318103</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>DNA</td>
<td>0.5039683</td>
<td>RNA, mDNA, miRNA</td>
<td>0.8140075 0.551076</td>
</tr>
<tr>
<td>LUNG</td>
<td>2nd Most Relevant</td>
<td>mDNA</td>
<td>0.3535389</td>
<td>RNA, RNA, miRNA, mDNA</td>
<td>0.9417093 0.7806842 0.9418778 0.6878184</td>
</tr>
<tr>
<td></td>
<td>3rd Most Relevant</td>
<td>miRNA</td>
<td>0.3533499</td>
<td>RNA, RNA, miRNA, mDNA</td>
<td>0.9417093 0.7806842 0.9418778 0.6878184</td>
</tr>
<tr>
<td></td>
<td>4th Most Relevant</td>
<td>miRNA</td>
<td>0.3104797</td>
<td>RNA, miRNA, mDNA</td>
<td>0.9417093 0.7806842 0.9418778 0.6878184</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>RNA</td>
<td>0.317991</td>
<td>RNA, miRNA, mDNA</td>
<td>0.9417093 0.7806842 0.9418778 0.6878184</td>
</tr>
<tr>
<td>KIDNEY</td>
<td>2nd Most Relevant</td>
<td>miRNA</td>
<td>0.5325763</td>
<td>miRNA, RNA, mDNA</td>
<td>0.9530685 0.8534490 0.9525102 0.7726162</td>
</tr>
<tr>
<td></td>
<td>3rd Most Relevant</td>
<td>miDNA</td>
<td>0.5909155</td>
<td>miRNA, RNA, mDNA</td>
<td>0.9530685 0.8534490 0.9525102 0.7726162</td>
</tr>
<tr>
<td></td>
<td>4th Most Relevant</td>
<td>miDNA</td>
<td>0.3906367</td>
<td>RNA, miRNA, mDNA</td>
<td>0.9530685 0.8534490 0.9525102 0.7726162</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>RNA</td>
<td>0.5838337</td>
<td>RNA, miRNA, mDNA</td>
<td>0.9530685 0.8534490 0.9525102 0.7726162</td>
</tr>
</tbody>
</table>

C. Importance of Relevance

The proposed algorithm first evaluates the relevance of each modality based on compactness of the cluster structure embedded within its left subspace. The relevance measure provides a linear ordering of the modalities, and the process of integration starts with the most relevant one. To establish the importance of relevance based ordering in data integration, the performance of clustering is studied for three other cases where the process of integration is initiated with the second, third, and fourth most relevant modalities, keeping all other components of the algorithm fixed. For different initiating modalities, different subset of modalities are selected during the construction of joint subspace, giving rise to different clustering solutions. The starting modality for other three cases, their corresponding subset of selected modalities and their comparative performance with the proposed approach are reported in Table II for different data sets.

The results in Table II show that for the LGG data set, only the proposed relevance ordering gives the best performance, while for other orderings the performance is degraded drastically. For the other data sets, however, one or more orderings have the same performance as that of the proposed algorithm. This is due to the presence of the concordance measure and the value of threshold $\tau$ selected for each of those orderings. For example, for the CESC data set, if the process starts with RNA, miRNA has the highest concordance and the remaining modalities have concordance below the optimal threshold $\tau$ selected for CESC. Again, starting with miRNA, only RNA has the highest concordance that exceeds
the optimal threshold. Hence, same subsets of modalities are selected for both the cases of CESC, giving rise to identical clustering performance. Similar cases occur for both GBM and KIDNEY data sets. For LUNG data set, for each different ordering, all four modalities get selected without degrading final clustering performance. However, the proposed ordering gives the best performance with smaller subset of modalities. So, the performance of the proposed relevance based ordering is at least as best as the other orderings.

<table>
<thead>
<tr>
<th>Different Data Sets</th>
<th>Algorithm Settings</th>
<th>External Evaluation Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F-Measure</td>
</tr>
<tr>
<td>CESC</td>
<td>Without SURE</td>
<td>0.8390298</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>0.8812028</td>
</tr>
<tr>
<td>GBM</td>
<td>Without SURE</td>
<td>0.8120413</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>0.8120413</td>
</tr>
<tr>
<td>LGG</td>
<td>Without SURE</td>
<td>0.6574834</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>0.7904750</td>
</tr>
<tr>
<td>LUNG</td>
<td>Without SURE</td>
<td>0.9417093</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>0.9417093</td>
</tr>
<tr>
<td>KIDNEY</td>
<td>Without SURE</td>
<td>0.9230685</td>
</tr>
<tr>
<td></td>
<td>SURE</td>
<td>0.9230685</td>
</tr>
</tbody>
</table>

D. Significance of Concordance

At each iteration of eigenspace update, the proposed algorithm considers the modality having maximum average concordance $\bar{C}$ or shared information with respect to the previously updated ones. However, if the value of $\bar{C}$ is below the optimal threshold of $\tau$, then it is not updated with the current joint eigenspace. To assess the significance of the concordance measure for modality selection, all the modalities are naively integrated based on their relevance ordering, and the clustering performance of the resulting subspace is studied. The comparative performance of this relevance-based subspace (without concordance $\bar{C}$) and the proposed SURE algorithm is reported in Table III. The results in Table III show that for CESC and LGG data sets, selection of a subset of modalities gives better performance compared to the naive integration of all modalities. For GBM, there are only three modalities and the proposed algorithm selects all of them. So, the performance on GBM is identical with or without concordance. For LUNG and KIDNEY data sets, the proposed algorithm selects only three modalities out of four using the concordance measure. However, the results in Table II show that, for these data sets, selection of all four modalities does not degrade the clustering performance. But, the concordance measure for modality selection gives better performance with smaller subset of modalities compared to relevance alone.

E. Performance Analysis of Different Algorithms

Finally, the performance of the proposed SURE algorithm is compared with that of seven existing integrative clustering approaches, namely, BCC [4], COCA [6], JIVE [7], iCluster [2], iCluster+ [3], LRAcluster [8], and PCA-Con [13]. Comparative results with respect to four external indices, survival analysis and execution time are reported in Table IV. The results in Table IV show that the SURE approach performs better than all the existing approaches with respect to the external indices, on four data sets, namely, CESC, GBM, LGG, and LUNG. For LGG data set, the performance of SURE algorithm is significantly better compared to all the existing algorithms. The better performance is attributed to the efficient selection of relevant modalities only during joint subspace construction. For KIDNEY data set, LRAcluster gives the best performance. However, the performance of the SURE on KIDNEY data set, considering only three modalities, is almost close to the best results. The JIVE, iCluster, iCluster+, LRAcluster, and PCA-Con are low-rank based approaches. The results in Table IV show that the joint subspace extracted by the proposed algorithm preserves better cluster structure compared to the ones extracted by these existing low-rank based approaches. This is because the proposed algorithm first truncates the individual eigenspaces at rank $k$, and then considers only the cluster information of top $k$ singular triplets for further integration; thus filtering out the inherent noise present in the $(n - k)$ remaining components. The existing low-rank based approaches, however, consider cluster as well as noisy information of all the modalities; thus giving poor cluster structure in the extracted subspace.

For GBM data, BIC based JIVE algorithm estimates the rank of joint structure to be 0, which implies that the four different modalities do not share any correlated information among them. On the other hand, for LGG and KIDNEY data, the joint rank estimated by JIVE is the same using both BIC and permutation tests. However, the overall performance differs due to difference in rank of the individual modalities estimated by these two criteria. The survival analysis results of Table IV show that the subtypes identified by all the algorithms for LGG and KIDNEY data have significantly different survival profiles. On the other hand, for the CESC and LUNG data sets, most of the algorithms fail to give statistically significant results at 5% significance level.

Comparing the execution time of different algorithms in Table IV, it is seen that SURE has the minimum execution time compared to all existing algorithms on three larger data sets, namely, LGG, LUNG, and KIDNEY, having 267, 671, and 737 samples, respectively. For two smaller data sets, namely, CESC and GBM having 124 and 168 samples, respectively, PCA-Con achieves the minimum execution time. Comparing the execution time of SURE with the state-of-the-art low-rank approaches such as iCluster, iCluster+, and LRAcluster in Table IV, it is evident that the SURE extracts the low-rank subspace in significantly lower time as compared to all these approaches for five data sets. Hence, the proposed algorithm is computationally more efficient compared to all the existing approaches considered in this work.

VI. Conclusion

This paper presents a novel algorithm to extract a low-rank joint subspace of the high dimensional multimodal data. The sample clustering is performed on the extracted subspace to find the subtypes of respective cancer. The problem of updating the SVD of a data matrix is formulated for
multimodal data, where new modalities are added for the same set of samples. The theoretical formulation introduced here enables the proposed SURE algorithm to extract the principal components in lesser time compared to performing PCA on the concatenated data. Some new quantitative indices are proposed to evaluate theoretically the gap between joint subspace extracted by the proposed algorithm and the principal subspace extracted by PCA. Theoretical analysis also shows that the extracted subspace converges to the full-rank subspace extracted by PCA, as the rank approaches full rank of the integrated data. Unlike the existing integrative clustering approaches, the proposed approach considers that each modality may not provide relevant and consistent information about the true subtypes; hence, it evaluates the quality of each modality before integration. The evaluation measures and eigenspace update based approach allow the proposed algorithm to efficiently select only relevant modalities, discarding the noisy and inconsistent ones. The effectiveness of the proposed algorithm for cancer subtype identification has been studied and compared with existing integrative clustering approaches on several real-life multimodal cancer data sets. The experimental results show that the proposed algorithm performs better than unimodal and multimodal approaches in identification of cancer subtypes.

### References


