Supplementary Material: Selective Updation of Relevant Eigenspaces for Integrative Clustering of Multimodal Data

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

The main article introduces a novel algorithm, termed as SURE (Selective Updation of Relevant Eigenspaces), to construct a low-rank joint subspace of a high dimensional multimodal data set. In this supplementary material, Section I contains description of the data sets and the data preprocessing steps. Section II evaluates the statistical power of the data sets for a clustering problem. Hypothesis testing is used to assess whether clusters are "really present" in the data sets as opposed to being artifacts of natural sampling variations. Section III evaluates the computational complexity of the proposed algorithm and also empirically establishes its efficiency over principal component analysis (PCA). Wedin's theorem from matrix perturbation theory, used to derive error bound on the principal sines between true and approximate eigenspaces is described in Section IV. Section V describes the external cluster evaluation measures used in this work to compare the performance of different algorithms. Section VI provides the detailed survival analysis of the the cancer subtypes identified by the proposed SURE approach. Finally, the experimental setup for the existing algorithms is outlined in Section VII.

I. Description of Data Sets

This section presents a brief description of the five multimodal omics data sets of The Cancer Genome Atlas (TCGA) [1] used in this work. All the data sets have been downloaded from the Genomic Data Commons (GDC) Data Portal [2]. The five different genomic modalities considered for the data sets are DNA methylation (DNA), gene expression (Gene), miRNA expression (miRNA), protein expression (Protein), and copy number variation(CNV). Publicly available clinical information for the all the data sets is retrieved using RTCGA.clinical package [3]. The data sets and their established subtypes are described next.

A. Data Sets

Five multimodal cancer data sets used in this work are as follows:

1) **CESC**: This the cervical cancer data set which consists of 124 samples. The recent integrative study by TCGA Research Network [4] has identified three molecular subtypes cervical cancer, namely, Keratin-low Squamous subgroup, Keratin-high Squamous subgroup, and Adenocarcinoma-rich subgroup. The data set consists of 37 samples of Keratin-low Squamous subgroup, 58 samples of Keratin-high Squamous subgroup, and 29 samples of Adenocarcinoma-rich subgroup.
2) **GBM**: Glioblastoma Multiforme (GBM) is the most common and malignant form of brain cancer and has four subtypes identified in the study by Veerhak *et al.* [5]. The subtypes are Proneural, Neural, Classical, and Mesenchymal. The data set consists of 168 samples from three genomic modalities, namely, Gene, miRNA, and CNV, as the DNA and the Protein modalities are available for a small number of samples. The data set contains 51, 24, 37, and 56 samples of Proneural, Neural, Classical, and Mesenchymal subtypes, respectively.
3) **LGG**: Lower-grade glioma (LGG) is a type of brain tumor originating from glial the cells of the brain. Three molecular subtypes of LGG have been identified in [6] by integrative genomic analysis. The LGG data set consists of 267 samples of lower-grade giloma. The subtypes are IDH mutation and no $1p/19q$ codeletion subtype, IDH mutation and $1p/19q$ codeletion subtype, and wild-type IDH subtype having 134, 84, and 49 samples, respectively.
4) **LUNG**: Based on the same primary site of origin, there are two subtypes of lung cancer in the TCGA project namely, lung adenocarcinoma and lung squamous cell carcinoma. The data set consists of 671 samples with 360 samples of lung adenocarcinoma and 311 samples of lung squamous cell carcinoma.
5) **KIDNEY**: There are three subtypes of kidney cancer under the TCGA project, namely, kidney renal clear cell carcinoma, kidney renal papillary cell carcinoma and kidney chromophobe. The data set consists of 737 samples of kidney cancer with 460 samples of kidney renal clear cell carcinoma, 214 samples of kidney renal papillary cell carcinoma, and 63 samples of kidney chromophobe.

A summary of the data sets in terms of the number of samples, the number of features in each modality, the sample-to-feature ratio, and the number of clusters is provided in Table I.

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

This 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.

TABLE I
SUMMARY OF DATA SETS

Different Data Sets	No. of Samples	No. of Features						Sample to Feature Ratio	No. of Clusters (k)
		mDNA	RNA	miRNA	Protein	CNV	Total		
CESC	124	2000	2000	311	219	2664	7194	0.017236	3
GBM	168	-	2000	534	-	2000	4534	0.037053	4
LGG	267	2000	2000	333	209	1544	6086	0.043871	3
LUNG	671	2000	2000	296	180	1572	6048	0.110945	2
KIDNEY	737	2000	2000	261	174	1544	5979	0.123264	3

B. Data Platforms and Pre-processing

For the CESC, LGG, LUNG, and KIDNEY data sets, five different modalities, namely, gene expression (RNA), DNA methylation (mDNA), copy number variation (CNV), miRNA expression (miRNA), and protein expression (Protein) are considered, while for the GBM data set three modalities namely Gene, CNV, and miRNA are considered as mDNA and Protein modalities are not available for most of the samples in the data set. In order to avoid considering features with too many missing values, for all the omic modalities, those features for which the corresponding omic expression value is not present for more than 5% of the total number of samples are excluded. For the remaining features, missing values are replaced using 0.

For all the data sets, CNV data from affymetrix SNP array 6.0 platform is used. The raw copy number segmented data is processed using the CNregions function of iCluster+ [7] R-package to reduce the redundant copy number regions. The CNregions function has a $epsilon$ parameter which denotes the maximum Euclidean distance between adjacent probes tolerated for defining a non-redundant region. The number of non-redundant copy number regions extracted for a data set depends on the value of the $epsilon$ parameter and is proportional to the number of samples in the data set. It is recomended in [7] to choose a value of $epsilon$ such that the reduced dimension is less than $10,000$. The default value of 0.005 is considered for the $epsilon$ parameter of the CNregions function for all the data sets.

For the CESC, LGG, LUNG, and KIDNEY data sets, sequence based RNA and miRNA expression data from Illumina HiSeq and Illumina GA platforms are used. The RNA and miRNA modalities contain expression signals for $20,502$ annotated genes and 1046 miRNAs, respectively. However, filtering out miRNAs with more than 5% missing values reduced the number miRNAs for the these data sets to around 300. For the GBM data set, array based gene and miRNA expression data is used. Gene expression data from three microarray platforms, namely, Affymetrix HT_HG-U133A GeneChips, and custom designed Agilent 244K arrays of G4502A_07_2 and AgilentG4502A_07_1 are used which contains $\log_2$ normalized gene expression level for $17,814$ genes. Array based miRNA expression from H-miRNA_8x15K platform is used which contains expression levels for 534 miRNAs. The underlying assumption of the proposed work is that the data follows multivariate Gaussian distribution. However, the sequence based gene and miRNA expression modalities of CESC, LGG, LUNG, and KIDNEY data sets contain normalized count data. Count data are known to follow a skewed distribution and have the property that the variance depends on the mean value [8]. It is observed that genes having larger mean expression values also tend to have larger variances and are not normally distributed. Log transformation is generally performed on the sequence based expression data to make the data more or less normally distributed [8]. The degree of normality attained depends on the skewness of the data before transformation. Therefore, for modalities with sequence based count data, the 0 entries are replaced by 1, and then the data is log-transformed using base 10.

For the all the data sets except GBM, DNA methylation beta values from two platforms Illumina Human Methylation 27K and 450K are used. Only the common set of $25,978$ CpG locations present in both the platforms are considered for each sample. Protein expression data from reverse-phase array based MDA_RPPA_Core platform is used for all the data sets which contains protein expression less than 230 annotated proteins. Finally, variance filtering is performed on the RNA and mDNA modalities of all the data sets to extract the most varying 2000 genes and CpG locations. The summary of the data sets in terms of their sample size, dimension of their individual modalities, and their number of clusters is provided in Table I.

II. STATISTICAL POWER OF DATA SETS

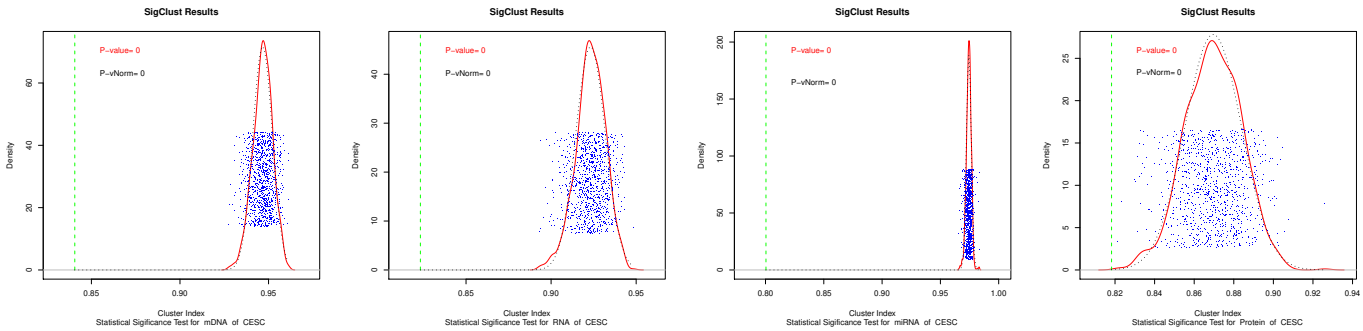
In this section hypothesis testing is performed on the multimodal data sets in order to evaluate whether clusters are "really present" in them as as opposed to being artifacts of the natural sampling variation. If the data set comes from only one Gaussian distribution, then any clustering operation that would split this data set is not significant; that is, there is no strong evidence for more than one cluster. This Gaussian null distributional assumption allows direct formulation of p values that effectively quantify significance clustering in a the data set. The SigClust algorithm [9] is used to assess the statistical significance of clustering in the multimodal data sets used in this work. The SigClust method assesses the significance of a two-way split the data set. In terms of hypothesis testing, SigClust tests the null hypothesis that the data set can be modeled as coming from a single multivariate Gaussian distribution. Accordingly, the null and the alternative hypotheses are as follows:

H_0 The data came from a single Gaussian distribution.
H_a The data came from a non-Gaussian distribution.

When H_0 is not rejected,then there is no strong evidence against the null assumption that the data came from a single Gaussian distribution. Hence, it cannot be concluded that the

TABLE II
STATISTICAL POWER OF DATA SETS

Different	p-value based on empirical quantiles (P-value)					p-value based on Gaussian quantiles (P-vNorm)				
Data Sets	mDNA	RNA	miRNA	Protein	CNV	mDNA	RNA	miRNA	Protein	CNV
CESC	0	0	0	0	-	3.063E-081	1.350E-30	0	1.704E-004	-
GBM	-	0	0	-	2.300E-002	-	5.291E-133	8.337E-020	-	1.429E-002
LGG	0	0	0	0	-	2.422E-231	1.097E-221	3.121E-096	1.135E-004	-
KIDNEY	0	0	0	4.000E-03	-	0	3.322E-240	0	4.281E-003	-



(a) mDNA (b) RNA (c) miRNA (d) Protein

Fig. 1. Evaluation of Statistical power of the modalities of CESC data set

given split of the data is real. As the test statistic, the k-means cluster index (CI), that is, the within-class sums of squares about the mean, divided by the total sum of squares about overall mean, in the case where $k = 2$. CI is given by

$$CI = \frac{\sum_{k=1}^{2} \sum_{j \in C_k} ||x_j - \bar{x}(k)||^2}{\sum_{j=1}^{n} ||x_j - \bar{x}||^2} \tag{1}$$

A lower value of CI indicates better clustering. The null distribution of the CI is approximated by simulating a single Gaussian distribution, fit to the data. For the calculation of p values the CI of the original data set is compared with the empirical distribution of the simulated CIs. The p value is the proportion of simulated CIs that are smaller than the CI for the original data set. This approach depends strongly on the number of simulated CIs. Therefore, as an alternative, it is observed that the distribution of the simulated CI usually is close to normal. Consequently, the probabilities computed using normal approximation of the distribution of simulated CI values can be used to compute the p value. The number of simulations of null distribution is taken to be $1,000$ and the level of significance is considered to be $\alpha = 0.05$. This hypothesis testing is performed on each modality of a multimodal data set to evaluate whether the cluster structure embedded in that modality is statistically significant or not.

The SigClust [9] R-package is used for to perform statistical hypothesis testing on the multimodal data sets. The p-value from the hypothesis test is calculated using both the empirical distribution of the data (denoted by P-value) as well as using the normal approximation to the distribution of simulated CI values (denoted by P-vnorm). The SigClust summary plots of simulated null distribution of CI values for different modalities of the CESC, LGG, LUNG, KIDNEY, and GBM data sets are shown in Fig. 1, 2, 3, 4, and 5, respectively. In these figures, the blue points represent the simulated CIs, while the green vertical dashed line represents the CI obtained corresponding to a 2-way partition of the original data set. The p is given by the proportion of blue dots that are present to the left of the vertical dashed line. Larger the separation between the vertical line (observed CI) and the blue dots (simulated CIs), lower is the p-value. A p-value less than 0.05 implies that the clusters present in the data set are statistically significant and are not artifacts of natural sampling variations. The p-values obtained by statistical significance test on the individual modalities of different data sets are reported in Table II. The results in Table II show that the p-values obtained for all the modalities of each data set is less than 0.05. Thus, clusters present in all the data sets are statistically significant at 5% significant level. Hence, it can be concluded that all the data sets, namely, CESC, GBM, LGG, LUNG, and KIDNEY, indicate the presence of real clusters within them as opposed to natural sampling variations.

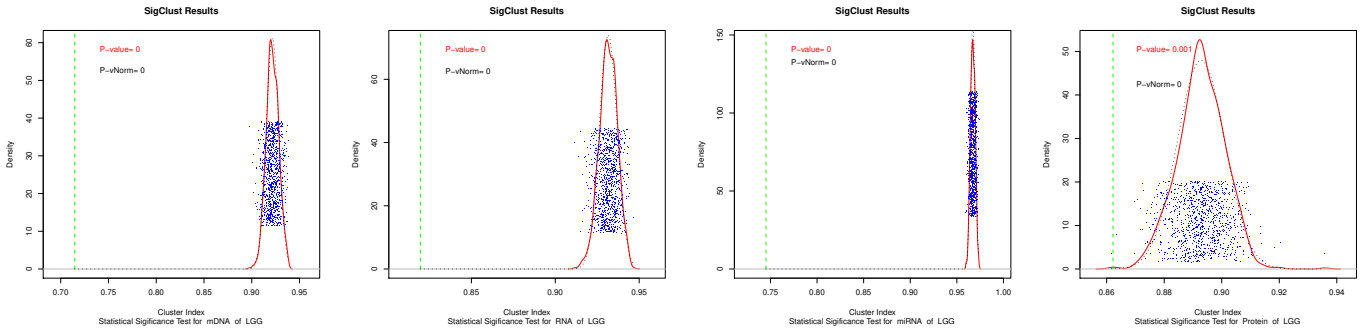
III. COMPUTATIONAL EFFICIENCY OF SURE

This section evaluates the computational complexity of the proposed algorithm and also empirically establishes its efficiency over PCA.

A. Computational Complexity of SURE

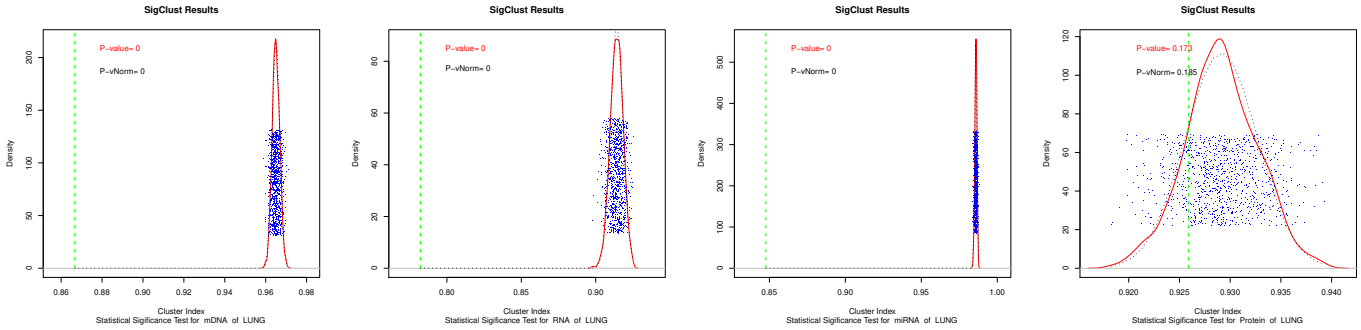
Let $X_1, \ldots, X_m, \ldots, X_M$, where $X_m \in \mathbb{R}^{n \times d_m}$, be M different modalities of a multimodal data set, all measured on the same set of n samples. Let $d_{max} = \max\{d_m\}$ and $d = \sum_{m=1}^{M} d_m$. Let the integrated data matrix $\widetilde{\mathbf{X}}_M$ obtained by concatenation of features from all the modalities be given by

$$\widetilde{\mathbf{X}}_M = \begin{bmatrix} X_1 & \ldots & X_m & \ldots & X_M \end{bmatrix}. \tag{2}$$



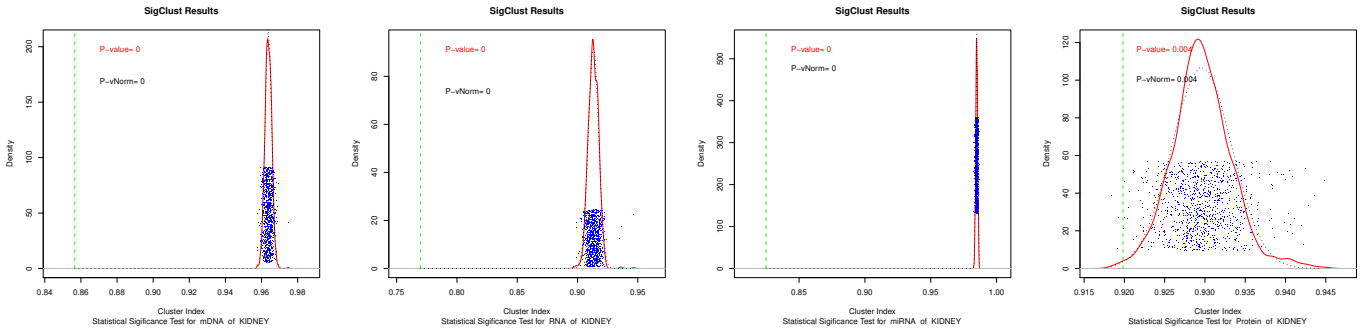
(a) mDNA (b) RNA (c) miRNA (d) Protein

Fig. 2. Evaluation of Statistical power of the modalities of LGG data set



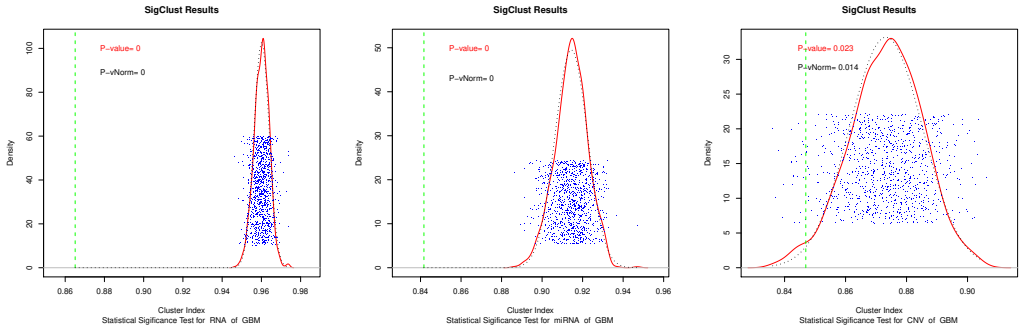
(a) mDNA (b) RNA (c) miRNA (d) Protein

Fig. 3. Evaluation of Statistical power of the modalities of LUNG data set



(a) mDNA (b) RNA (c) miRNA (d) Protein

Fig. 4. Evaluation of Statistical power of the modalities of KIDNEY data set



(a) RNA (b) miRNA (c) Protein

Fig. 5. Evaluation of Statistical power of the modalities of GBM data set

The proposed SURE algorithm to construct the joint eigenspace $\Psi(\widetilde{\mathbf{X}}_M)$ of the integrated data is described in Section III-C of the main article. The computational complexity of the SURE algorithm is analyzed as follows: The exact SVD of a $(n \times z)$ matrix has time complexity of $\mathcal{O}(\min\{nz^2, n^2z\})$. In the proposed algorithm, for each modality X_m, a SVD problem of size $n \times d_m$ is solved in step 2. The SVD problems on the individual modalities are independent of each other and can be computed parallely for all the modalities. This time complexity is bounded by the time required for the largest modality, that is, $\mathcal{O}(\min\{nd_{max}^2, n^2 d_{max}\}) = \mathcal{O}(n^2 d_{max})$, assuming $n < d_{max}$ due to the high dimension low sample size nature of the data sets. Similarly performing k-means on the left subspace $U(X_m)$ of X_m and computation of its relevance $\mathrm{Rel}(X_m)$ from the clustering solution, can be done for all the modalities in parallel. k-means clustering on $n \times k$ matrix $U(X_m)$ has time complexity of $\mathcal{O}(t_{max} n k^2)$, where t_{max} is the maximum number of iterations the k-means algorithm runs and $k << n$. Computation of $\mathrm{Rel}(X_m)$ takes $\mathcal{O}(n)$ time, owing to the computation of within-cluster variance in $U(X_m)$. Thus, for M modalities, the time complexity of Steps 1-5 is bounded by the that of the largest modality, that is $\left(\mathcal{O}\left(n^2 d_{max} + t_{max} n k^2 + n\right) =\right) \mathcal{O}(n^2 d_{max})$.

After computation of individual eigenspaces in Steps 1-5, concordance C between every pair of modalities is computed in Step 6. This involves computation of normalized mutual information which takes $\mathcal{O}(k^2)$ time. Step 7, has time complexity of $\mathcal{O}(M)$ to find the modality with maximum relevance. Steps 8-9 are assignments operations which take $\mathcal{O}(1)$ time. For $(M-1)$ remaining modalities, the loop in step 10 can execute at most $(M-1)$ times. On m-th execution of the loop, there are $(M-m)$ candidate modalities for updation. For each candidate modality, its average concordance $\bar{\mathsf{C}}$ with the formerly updated ones is computed in Step 12. This has a complexity of $\mathcal{O}(m)$. For $(M-m)$ candidate modalities, the total complexity of Steps 11- 13 is $\mathcal{O}(m(M-m))$. The one with maximum average concordance is chosen in $\mathcal{O}(M-m)$ time. If its average concordance $\bar{\mathsf{C}}$ is greater than threshold τ then the eigenspace is updated in steps 16-27.

For eigenspace updation, Steps 17-19 consist of concatenation and union operations which take at most $\mathcal{O}(d_{max})$ time. Step 20 takes $\mathcal{O}(nk^2)$ time to compute the matrices $\mathcal{I}$, $\mathcal{P}$ and $\mathcal{Q}$. The Gram-Schmidt orthogonalization in step 21 has complexity of $\mathcal{O}(nk^2)$ for $n \times k$ matrix $\mathcal{Q}$. To find t in step 22, the norm of the columns of $\mathcal{Q}$ is computed, which takes $\mathcal{O}(nk)$ time. Step 24 requires solving the SVD problem of (23) of the main article, which is of size at most $2k \times d$ and has time complexity of $\mathcal{O}(k^2 d)$. $U(\widetilde{\mathbf{X}}_{m+1})$ in step 25 computed in $\mathcal{O}(nk^2)$ time. Steps 26 and 27 have constant complexity of $\mathcal{O}(1)$. Hence, the total complexity of steps 16-27 for updating the eigenspace is $\left(\mathcal{O}(d_{max} + nk^2 + nk + nk^2 + k^2 d + nk^2) =\right) \mathcal{O}(k^2 d)$. Therefore, time complexity of updating the eigenspace in m-th iteration of the loop in Step 10 is $\left(\mathcal{O}(m(M-m) + k^2 d) =\right) \mathcal{O}(k^2 d)$. Step 10 is executed at most $(M-1)$ times which gives a total complexity of $\mathcal{O}(Mk^2 d)$. Hence, the overall computational complexity of the proposed SURE algorithm, to extract the joint eigenspace of the integrated data is $\left(\mathcal{O}(n^2 d_{max} + M + k^2 + Mk^2 d) =\right) \mathcal{O}(n^2 d_{max})$, assuming $M, k << n < d_{max}$. Thus the complexity of the proposed algorithm is bounded by that of individual eigenspace construction in Steps 1-5.

The comparative execution time of the proposed SURE algorithm with PCA computed using eigen value decomposition (EVD) of the covariance matrix is shown is Fig. 6, while that using SVD of mean-centered data matrix is shown in Fig. 7, for CESC and LGG data sets. The plots in Fig. 6 show that the execution time of PCA-NI computed using EVD increases quadratically with respect the proposed SURE approach. This is because PCA-NI using EVD takes $\mathcal{O}(d^3)$ time which is significantly higher compared to $\mathcal{O}(n^2 d_{max})$. Fig. 7 shows that the execution time of PCA using SVD as well as of the proposed SURE algorithm increases linearly with increase in number of features. However, SURE takes significantly lesser time to extract the principal components as compared to PCA-NI using SVD as seen in Fig. 7.

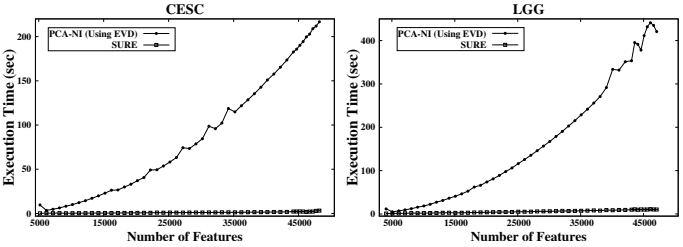


Fig. 6. Comparison of execution time for PCA-NI computed using EVD and the proposed SURE approach on CESC and LGG data sets

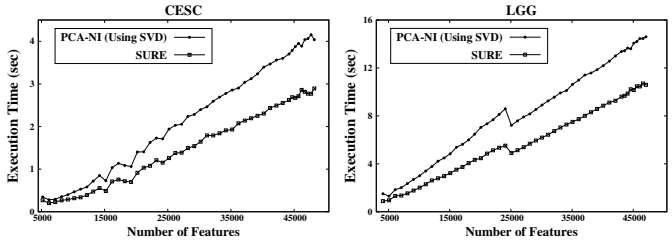


Fig. 7. Comparison of average execution time for PCA-NI computed using SVD and the proposed SURE approach on CESC and LGG data sets

IV. Wedin's $\sin\Theta$ Theorem

Matrix perturbation theory [10] is used to estimate how the spectrum of a matrix changes when it is subjected to perturbations. More specifically, for an approximately low-rank matrix A and a perturbation matrix E, perturbation theory analyzes how much the left and right singular subspaces of A and $\widetilde{A} = A + E$ differ from each other. Wedin's $\sin\Theta$ theorem [11] provides perturbation bounds for both the left and right singular subspaces in terms of the gap between singular values and the perturbation level.

Let A and $\widetilde{A}$ be two complex $(n \times d)$ matrices with conformally partitioned SVDs as follows:

$$A = \begin{bmatrix} U_1 & U_2 & U_3 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^*; \tag{3}$$

$$\widetilde{A} = \begin{bmatrix} \widetilde{U}_1 & \widetilde{U}_2 & \widetilde{U}_3 \end{bmatrix} \begin{bmatrix} \widetilde{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \widetilde{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \widetilde{V}_1 & \widetilde{V}_2 \end{bmatrix}^*, \tag{4}$$

where $U_1, \widetilde{U}_1 \in \mathbb{C}^{n\times r}$, $V_1, \widetilde{V}_1 \in \mathbb{C}^{d\times r}$, and

$$\Sigma_1 = diag(\sigma_1, \ldots, \sigma_r), \quad \Sigma_2 = diag(\sigma_{r+1}, \ldots, \sigma_n), \tag{5}$$

$$\widetilde{\Sigma}_1 = diag(\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_r), \quad \widetilde{\Sigma}_2 = diag(\widetilde{\sigma}_{r+1}, \ldots, \widetilde{\sigma}_n). \tag{6}$$

B^* is the conjugate transpose of B for any matrix B. There is no particular assumption about the order of the singular values. Let $\mathcal{C}(B)$ denote the column space of B, while $\| B \|_F^2$ denotes the squared Frobenius norm of B. Let $\Theta(U_1, \widetilde{U}_1)$ be the set of principal angles between the left subspaces $\mathcal{C}(U_1)$ and $\mathcal{C}(\widetilde{U}_1)$, while $\Theta(V_1, \widetilde{V}_1)$ be that between the pair of right subspaces $\mathcal{C}(V_1)$ and $\mathcal{C}(\widetilde{V}_1)$. Let the sum of squared principal sines between the pair of left and right subspaces be given by $\| \sin\Theta(U_1, \widetilde{U}_1) \|_F^2$ and $\| \sin\Phi(V_1, \widetilde{V}_1) \|_F^2$, respectively. Let the following residuals be defined as:

$$\mathbb{R}_L = A\widetilde{V}_1 - \widetilde{U}_1\Sigma_1 = (A - \widetilde{A})\widetilde{V}_1 \quad \text{and} \tag{7}$$

$$\mathbb{R}_R = A^*\widetilde{U}_1 - \widetilde{V}_1\Sigma_1 = (A^* - \widetilde{A}^*)\widetilde{U}_1. \tag{8}$$

Theorem 1. ***Wedin's*** $\sin\Theta$ ***Theorem*** *[11] Let A and $\widetilde{A}$ be two complex matrices with SVDs partitioned as in (3) and (4), respectively. If*

$$\delta \stackrel{\text{def}}{=} \min \left\{ \min_{1\le i\le r, 1\le j\le (n-r)} |\sigma_i - \widetilde{\sigma}_{r+j}|, \min_{1\le i\le r} \sigma_i \right\} > 0, \tag{9}$$

$$\textit{then} \quad \sqrt{\| \sin\Theta(U_1, \widetilde{U}_1) \|_F^2 + \| \sin\Theta(V_1, \widetilde{V}_1) \|_F^2} \le \frac{\sqrt{\| \mathbb{R}_L \|_F^2 + \| \mathbb{R}_R \|_F^2}}{\delta}. \tag{10}$$

V. External Cluster Evaluation Indices

Let $\mathcal{T} = \{t_1, \ldots, t_i, \ldots, t_k\}$ be the true partition of n samples of a data set into k clusters. Let $\mathcal{C} = \{c_1, \ldots, c_j, \ldots, c_k\}$ be the k clusters returned by a clustering algorithm. The external evaluation indices measure how close is the clustering $\mathcal{C}$ with respect to true partition $\mathcal{T}$. The external evaluation indices used in this work are defined as follows.

1) F-measure [12] of a cluster c_i with respect to a class t_j assess how well cluster cluster c_i describes class t_j and is given by the harmonic mean of precision and recall.

$$Precision\ P_{ij} = \frac{|c_i \cap t_j|}{|c_i|}. \tag{11}$$

$$Recall\ R_{ij} = \frac{|c_i \cap t_j|}{|t_j|}. \tag{12}$$

$$F - measure\ \mathcal{F}(t_j, c_i) = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}} \tag{13}$$

$$= \frac{2|c_i \cap t_j|}{|c_i| + |t_j|}. \tag{14}$$

The overall F-measure is given by the weighted average of the maximum F-measure over the clusters in $\mathcal{C}$.

$$F{-}measure\,(\mathcal{C}, \mathcal{T}) = \frac{1}{n}\sum_{j=1}^{k} n_j \max_i \{\mathcal{F}(t_j, c_i)\}. \tag{15}$$

2) Adjusted Rand Index (ARI) [13] is an adjustment of the rand index, given by,

$$ARI\,(\mathcal{C}, \mathcal{T}) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} \binom{|c_i \cap t_j|}{2} - n_3}{\frac{1}{2}(n_1 + n_2) - n_3}. \tag{16}$$

where $n_1 = \sum_{i=1}^{k} \binom{|c_i|}{2}$, $n_2 = \sum_{j=1}^{k} \binom{|t_j|}{2}$, $n_3 = \frac{2n_1n_2}{n(n-1)}$.

3) Purity [14] measures the extent to which each cluster contains samples primarily from one class. Each cluster is first assigned with the true class which is most frequent in the cluster and then the purity of the clustering solution is assessed by the proportion of correctly assigned samples. Formally it is given by,

$$Purity\,(\mathcal{C}, \mathcal{T}) = \frac{1}{n}\sum_{i=1}^{k} \max_j \{|c_i \cap t_j|\}. \tag{17}$$

In general, higher the value of purity, better is the cluster solution. However, purity does not penalize large number of clusters.

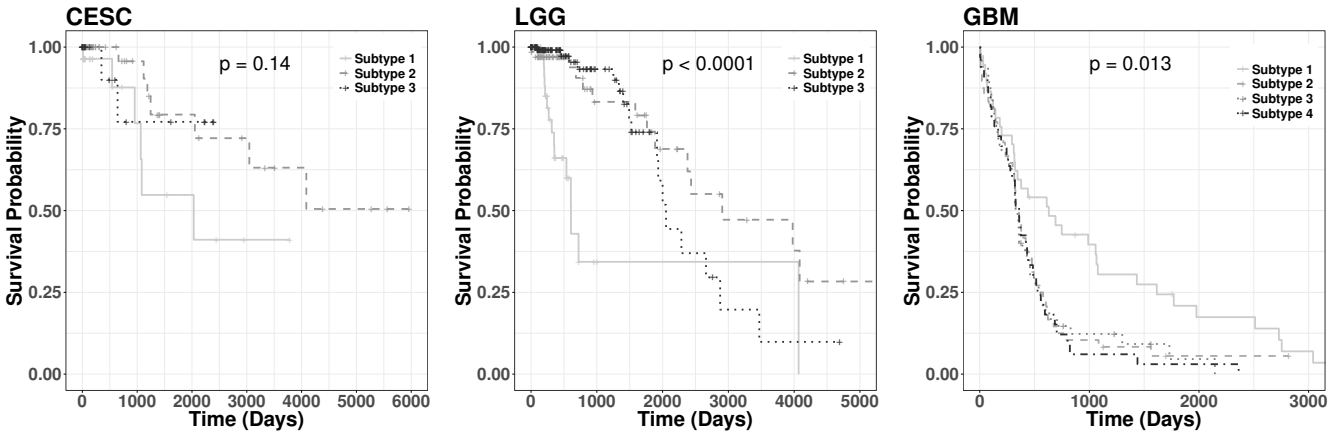
4) Information theoretic measure : The Normalized Mutual Information (NMI) [15] which is the concordance measure used in the proposed SURE approach is used measure the concordance of cluster assignments in $\mathcal{T}$ and $\mathcal{C}$.

The value of ARI lies in $[-1, 1]$, while, the value of the other indices, namely, purity, F-measure and NMI lies in $[0, 1]$. For all the indices, a value closer to one indicates better clustering.

VI. Survival Analysis

Clinical information of the samples is used to analyze the survival profiles of the subtypes identified by the proposed SURE algorithm on different data sets. The survival profiles of the subtypes in a data set are compared using Kaplan-Meier survival plots, median survival times of the subtypes, survival probability of the samples in a subtype after two, three, and seven years of diagnosis of the disease, and log-rank test p-value from pairwise comparison of subtypes. Median survival time is a statistic that refers to how long patients are expected to survive with a disease. It is the time expressed in months or years, when half of the patients in a group of patients diagnosed with the disease are still alive. It gives an approximate indication of the survival as well as the prognosis of a group of patients with the disease. The median survival time for a disease subtype is given by the time period where the Kaplan-Meier curve for the subtype crosses the survival probability of 0.5, and it is not available for subtypes whose survival curves end before the survival probability of 0.5 due to low sample count or presence of censored samples. The total number of deaths in each subtype, the number of samples at risk and the number of events of death at two, three and seven years of diagnosis is also observed to study the prognosis of respective cancer with time.

The Kaplan-Meier plot and survival analysis results for the LGG data set are given in Fig. 8b and Table III, respectively. The p-values for the log-rank test and the generalized Wilcoxon test are $2.125E-07$ and $5.901E-09$,



(a) (b) (c)

Fig. 8. Kaplan-Meier survival plots for subtypes identified by SURE on LGG, CESC, and GBM data sets

TABLE III
SURVIVAL ANALYSIS FOR SUBTYPES IDENTIFIED BY SURE ON THE LGG DATA SET

Different Subtypes	No. Of Samples	Total No. Of Deaths	Median Survival Time (Years)	Time (Years)	No. Of Risks	No. Of Events Of Death	Survival Probability	Standard Error (in probabilities)
Subtype 1	51	15	1.66	2	3	14	0.343	0.126
				5	1	0	0.343	0.126
				7	1	0	0.343	0.126
Subtype 2	73	14	7.96	2	28	4	0.906	0.0482
				5	15	4	0.741	0.0853
				7	8	3	0.551	0.1152
Subtype 3	143	17	5.62	2	43	4	0.933	0.0340
				5	10	5	0.740	0.0825
				7	5	5	0.370	0.1240

TABLE IV
SURVIVAL ANALYSIS FOR SUBTYPES IDENTIFIED BY SURE ON THE CESC DATA SET

Different Subtypes	No. Of Samples	Total No. Of Deaths	Median Survival Time	Time (Years)	No. Of Risks	No. Of Events Of Death	Survival Probability	Standard Error (in probabilities)
Subtype 1	33	6	5.57	2	8	2	0.877	0.0895
				5	4	3	0.548	0.1601
				7	2	1	0.411	0.1688
Subtype 2	70	7	NA	2	21	1	0.957	0.0425
				5	11	3	0.794	0.0928
				7	9	1	0.721	0.1089
Subtype 3	21	2	NA	2	6	2	0.771	0.144
				5	4	0	0.771	0.144

respectively. These p-values show that there is a statistically significant difference in survival profiles of the subtypes of LGG, identified by the SURE algorithm. Table III shows that subtype 2 and subtype 3 have median survival times of 7.96 and 5.62 years, respectively. Hence, subtype 2 and Subtype 3 have much better prognosis than subtype 1 which has survival time of 1.66 years. The survival risk is also very high for subtype 1, as the number of death is 15 out of 51 samples and the survival probability is only 0.343 at two, three, and five years of diagnosis. The p-value from pairwise log-rank test comparing subtypes 1 and 2 is $5.117E-05$, comparing subtypes 1 and 3 is $5.915E-06$, while the p-value between subtypes 2 and 3 is 0.32947. Thus, the difference between survival profiles of subtypes 1 and 2 and subtypes 1 and 3 are statistically significant, while the difference is not statistically significant between subtypes 2 and 3. Both the subtypes 2 and 3 have similar survival probabilities at two and five years of diagnosis. However, the survival probability for subtype 3 is 0.370 which is very low compared to subtype 3 having probability 0.551 after seven years of diagnosis of cancer.

The survival plots and analysis for the CESC data set are given in Fig. 8a and Table IV, respectively. The median survival time is not reached for subtypes 2 and 3, while for subtype 1 the median survival time is 5.57 years. Moreover,

TABLE V
SURVIVAL ANALYSIS FOR SUBTYPES IDENTIFIED BY SURE ON THE GBM DATA SET

Different Subtypes	No. Of Samples	Total No. Of Deaths	Median Survival Time (Years)	Time (Years)	No. Of Risks	No. Of Events Of Death	Survival Probability	Standard Error (in probabilities)
Subtype 1	37	34	1.726	2	16	20	0.455	0.0825
				5	6	8	0.209	0.0707
				7	4	2	0.139	0.0620
Subtype 2	48	45	0.944	2	6	42	0.125	0.0477
				5	1	3	0.055	0.0350
				7	1	0	0.055	0.0350
Subtype 3	50	46	0.921	2	7	42	0.147	0.0511
				5	1	3	0.046	0.0395
Subtype 4	33	33	0.984	2	4	29	0.1212	0.0568
				5	1	3	0.0303	0.0298

subtypes 2 and 3 have 7 and 2 deaths out of 70 and 21 samples, respectively. On the other hand, subtype 1 has 3 death cases out of 33 samples. The survival probability after seven years of diagnosis is only 0.411 for subtype 1, while the probabilities are 0.721 and 0.771 for subtypes 2 and 3, respectively. These results show that subtype 2 and subtype 3 have better prognosis compared to subtype 1. The pairwise log-rank test p-values for subtypes 1 and 2 is 0.04712, for subtypes 1 and 3 is 0.29749, and for subtypes 2 and 3 is 0.78188. The difference in survival profiles is statistically significant only for subtypes 1 and 2 and is not significant for other pairs.

Table V reports the survival analysis results for the GBM data set and the Kaplan-Meier plot for the GBM subtypes identified by the proposed SURE approach is given in 8c. For the GBM data set, the overall log-rank p-value is 0.0137, which shows that the subtypes have significant difference in their survival profiles. The median survival times for subtypes 1, 2, 3, and 4 are 1.726, 0.944, 0.921, and 0.984, respectively. Comparative results from survival analysis of other data sets show that the GBM subtypes have significantly poor prognosis compared to subtypes of other cancers. Moreover, across all the subtypes the number of deaths is very close to the total number of samples. Death rate is most severe for subtype 4, where death occurs for all the 33 samples of the subtype. The p-values for pairwise log-rank test for subtypes 1 and 2 is 0.01413, for subtypes 1 and 3 is 0.00743, and for subtypes 1 and 4 is 0.00290. The pairwise survival difference between subtype 1 and the other subtypes is statistically significant. On the other hand, the pairwise log-rank test p-values for subtypes 2 and 3 is 0.95869, for subtypes 2 and 4 is 0.71164, and for subtypes 3 and 4 is 0.86016, which show no significant difference among survival profiles of subtypes 2, 3, and 4.

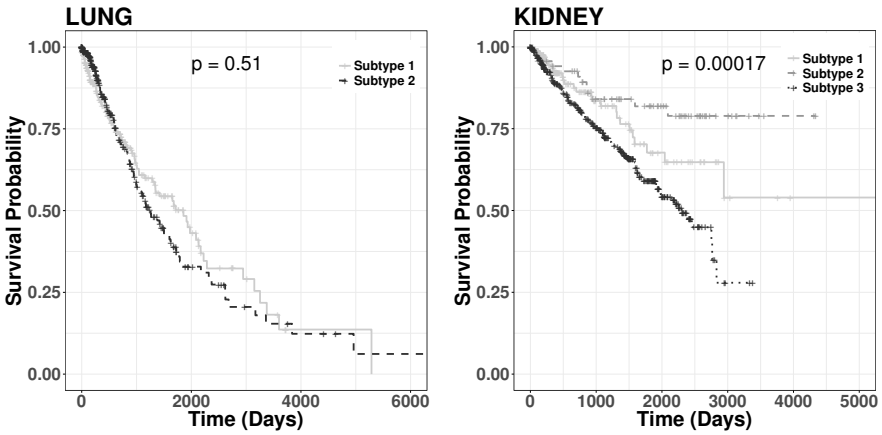
The Kaplan-Meier plot and survival analysis results for the LGG data set are given in Fig. 9a and Table VI, respectively. The median survival time for subtype 1 is 5.08 years, while for subtype 2 the median survival time is worse, that is, 3.45 years. The log-rank p-value for survival difference is 0.51, which does not show statistical significance. However, the survival probabilities for subtype 1 and subtype 2 after five years of diagnosis is 0.501 and 0.329, respectively, and after seven years of diagnosis, the survival probabilities are 0.323 and 0.274, respectively. This shows increased survival risk for subtype 2 compared to subtype 1.

For the KIDNEY data set, the survival curves are plotted in Fig. 9b and the results are reported in Table VII. In the KIDNEY data set, for both the subtypes 1 and 2, the survival curves end before the median survival probability of 0.5. Moreover, the survival probabilities for subtypes 1 and 2 after seven years of diagnosis are 0.648 and 0.789, respectively, while for subtype 3, this probability drops to 0.449. This indicates that subtypes 1 and 2 have better prognosis than compared to subtype 3 which has a median survival time of 6.3 years. The p-value from pairwise log-rank test comparing subtypes 1 and 2 is 0.124566, comparing subtypes 1 and 3 is 0.01657646, and for subtypes 2 and 3 is 0.0001816. The p-values show statistical significance when compared between subtypes 3 and 1 and between subtypes 3 and 2. The overall log-rank p-value is 0.00017 when the profiles of all the three subtypes are compared together, which is statistically significant.

VII. EXPERIMENTAL SETUP

In the proposed algorithm, concordance between two modalities is computed in terms of NMI between the cluster assignments of two modalities. However, real-life omic modalities like gene expression, DNA methylation, protein expression are highly heterogeneous in nature in terms of unit, variance, and scale, and are likely to have very disparate cluster structures. The pairwise NMI values for real-life data sets are usually less than 0.5 indicating low concordance. So, in Step 12 of the proposed algorithm, the pairwise concordance values are rescaled to have maximum concordance of 1 between two distinct modalities. Moreover, the maximum and minimum concordance values differ for different data sets. Rescaling the maximum concordance to 1 gives a uniform interpretation of the concordance threshold τ across different data sets. The minimum concordance, however, has not been transformed to 0 as that would imply completely disparate cluster structure with no concordance at all between the respective modalities.

The performance of clustering on the joint subspace extracted by the proposed algorithm is compared with two two-stage clustering approaches, namely, Bayesian consensus cluster (BCC) [16] and cluster of cluster analysis (COCA) [17], and five low-rank direct integrative clustering approaches,



(a) Survival plots for subtypes of LUNG (b) Survival plots for subtypes of KIDNEY

Fig. 9. Kaplan-Meier survival plots for subtypes identified by SURE on LUNG and KIDNEY data sets

TABLE VI
SURVIVAL ANALYSIS FOR SUBTYPES IDENTIFIED BY SURE ON THE LUNG DATA SET

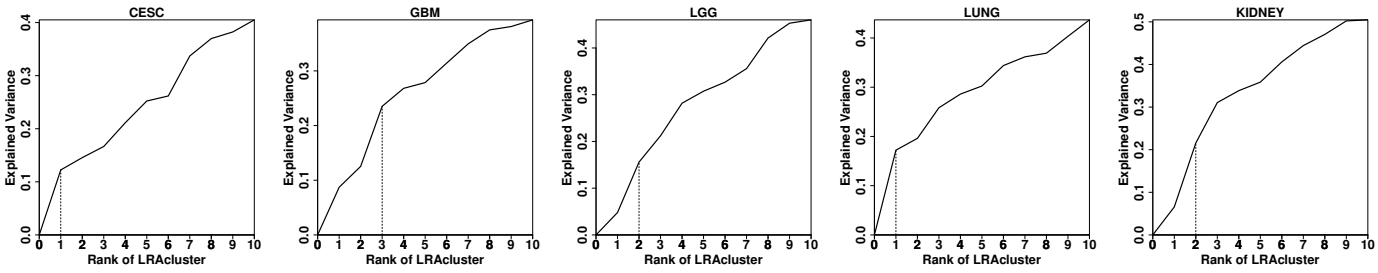
Different Subtypes	No. Of Samples	Total No. Of Deaths	Median Survival Time (Years)	Time (Years)	No. Of Risks	No. Of Events Of Death	Survival Probability	Standard Error (in probabilities)
Subtype 1	285	86	5.08	2	92	50	0.717	0.0350
				5	31	21	0.501	0.0476
				7	13	9	0.323	0.0575
Subtype 2	363	105	3.45	2	98	56	0.703	0.0348
				5	22	39	0.329	0.0473
				7	13	3	0.274	0.0488

TABLE VII
SURVIVAL ANALYSIS FOR SUBTYPES IDENTIFIED BY SURE ON THE KIDNEY DATA SET

Different Subtypes	No. Of Samples	Total No. Of Deaths	Median Survival Time (Years)	Time (Years)	No. Of Risks	No. Of Events Of Death	Survival Probability	Standard Error (in probabilities)
Subtype 1	214	28	NA	2	73	16	0.864	0.0328
				5	27	10	0.677	0.0595
				7	13	1	0.648	0.0638
Subtype 2	74	12	NA	2	55	6	0.909	0.0356
				5	35	5	0.818	0.0503
				7	18	1	0.789	0.0563
Subtype 3	445	140	6.3	2	263	70	0.811	0.0205
				5	91	55	0.591	0.0304
				7	14	12	0.449	0.0462

namely, joint and individual variation explained (JIVE) [18], iCluster [19], iCluster+ [20], LRAcluster [21], and PCA [22] on the naively concatenated data (PCA-NI). The experimental setup used for these algorithms is briefly outlined as follows:

- **BCC** [16]: The BCC approach uses Bayesian framework for simultaneous estimation of the overall consensus and source-specific clusterings. It assumes Dirichlet distribution for the prior probabilities of the k clusters and uses Gibbs sampling to estimate the posterior distribution of the model parameters and, the overall and source-specific clusterings. However, each Markov Chain Monte Carlo (MCMC) iteration of Gibbs sampling produces different realizations of the consensus and source specific clusters. The R code developed by the authors which is available at *http://ericfrazerlock.com/Software.html* is used to study the performance of BCC. The BCC algorithm is executed at the default setting which uses 1000 MCMC draws and initialization of the source-specific clusters using k-means. The values of hyper-parameters a and b in the $Beta(a, b)$ prior distribution on the model parameter α are assigned values 1 and 1, respectively, under the default setting. The BCC algorithm additionally has Dirichlet prior concentration parameter β_0 having default value of 1. However this default value yields less than k clusters. Therefore, the prior concentration parameter β_0 is varied between 1 to 10, where higher value of β_0 favors larger number of clusters and more equal proportions for each cluster. The optimal values of β_0 is selected



(a) (b) (c) (d) (e)

Fig. 10. Optimal rank estimation of LRAcluster for different data sets

using an adherence based statistic $\alpha*$, as proposed by the authors. The optimal concentration parameter β_0 selected for CESC, GBM, LGG, LUNG, and KIDNEY data sets are 6, 6, 4, 4, and 10, respectively.

- **COCA** [17]: For the COCA approach, k-means clustering is first performed on each modality separately with k clusters. Clusters identified from each modality are encoded into a series of indicator variables for each cluster. Consensus clustering [23] is then performed on the indicator matrix of 0's and 1's using ConsensusClusterPlus R package [24] version $1.40.0$. Parameters used for consensus clustering are 80% sample resampling with 1000 iterations of hierarchical clustering based on a Pearson correlation distance metric, as suggested in [17].
- **JIVE** [18]: The JIVE algorithm extracts two low-rank representations for each modality, one encodes the shared joint structure, while the other encodes modality specific structure. The ranks of the joint and the individual structures are automatically determined using two different criteria: one based on permutation test (PERM), and the other based on Bayesian information criteria (BIC). After obtaining the joint rank, say j, and the joint and individual structures for each modality, the integrated joint structure from all the modalities is obtained by concatenating the j largest principal components of the joint structure obtained from each of the modalities. Then k-means clustering is performed on the integrated joint structure to get the final clusters. The joint and individual ranks obtained by the JIVE algorithm using the permutation and BIC based rank selection criteria are given in Table VIII for different data sets.
- **iCluster** [19]: This is a low-rank based approach which uses Gaussian latent variable model to extract a $(k-1)$ dimensional joint subspace of a multimodal data set, where k is the number of clusters in the data set. The k-means clustering is performed in the $(k-1)$ dimensional joint subspace extracted by the iCluster algorithm Hence, the dimensions of low-rank subspaces extracted by iCluster for CESC, GBM, LGG, LUNG, and KIDNEY data sets are 2, 3, 2, 1, and 3, respectively. The iCluster R-package available at *https://CRAN.R-project.org/package=iCluster* is used to evaluate the performance of the iCluster algorithm. For each modality, iCluster has a lasso penalty parameter (λ), which varies between 0 and 1. The value 0 represents the non-sparse solution where all features are selected, while 1 represents the null model where no features are included. The optimal value of λ is selected using the proportion of deviance (POD) statistic [19]. The POD statistic lies between 0 and 1. Small values of POD indicate strong cluster separability, and large values of POD indicate poor cluster separability. The value of λ that minimizes the POD statistic is selected to be the optimal one. The uniform sampling design (UD) approach of Fang and Wang [25] is used to generate different combination of λ values that are scattered uniformly across the search domain as suggested in [26]
- **iCluster+** [20]: iCluster+ extracts a $(k-1)$ dimensional low-rank subspace of a multimodal data set. It uses different distributions to model the different modalities of a multimodal data set. As suggested by the authors, Gaussian distribution is used to model the real-valued array based mDNA, CNV, and Protein modalities of CESC, LGG, LUNG, and KIDNEY data sets, while Poisson distribution is used to model the count based RNA and miRNA modalities. For the GBM data set, all the modalities are observed on array based platforms, so Gaussian distribution is used to model them. The iCluster+ R-package [7] is used to used to study the performance of iCluster+. The default parameter setting is used for the iCluster+ algorithm for all the data sets.
- **LRAcluster** [21]: This is a low-rank based approach which models each modality of a multimodal data set using a separate probability distribution having its own set of parameters. Similar to iCluster+, Gaussian distribution is used to model the real-valued array based mDNA, CNV, and Protein modalities of CESC, LGG, LUNG, and KIDNEY data sets, and all the modalities of GBM dat set. Poisson distribution is used to model the count based RNA and miRNA modalities of CESC, LGG, LUNG, and KIDNEY data sets. For LRAcluster, the rank of the lower dimensional subspace is optimized using the likelihood based "explained variation" criteria [21], as suggested by the authors. According to this criteria, the value of explained variance is observed for different values of rank varying between 0 to 10. The optimal value of rank is chosen to be the one having the maximum change in explained variance. The change in explained variance for different values of rank is given in Fig. 10 for different data sets. Based on this criteria, the optimal rank obtained for the CESC, GBM, LGG, LUNG, and KIDNEY data

TABLE VIII
JOINT AND INDIVIDUAL RANKS OBTAINED BY JIVE ALGORITHM

Different Datasets	Algorithm	Joint Rank	Individual Ranks					Algorithm	Joint Rank	Individual Ranks				
			mDNA	RNA	miRNA	Protein	CNV			mDNA	RNA	miRNA	Protein	CNV
CESC	JIVE (PERM)	5	15	21	13	10	-	JIVE (BIC)	1	1	0	1	1	-
GBM		4	-	27	19	-	35		0	-	-	-	-	-
LGG		2	12	23	18	12	-		2	1	2	2	2	-
LUNG		1	30	29	22	16	-		2	4	5	2	2	-
KIDNEY		3	30	41	24	17	-		3	4	4	2	1	-

sets are 1, 3, 2, 1, and 2, respectively. After obtaining the optimal low-rank subspace, k-means clustering is performed in that subspace to identify the clusters.

- **PCA-NI** [22]: In the PCA-NI approach, genomic features from all the modalities are concatenated and then PCA is performed on the concatenated data to extract the principal subspace. For a comparative study, the number of principal components considered for PCA-NI approach is same as the dimension of the joint subspace extracted by the proposed approach, that is, the number of clusters k. For all the low-rank based approaches, namely, JIVE, iCluster, iCluster+, LRAcluster, PCA-NI, and the proposed approach, k-means clustering is performed 30 times and the cluster solution corresponding to the minimum objective function is used for comparative analysis.

The BCC and iCluster+ algorithms use Gibbs sampling with MCMC iterations, while COCA uses resampling based consensus clustering technique to find the final joint clusters. The results of BCC, iCluster+, and COCA algorithms can vary on different executions of these algorithms. So, the average performance of these algorithms over 10 executions is reported in this work.

REFERENCES

[1] TCGA Research Network, "http://cancergenome.nih.gov/,"

[2] GDC Data Portal, "https://gdc-portal.nci.nih.gov/,"

[3] M. Kosinski, *RTCGA.clinical: Clinical datasets from The Cancer Genome Atlas Project*, 2016. R package version 20151101.6.0.

[4] TCGA Research Network, "Integrated genomic and molecular characterization of cervical cancer," *Nature*, vol. 543, pp. 378–384, 2017.

[5] R. G. W. Verhaak, K. A. Hoadley, *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, no. 17, pp. 98–110, 2010.

[6] TCGA Research Network, "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas," *The New England Journal of Medicine*, vol. 372, no. 26, pp. 2481–2498, 2015.

[7] Q. Mo and R. Shen, *iClusterPlus: Integrative clustering of multi-type genomic data*, 2016. R package version 1.12.1.

[8] I. Zwiener, B. Frisch, and H. Binder, "Transforming rna-seq data to improve the performance of prognostic gene signatures," *PloS one*, vol. 9, no. 1, p. e85150, 2014.

[9] H. Huang, Y. Liu, M. Yuan, and J. S. Marron, "Statistical Significance of Clustering using Soft Thresholding," *J Comput Graph Stat*, vol. 24, no. 4, pp. 975–993, 2015.

[10] G. W. Stewart and J.-g. Sun, *Matrix perturbation theory*. Academic press New York, 1990.

[11] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.

[12] B. Larsen and C. Aone, "Fast and effective text mining using linear time document clustering," in *In Proc. Knowledge Discovery and Data mining*, (San Diego, USA), pp. 16–22, 1999.

[13] P. Arabie and L. Hubert, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[14] Y. Zhao and G. Karypis, "Criterion functions for document clustering: Experiments and analysis," *Technical Report #01-40, University of Minnesota*, 2001.

[15] A. L. Fred and A. K. Jain, "Robust data clustering," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 3, pp. 128–136, 2003.

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

[17] 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.

[18] E. F. Lock *et al.*, "Joint and individual variation explained (jive) for integrated analysis of multiple data types," *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 523–542, 2013.

[19] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, no. 25(22), pp. 2906–2912, 2009.

[20] Q. Mo, S. Wang, *et al.*, "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 11, pp. 4245–4250, 2013.

[21] 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*, 2015.

[22] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 18, pp. 10101–10106, 2000.

[23] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, pp. 91–118, 2003.

[24] Wilkerson, M. D., Hayes, and D. Neil, "Consensusclusterplus: a class discovery tool with confidence assessments and item tracking," *Bioinformatics*, vol. 26, no. 12, pp. 1572–1573, 2010.

[25] K. T. Fang and Y. Wang, *Number-Theoretic Methods in Statistics*. Chapman and Hall/CRC, 1993.

[26] R. Shen, Q. Mo, N. Schultz, V. E. Seshan, A. B. Olshen, J. Huse, M. Ladanyi, and C. Sander, "Integrative subtype discovery in glioblastoma using icluster," *PloS one*, vol. 7, no. 4, p. e35236, 2012.