Appendix: Selective Update 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 Update of Relevant Eigenspaces), to construct a low-rank joint subspace of a high dimensional multimodal data set. This appendix file describes the multimodal data sets and the experimental setup used in this work, along with survival analyses of the cancer subtypes identified by the proposed algorithm. It also outlines Wedin's theorem, used in the main paper to derive error bound on the principal sines between full-rank and approximate eigenspaces.

I. Experimental Setup

This section describes the five multimodal cancer data sets, their pre-processing steps and statistical power, and the experimental setup used for the existing integrative clustering algorithms.

A. Description of Data Sets

Five multimodal cancer data sets from The Cancer Genome Atlas (TCGA) [1] are 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 (mDNA), gene expression (RNA), miRNA expression (miRNA), reverse phase protein array expression (RPPA), and copy number variation(CNV). Publicly available clinical information for the all the data sets is retrieved using RTCGA.clinical package [3]. The five multimodal cancer data sets used in this work are as follows:

1) Cervical carcinoma (**CESC**): This cancer accounts for 528,000 new cases and 266,000 deaths worldwide each year, more than any other gynecological tumour [4]. By comprehensive integrated analysis, TCGA research network has identified three subtypes in CESC [5]. The CESC data set consists of 124 samples: 37 samples of keratin-low squamous subgroup, 58 samples of keratin-high squamous subgroup, and 29 samples of adenocarcinoma-rich subgroup.
2) Glioblastoma Multiforme (**GBM**): It is the most common and malignant form of brain cancer and has four subtypes identified in the study by Veerhak *et al.* [6]. The subtypes are proneural, neural, classical, and mesenchymal. The data set consists of 168 samples from three genomic modalities, namely, RNA, miRNA, and CNV, as the mDNA and the RPPA 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) Lower-grade glioma (**LGG**): This is a type of brain tumor originating from glial the cells of the brain. Diffuse low-grade and intermediate-grade gliomas which together make up the lower-grade gliomas have highly variable clinical behaviour that is not adequately predicted on the basis of histological class. Integrative analysis of data from RNA, DNA-copy-number, and DNA-methylation platforms has uncovered three prognostically significant subtypes of lower-grade glioma [7]. The LGG data set consists of 267 samples. The first subtype has 134 samples which exhibit IDH mutation and no 1p/19q codeletion. The second subtype exhibits both IDH mutation and 1p/19q codeletion and has 84 samples. The third one is called the wild-type IDH subtype and has 49 samples.
4) Lung Carcinoma (**LUNG**): Based on the same primary site of origin, lung cancer set can be categorized in two subtypes, namely, adenocarcinoma and squamous cell carcinoma. These were also the two major subtypes of lung cancer in 2015 WHO classification [8]. The LUNG data set consists of 671 samples with 360 samples of lung adenocarcinoma and 311 samples of lung squamous cell carcinoma.
5) Kidney Carcinoma (**KIDNEY**): The kidney cancer data set has three histological subtypes, namely, renal clear cell carcinoma, renal papillary cell carcinoma, and kidney chromophobe. These subtypes were included in the 2004 World Health Organization (WHO) classification of adult renal tumors [9]. The KIDNEY 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 the rare kidney chromophobe subtype.

A summary of the data sets in terms of the number of samples, number of features in each modality, sample-to-feature ratio, and number of clusters is provided in Table S1.

B. Data Platforms and Pre-processing

For the CESC, LGG, LUNG, and KIDNEY data sets, four different modalities, namely, RNA, mDNA, miRNA, and

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. (Corresponding author: Pradipta Maji)

TABLE S1: Summary of Data Sets

Different Data Sets	No. of Samples	No. of Features						Sample to Feature Ratio	No. of Clusters (k)
		mDNA	RNA	miRNA	RPPA	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

RPPA are considered, while for the GBM data set three modalities namely RNA, CNV, and miRNA are considered as mDNA and RPPA modalities are not available for a majority 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 the GBM data set, 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+ [10] 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 [10] 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 GBM data set.

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 [11]. 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 [11]. 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 S1.

C. Statistical Power of Data Sets

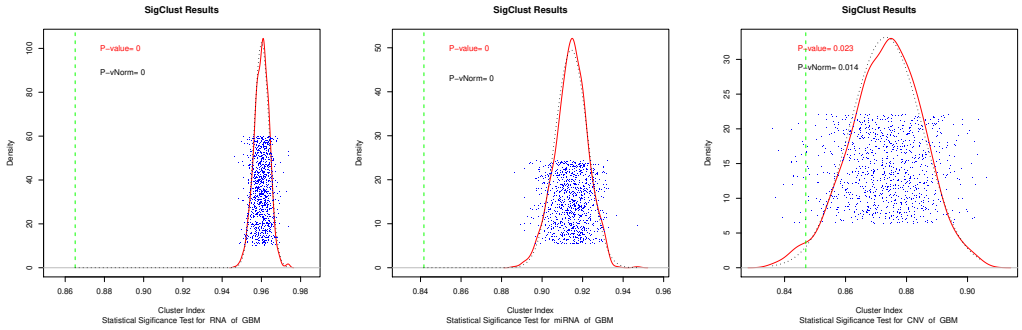
In this section hypothesis testing is performed on the multimodal data sets in order to evaluate whether clusters are "really present" in them as opposed to being artifacts of the natural sampling variation. If the data set comes from only one Gaussian distribution, then any clustering operation that 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 [12] 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 given split of the data is real. The test statistic is the k-means cluster index (CI). It is given by the within-class sums of squares divided by the total sum of squares, in the case where

TABLE S2: Statistical Power of Different Data Sets

Different	p-value based on empirical quantiles					p-value based on Gaussian quantiles (p-vNorm)				
Data Sets	mDNA	RNA	miRNA	RPPA	CNV	mDNA	RNA	miRNA	RPPA	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	-
LUNG	0	0	0	1.730E-001	-	0	7.309E-210	0	1.846E-001	-
KIDNEY	0	0	0	4.000E-03	-	0	3.322E-240	0	4.281E-003	-



(a) RNA (b) miRNA (c) RPPA

Fig. S1: Evaluation of Statistical power of the individual modalities of GBM data set.

the number of clusters is 2, that is

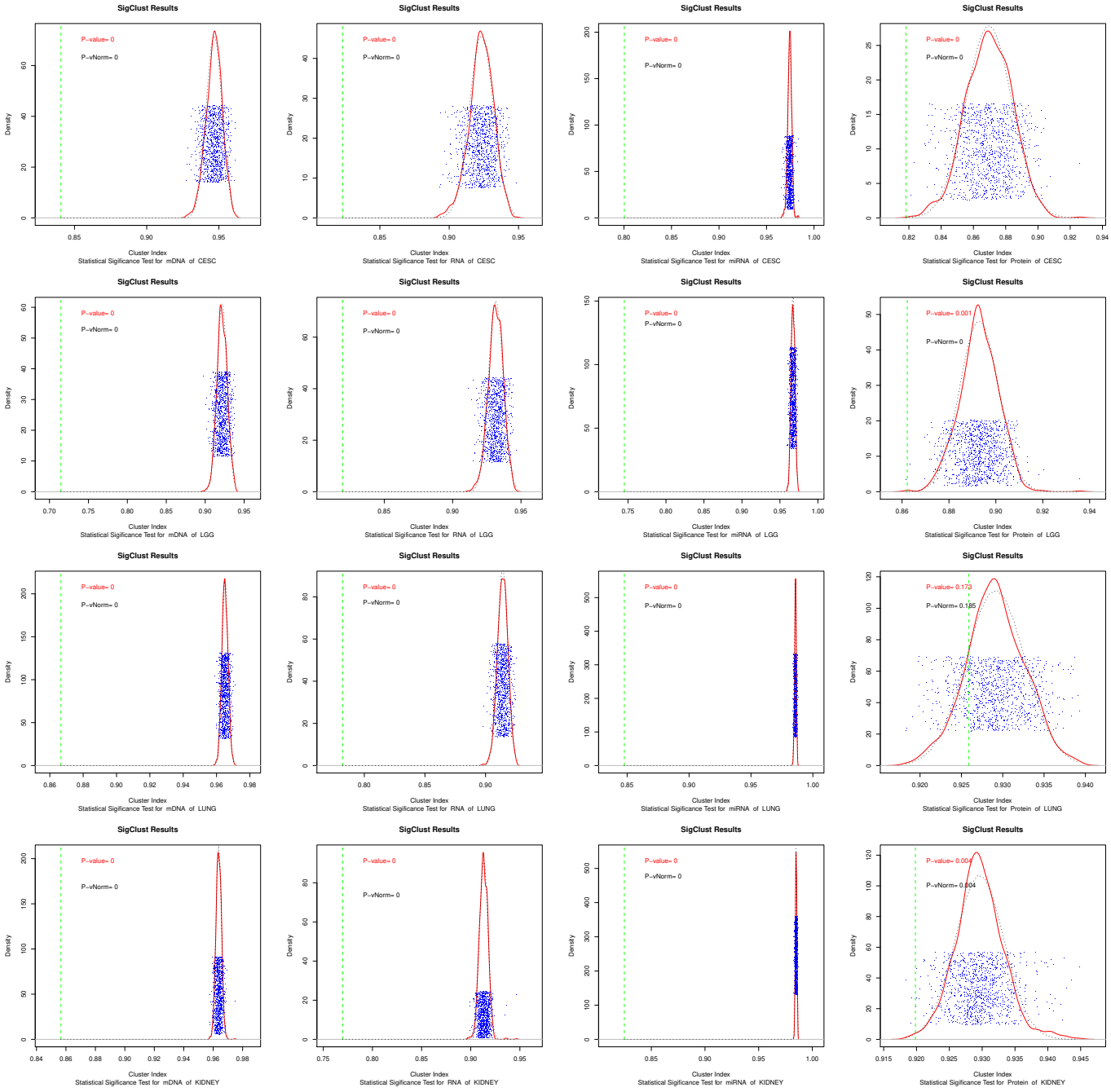
$$CI = \frac{\sum_{j=1}^{2} \sum_{i \in C_k} ||x_i - \bar{x}_{(j)}||^2}{\sum_{i=1}^{n} ||x_i - \bar{x}||^2}, \quad (1)$$

where x_i is the i-th sample of the data set, $\bar{x}$ is the global mean, and $\bar{x}_{(j)}$ is the j-th cluster centroid. 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 [12] 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 as well as using the normal approximation to the distribution of simulated CI values (denoted by p-vNorm). The SigClust summary plots for the simulated null distribution of CI values for different modalities of the GBM data sets are shown in Fig. S1, while for the other four data sets it is shown in S2. 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-value 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 S2. The results in Table S2 show that the p-values obtained for all the modalities of each data set is less than 0.05, except for the RPPA modality of LUNG data set. Thus, mostly the clusters present in individual modalities of 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.

D. Experimental Setup for Existing Algorithms

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



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

Fig. S2: Evaluation of Statistical power of the individual modalities of different data sets: CESC (top row), LGG (second row), LUNG (third row), KIDNEY (bottom row).

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) [13] and cluster of cluster analysis (COCA) [14], and five low-rank direct integrative clustering approaches, namely, joint and individual variation explained (JIVE) [15], iCluster [16], iCluster+ [17], LRAcluster [18], and PCA [19] on the naively concatenated data (PCA-Con). The experimental setup used for these algorithms is briefly outlined as follows:

- **BCC** [13]: 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

TABLE S3: Joint and Individual Ranks Obtained by JIVE Algorithm

Different Data Sets	Algorithm	Joint Rank	Individual Ranks					Algorithm	Joint Rank	Individual Ranks				
			mDNA	RNA	miRNA	RPPA	CNV			mDNA	RNA	miRNA	RPPA	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	-

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 both assigned to 1, under the default setting. The BCC algorithm additionally has Dirichlet prior concentration parameter β_0 having default value of 1. However this default value sometimes tend to yield 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 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** [14]: 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 [20] is then performed on the indicator matrix of 0's and 1's using ConsensusClusterPlus R package [21] 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 [14].
- **JIVE** [15]: 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 S3 for different data sets.
- **iCluster** [16]: 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 [16]. 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 approach of Fang and Wang [22] is used to generate different combination of λ values that are scattered uniformly across the search domain as suggested in [23].
- **iCluster+** [17]: 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, and RPPA 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 [10] 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** [18]: 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, and RPPA 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

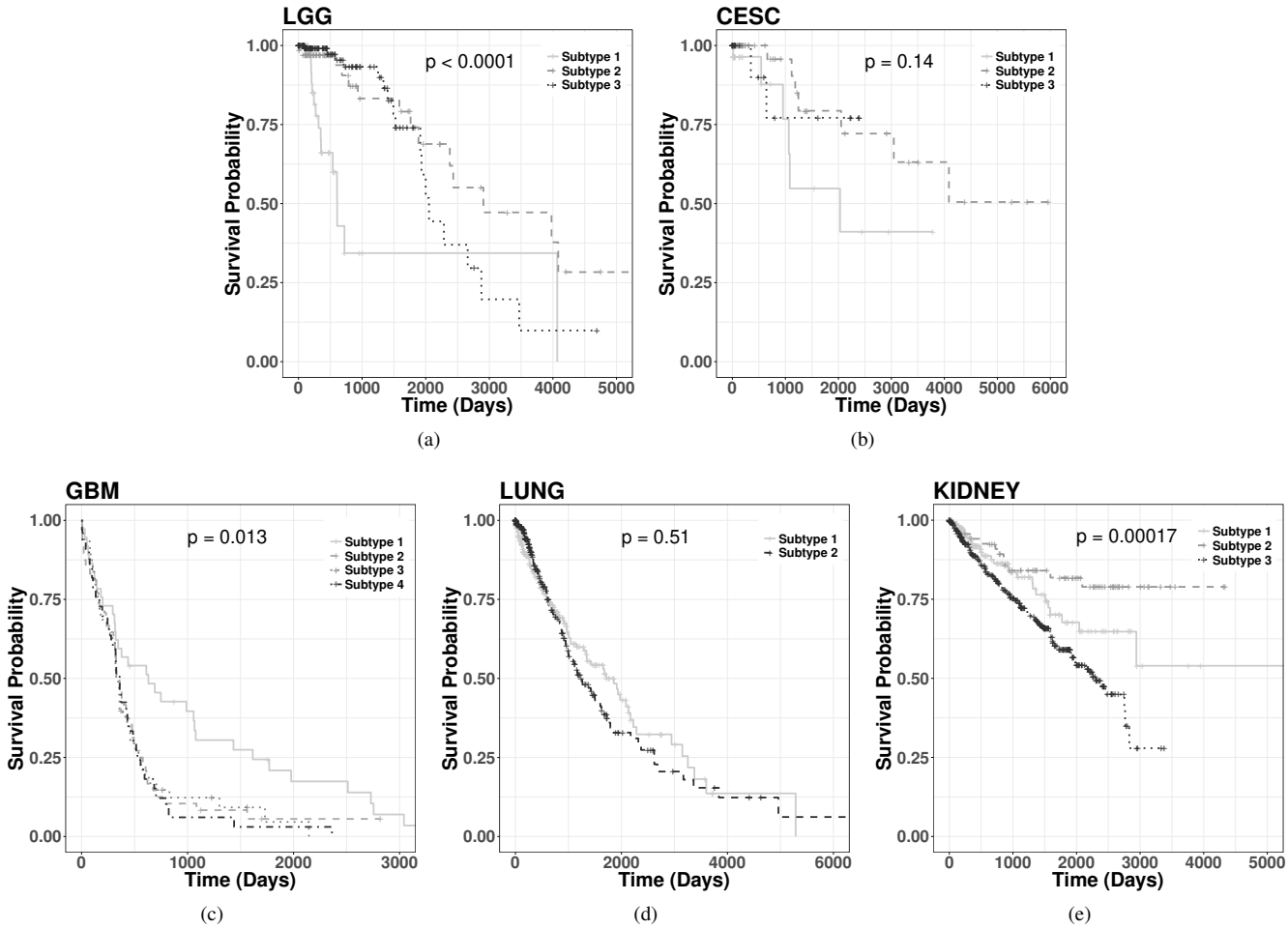


Fig. S3: Kaplan-Meier survival plots for subtypes identified by SURE on different data sets.

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 [18], 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. Based on this criteria, the optimal rank obtained for the CESC, GBM, LGG, LUNG, and KIDNEY data sets are 1, 3, 2, 1, and 2, respectively. After obtaining the optimal low-rank subspace, k-means clustering is performed in that subspace.

- **PCA-Con** [19]: In the PCA-Con 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-Con approach is same as the dimension of the joint subspace extracted by the proposed approach, that is, the number of clusters k. The PCA is performed using SVD of the integrated data matrix.

For all the low-rank based approaches, namely, JIVE, iCluster, iCluster+, LRAcluster, PCA-Con, 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.

II. Survival Analysis

Clinical information of the samples, retrieved from the RTCGA.clinical package [3], 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 are compared using Kaplan-Meier survival plots, median survival times, survival probability of the samples within a subtype after two, five, 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

TABLE S4: Survival Analysis for Subtypes Identified by SURE on Different Data Sets

Different Data Sets	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)
LGG	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
CESC	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
GBM	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
LUNG	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
KIDNEY	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

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, five, and seven years of diagnosis is also observed to study the prognosis of respective cancer with time. The survival results are reported in Fig. S3 and Table S4.

The Kaplan-Meier plot for the subtypes of LGG data set is given in Fig. S3a. The p-values for the log-rank test and the generalized Wilcoxon test are $2.125E-07$ and $5.901E-09$, 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 S4 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 after two 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 2 having probability 0.551 after seven years of diagnosis of cancer.

The survival plot for the CESC data is are given in Fig. S3b. Fig. S3b and Table S4 show that 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, 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 subtypes 2 and 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 S4 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 S3c. 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, years respectively. Comparative results from survival analysis of other data sets in Table S4 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 LUNG data set are given in Fig. S3d and Table S4, 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. S3e and the results are reported in Table S4. 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 are statistically significant 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.

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

Matrix perturbation theory [24] 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 [25] 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:

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^*; \tag{2}$$

$$\widetilde{A} = \begin{bmatrix} \widetilde{U}_1 & \widetilde{U}_2 \end{bmatrix} \begin{bmatrix} \widetilde{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \widetilde{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \widetilde{V}_1 & \widetilde{V}_2 \end{bmatrix}^*, \tag{3}$$

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

$$\Sigma_1 = diag(\sigma_1, \ldots, \sigma_r), \quad \Sigma_2 = diag(\sigma_{r+1}, \ldots, \sigma_n), \tag{4}$$

$$\widetilde{\Sigma}_1 = diag(\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_r), \quad \widetilde{\Sigma}_2 = diag(\widetilde{\sigma}_{r+1}, \ldots, \widetilde{\sigma}_n). \tag{5}$$

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 $\Phi(V_1, \widetilde{V}_1)$ 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:

$$\mathbb{R}_L = A\widetilde{V}_1 - \widetilde{U}_1\Sigma_1 = (A - \widetilde{A})\widetilde{V}_1 \quad \text{and} \tag{6}$$

$$\mathbb{R}_R = A^*\widetilde{U}_1 - \widetilde{V}_1\Sigma_1 = (A^* - \widetilde{A}^*)\widetilde{U}_1. \tag{7}$$

Theorem 1. ***Wedin's*** $\sin\Theta$ ***Theorem*** *[25] Let A and $\widetilde{A}$ be two complex matrices with SVDs partitioned as in (2) and (3), respectively. If*

$$\delta \stackrel{\text{def}}{=} \min \left\{ \min_{1\le i\le r, 1\le j\le (n-r)} |\sigma_i - \widetilde{\sigma}_{r+j}|, \min_{1\le i\le r} \sigma_i \right\} > 0, \tag{8}$$

$$\textit{then} \quad \sqrt{\| \sin\Theta(U_1, \widetilde{U}_1) \|_F^2 + \| \sin\Phi(V_1, \widetilde{V}_1) \|_F^2} \leq \frac{\sqrt{\| \mathbb{R}_L \|_F^2 + \| \mathbb{R}_R \|_F^2}}{\delta}. \tag{9}$$

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] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," *Int. J. Cancer*, vol. 136, pp. E359–386, Mar 2015.
[5] TCGA Research Network, "Integrated genomic and molecular characterization of cervical cancer," *Nature*, vol. 543, pp. 378–384, 2017.
[6] 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.
[7] 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.
[8] W. D. Travis *et al.*, "Introduction to The 2015 World Health Organization Classification of Tumors of the Lung, Pleura, Thymus, and Heart," *J Thorac Oncol*, vol. 10, pp. 1240–1242, Sep 2015.
[9] S. R. Prasad *et al.*, "Common and uncommon histologic subtypes of renal cell carcinoma: imaging spectrum with pathologic correlation," *Radiographics*, vol. 26, no. 6, pp. 1795–1806, 2006.
[10] Q. Mo and R. Shen, *iClusterPlus: Integrative clustering of multi-type genomic data*, 2016. R package version 1.12.1.
[11] 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.
[12] 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.
[13] E. F. Lock and D. B. Dunson, "Bayesian consensus clustering," *Bioinformatics*, vol. 29, no. 20, pp. 2610–2616, 2013.
[14] 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.
[15] 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.
[16] 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*, vol. 25, no. 22, pp. 2906–2912, 2009.
[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] K. T. Fang and Y. Wang, *Number-Theoretic Methods in Statistics*. Chapman and Hall/CRC, 1993.
[23] 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.
[24] G. W. Stewart and J.-g. Sun, *Matrix perturbation theory*. Academic press New York, 1990.
[25] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.