Supplementary Material: Low-Rank Joint Subspace Construction for Cancer Subtype Discovery

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

In this supplementary material, Section 1 contains an illustration of the working principle of the proposed algorithm using the real-life cervical cancer (CESC) data set. Section 2 contains description of the data sets and the data pre-processing steps. The experimental setup and parameter tunning approaches used for the existing algorithms are outlined in Section 3. Section 4 describes the cluster evaluation measures used in this work to evaluate the clustering performance of different algorithms.

1 Illustrative Example for Proposed Method on CESC Data Set

The proposed algorithm uses multivariate normality to the estimate rank and relevance of the individual modalities. The rank and relevance of the modalities are reported in Table 1 for different data sets. Table 1 shows that the relevance and rank of the modalities vary among the data sets, and hence different subsets of modalities are selected for different data sets. A modality having zero rank indicates that its first two principal components are normally distributed, and the modality contains only the noise component. This automatically eliminates noisy modalities having zero rank and low relevance values (like Gene and miRNA modalities of LGG, and DNA modality of OV) from integrating into the joint subspace. For CESC data set, initially all the modalities have non-zero ranks and all are considered for joint subspace construction. However, during integration, a majority of the residual components from different modalities turn out to be normal with respect to the existing joint subspace. Hence, they are not integrated into the final subspace, thus performing a second level of noise removal.

The working principle of the proposed algorithm is illustrated using the CESC data set as an example. Table 1 shows that for the CESC data set, the rank r of DNA, Gene, miRNA, and Protein are 3, 2, 5, and 4, respectively. Fig. 1 and Fig. 2 show density plots, quantile-quantile (Q-Q) plots, and p-values for the first 5 principal components of Gene and DNA modalities, respectively of CESC data set. These figures show that third, fourth, and fifth components of the Gene, and fourth and fifth components of DNA are normally distributed, depicting the random Gaussian noise component of these modalities. On the other hand, the first two components of Gene in Fig. 1 show deviation from normality, indicating the presence of clusters. For DNA, Fig. 2 shows that the second principal component abruptly follows a normal distribution, while both first and third components show deviation from normality. Additionally, the remaining components from 4 onwards are normally distributed. So, the rank of DNA is estimated to be 3. The density plots in Fig. 1 and 2 also show that the first component of both Gene and DNA have a bimodal distribution, indicating multiple clusters. According to the relevance values in Table1, four modalities of the CESC data set can be ordered as Gene followed by Protein, miRNA, and DNA. Therefore, the joint subspace construction begins with Gene. Although DNA is the modality with lowest relevance, it has the maximum shared information with Gene, according to the dependency measure. So, DNA is selected next for integration. Fig. 3 shows the density and Q-Q plots of the residuals of DNA with respect to the current joint subspace of Gene. The figure shows that the residuals of the first and second component of DNA are normally distributed with p-values 0.284 and 0.246, respectively, while the third component deviates from normality (p-value is 0.0348). Therefore, only the third principal component of DNA is integrated into the joint subspace. The modality selected next for integration is miRNA whose estimated rank is 5. The density and Q-Q plots for principal components of miRNA and their residuals with respect to the current joint subspace are given in Fig. 4 and 6(a), respectively. The plots of the residuals in 6(a) show that the residual of only the fourth principal component of miRNA shows significant divergence from normality and is selected for integration into the joint subspace. Finally, Protein is selected for

• *A. Khan and P. Maji are with the Biomedical Imaging and Bioinformatics Lab, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: {aparajitak_r, pmaji}@isical.ac.in.*

TABLE 1
Relevance and Rank of Each Modality and Selected Modalities

Different Modalities	CESC		LGG		OV		BRCA	
	Relevance	Rank	Relevance	Rank	Relevance	Rank	Relevance	Rank
DNA	0.1884817	3	0.4320317	10	0.0230986	0	0.2373227	5
Gene	0.2921399	2	0.0289518	0	0.4936741	3	0.2947759	3
miRNA	0.1990886	5	0.0056958	0	0.2474369	5	0.1602746	4
Protein	0.2006048	4	0.2428867	6	0.0579902	2	0.2464338	6
Selected	Gene, DNA, miRNA, Protein		DNA, Protein		Gene, miRNA, Protein		Gene, DNA, miRNA, Protein	

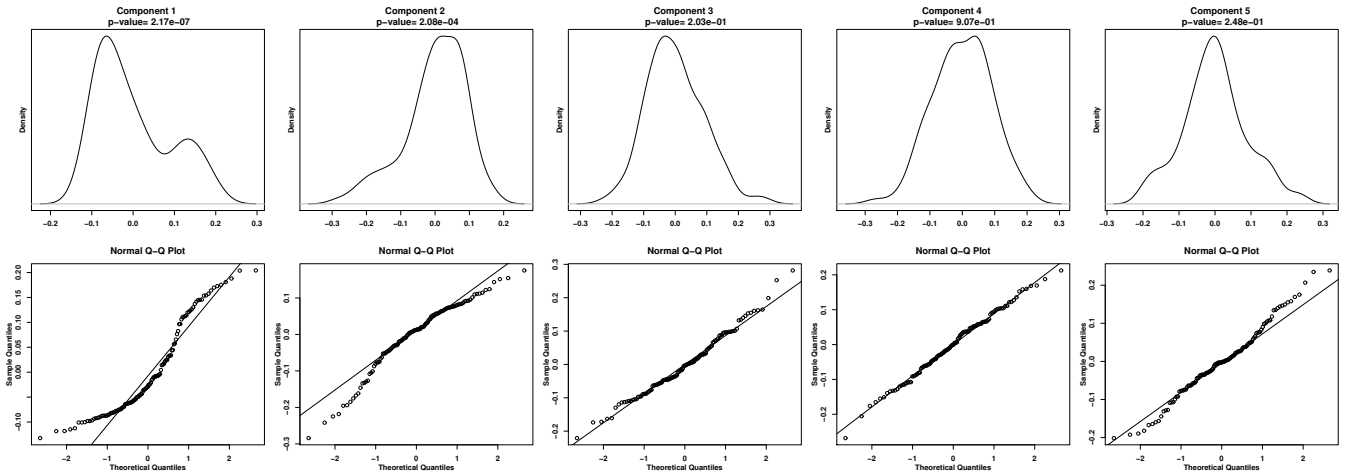


Fig. 1. Density and Q-Q plots for first five components of gene expression data of CESC

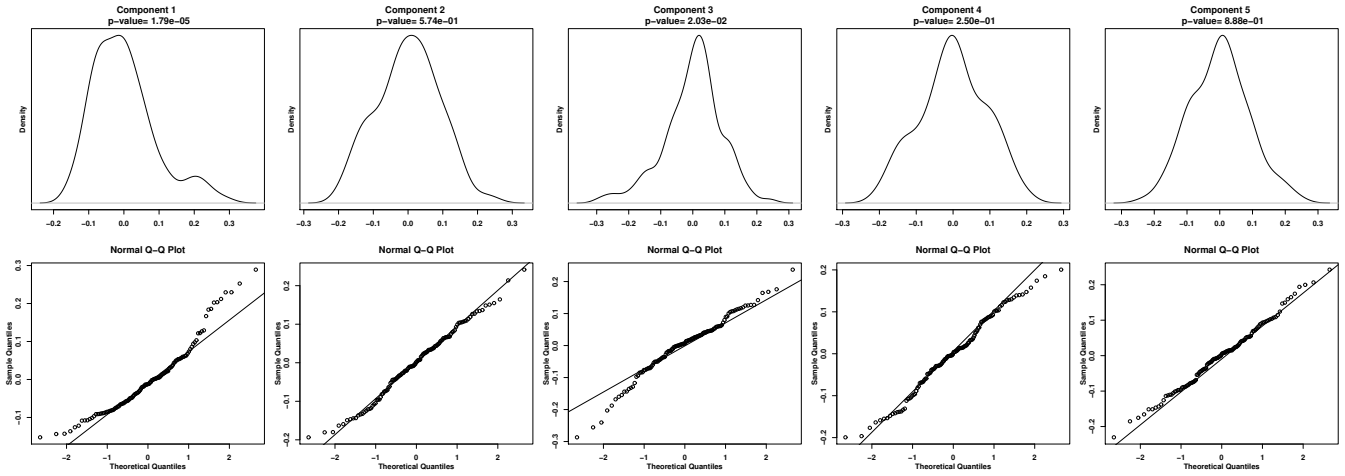


Fig. 2. Density and Q-Q plots for first five components of DNA methylation data of CESC

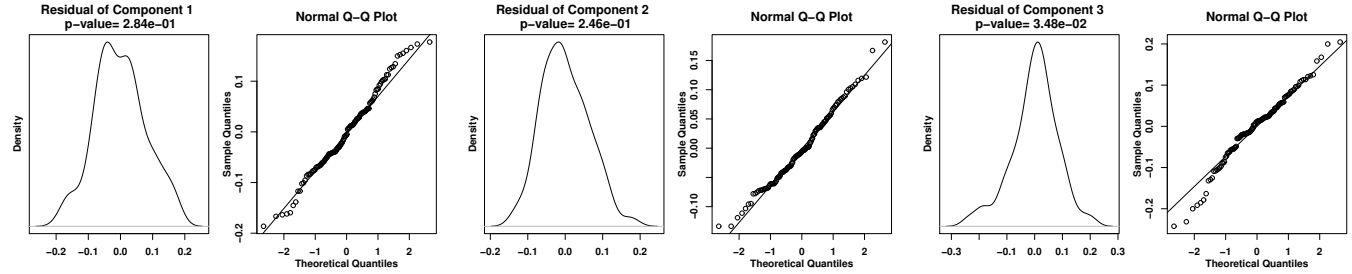
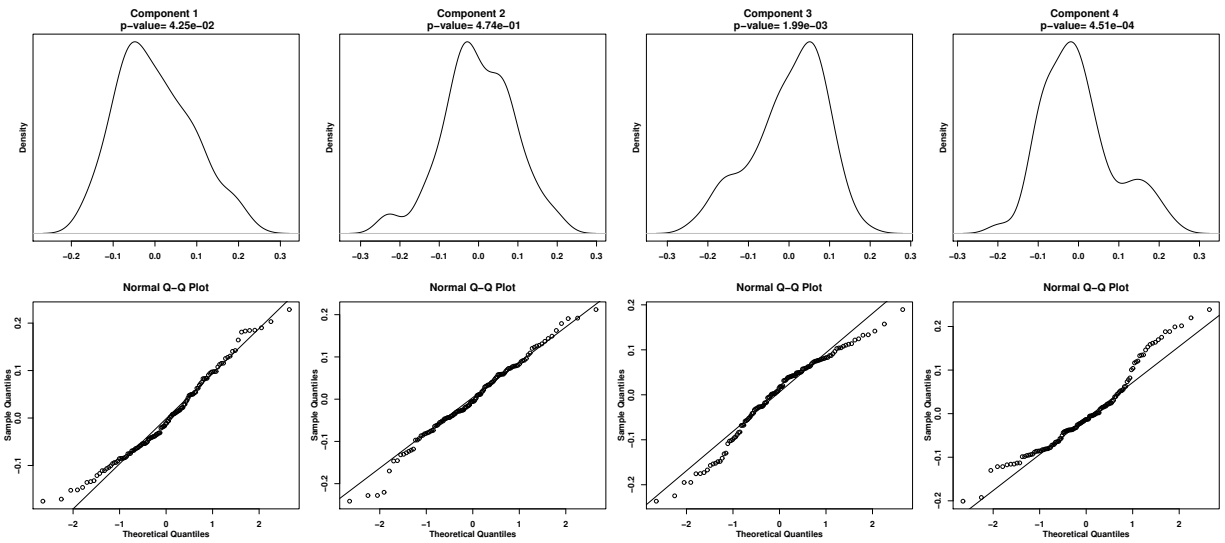
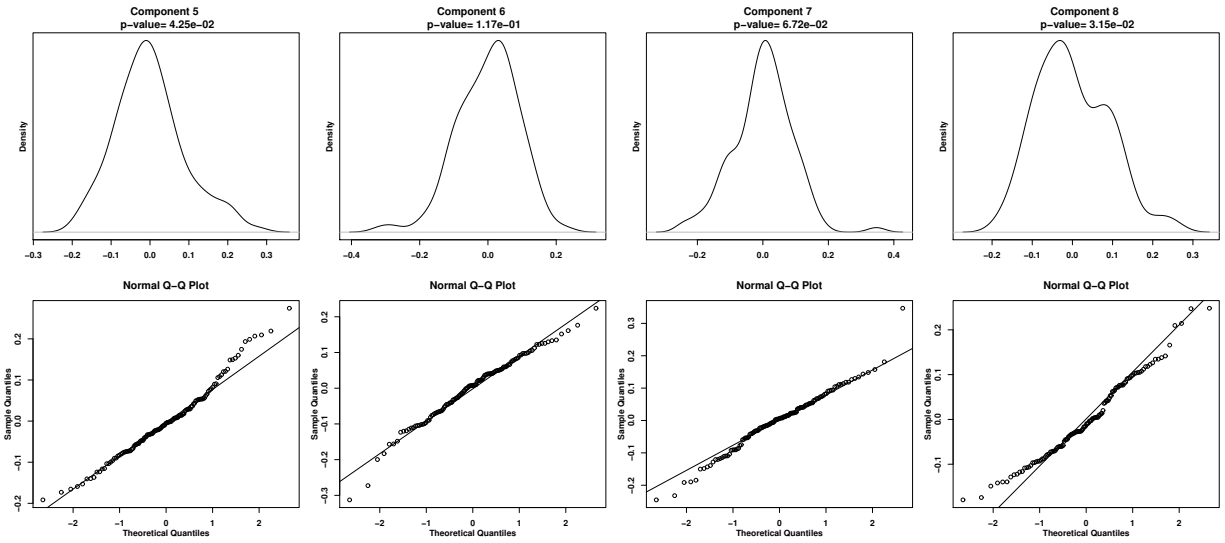


Fig. 3. Density and Q-Q plots for the residual components the DNA methylation data of CESC



(a)



(b)

Fig. 4. Density and Q-Q plots for first 8 components of miRNA modality of CESC data set

integration whose estimated rank is 4. The density and Q-Q plots for principal components of Protein and their residuals are given in Fig. 5 and 6(b), respectively. Fig. 6(b) shows that out of the top four principal components of Protein, the residuals of only the first and second components show deviation from normality. Thus only these two components of the Protein modality are integrated into the joint subspace and the rest are eliminated as noisy ones, thus forming a six dimensional joint subspace for CESC.

2 Description of Datasets

The descriptions of four real-life multimodal cancer data sets from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/), which are used in this study 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 [1]. By comprehensive integrated analysis, TCGA research network has identified three subtypes in CESC [2]. 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) Lower grade glioma (LGG): Diffuse low-grade and intermediate-grade gliomas which together make up the lower-grade gliomas have highly variable clinical behavior 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 [3]. The LGG data set consists of 267 samples. The first subtype has 134 samples which

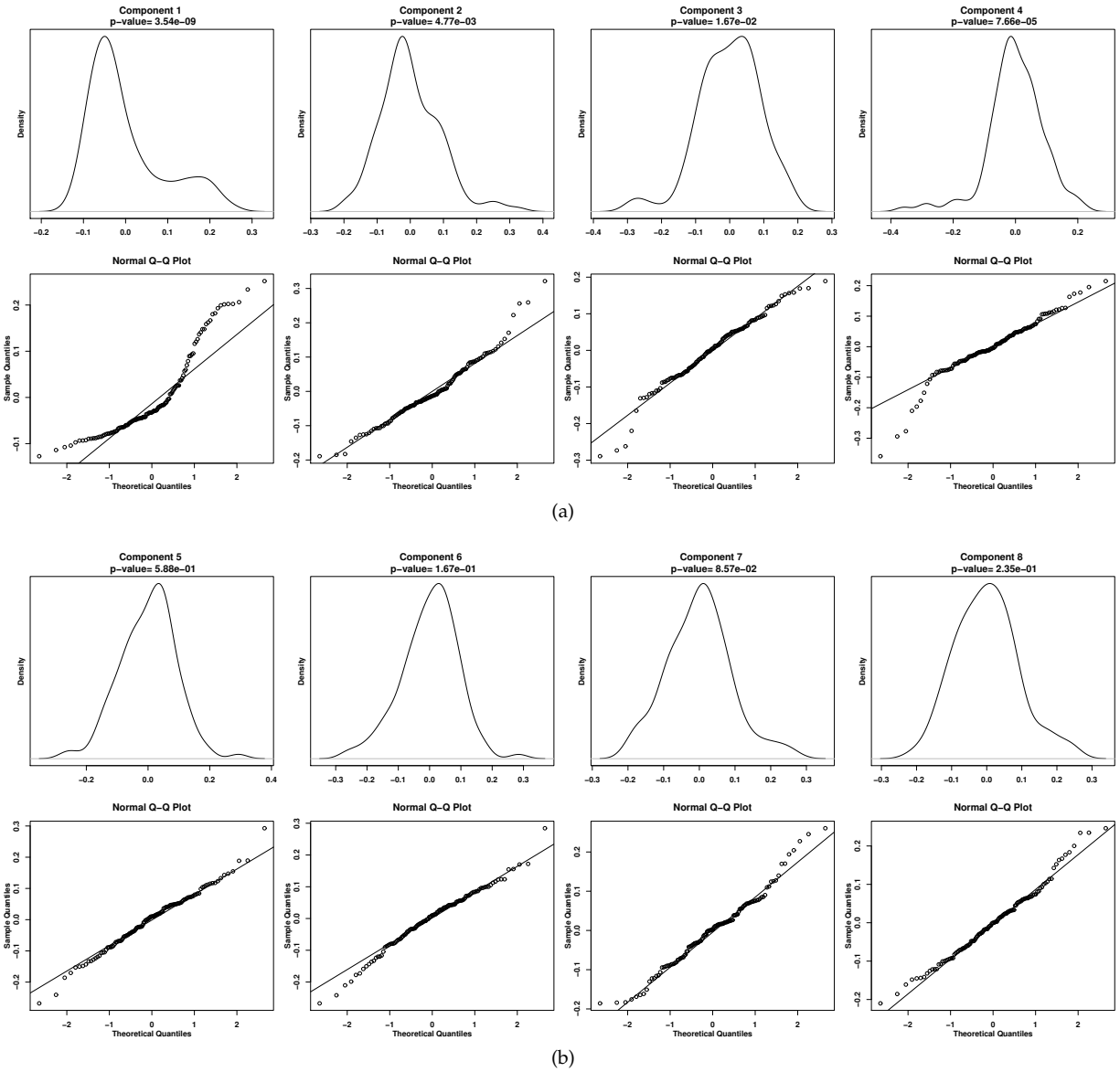


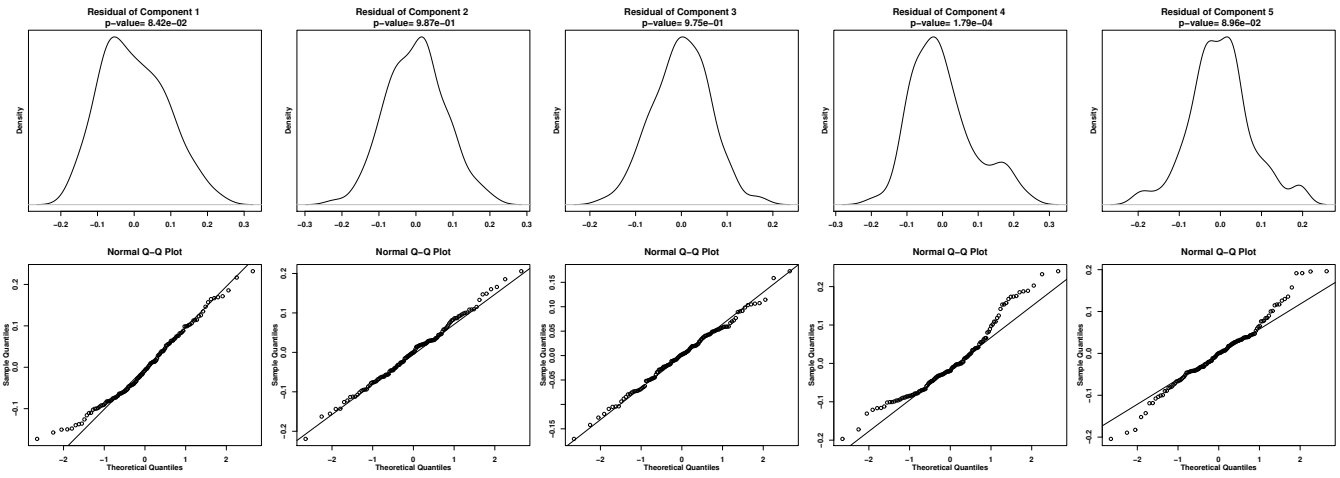
Fig. 5. Density and Q-Q plots for first 8 components of Protein modality of CESC data set

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.

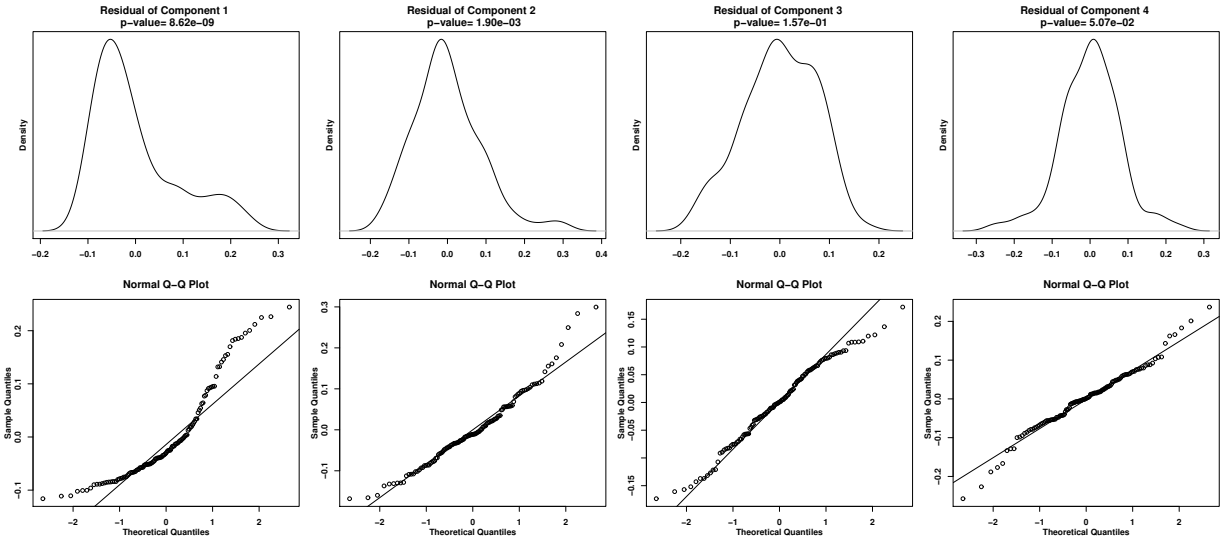
3) Ovarian carcinoma (OV): Ovarian cancer is the eighth most commonly occurring cancer in women and there were nearly 300,000 new cases in 2018 [4]. Ovarian cancer encompasses a heterogeneous group of malignancies that vary in etiology, molecular biology, and numerous other characteristics. TCGA researchers have identified four robust expression subtypes of high-grade serous ovarian cancer [5]. The OV data set consists of 334 samples. The four subtypes are termed as immunoreactive, differentiated, proliferative, and mesenchymal, consisting of 74, 91, 90, and 79 samples, respectively.
4) Breast invasive carcinoma (BRCA): Breast cancer is one of the most common cancers with greater than 1,300,000 cases and 450,000 deaths each year worldwide [6]. During the last 15 years, four intrinsic molecular subtypes of breast cancer (Luminal A, Luminal B, HER2-enriched, and Basal-like) have been identified and intensively studied [7], [8], [6]. The BRCA data set consists of 398 samples comprising of 171, 98, 49, and 80 samples of LuminalA, LuminalB, HER2-enriched, and Basal-like subtype, respectively.

These subtypes have been shown to be clinically relevant and provide roadmap for patient stratification and trials of targeted therapies. For all the data sets, four different omic modalities are considered, namely, DNA methylation (DNA), gene expression (Gene), microRNA expression (miRNA), and protein expression (Protein).

Data pre-processing: For the DNA methylation modal-



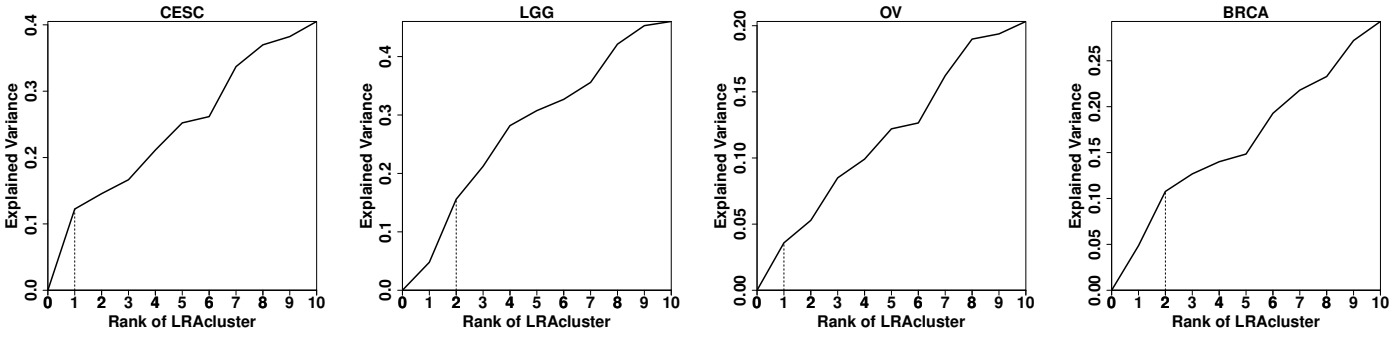
(a) Residuals of miRNA



(b) Residuals of Protein

Fig. 6. Density and Q-Q plots for the residual components the miRNA and Protein modalities of CESC data set

ity, methylation β-values from Illumina HumanMethylation450 and HumanMethylation450 beadarray platforms are used. The HumanMethylation450 beadarray gives methylation β-values of approximately 450,000 CpG sites, while HumanMethylation27 beadarray covers 27,000 CpG sites. These two platforms share a common set of 25,978 CpG locations. For all the data set, methylation data across those common 25,978 CpG locations are considered. Additionally, CpG locations with missing gene information were filtered out from the study. The top 2,000 most variable CpG sites are used for clustering. For the Gene modality of CESC, LGG, and BRCA data sets, RNA-sequence data from Illumina HiSeq platform is used which contains normalized RPKM (reads per kilobase of exon per million) counts for 20,531 genes. The data is then log transformed and 2,000 most variable genes based on their expression profile across the samples are considered. Sequence based microRNA expression data from Illumina HiSeq platform is used for CESC, LGG, and BRCA data sets, which contains RPM (reads per million miRNA mapped) values for 1045 miRNAs. The miRNA sequence data is also log transformed and only those miRNAs for which the expression value is present for 95% of the samples are considered. On the other hand, for the OV data set, array based gene and miRNA expression data from AgilentG4502A_07_3 and H-miRNA_8x15Kv2 platforms are used. The Gene modality of OV data set consists of log-ratio based expression data for 17,814 genes amongst which 2,000 most variable genes are considered. The miRNA expression data is available for 799 microRNAs. For protein modality of all the data sets, reverse phase protein array data from the MDA_RPPA_Core platform having approximately 220 proteins is used. These four modalities, measured on different platforms represent a wide variety of biological information.



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

Fig. 7. Optimal rank estimation of LRAcluster for different data sets

3 Experimental Setup for Existing Algorithms

The performance of the proposed algorithm is compared with nine existing integrative clustering based approaches, namely, cluster of cluster analysis (COCA) [9], Bayesian consensus clustering (BCC) [10], Bayesian correlated clustering (referred to as MDI) [11], and clusternomics [12], LRAcluster [13], joint and individual variance explained (JIVE) [14], iCluster [15], iCluster2 [16], and principal component analysis (PCA) on concatenated data (PCA-con) [17].

The experimental setup used for these algorithms is briefly outlined as follows:

- **COCA** [9]: This is a consensus clustering based approach which first cluster each modality separately and the individual clustering solutions are then combined to get the final cluster assignments. Subtypes identified from each modality are encoded into a series of indicator variables for each subtype. Consensus clustering is performed on the indicator matrix of 0's and 1's using ConsensusClusterPlus R-package [18] to identify structure and relationship of the samples. The consensus clustering algorithm uses re-sampling based technique to find the clusters, so its performance varies on different executions of the algorithm. The average performance of the COCA algorithm over 10 executions is reported in this work. Parameters for consensus cluster are 80% sample re-sampling with 1000 iterations of hierarchical clustering based on a Pearson correlation distance metric [9].
- **LRAcluster** [13]: 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. In this work, four omic modalities are considered for each cancer data set. For Gene and miRNA modalities of CESC, LGG, and BRCA data sets, sequence based count data are considered, while for DNA and Protein modalities, array based expression data is considered. Therefore, as suggested by the authors [13], count based Gene and miRNA modalities are not log transformed and are modeled using Poisson distribution, while array based DNA and Protein modalities are modeled using Gaussian distribution. On the other hand, for the OV data set, for all the four modalities, array based data is considered which is modeled using Gaussian distribution. For LRAcluster, the rank of the lower dimensional subspace is optimized using the likelihood based "explained variation" criteria [13], 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. 7 for different data sets. Based on this criteria, the optimal rank obtained for the CESC, LGG, OV, and BRCA data sets are 1, 2, 1, and 2, respectively. After obtaining the optimal low-rank subspace, k-means clustering is performed in that subspace to identify the clusters.
- **JIVE** [14]: 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 2 for different data sets.
- **iCluster** [15] and **iCluster2** [16]: These are low-rank based approach which use Gaussian latent variable model to extract a $(k-1)$ dimensional

TABLE 2
Joint and Individual Ranks Obtained by JIVE Algorithm

Different Datasets	Algorithm	Joint Rank	Individual Ranks				Algorithm	Joint Rank	Individual Ranks			
			DNA	Gene	miRNA	Protein			DNA	Gene	miRNA	Protein
CESC	JIVE (PERM)	5	15	21	13	10	JIVE (BIC)	1	1	0	1	1
LGG		2	12	23	18	12		2	1	2	2	2
OV		8	34	50	33	23		0	2	1	2	1
BRCA		3	30	36	15	15		1	3	4	1	0

TABLE 3
Lasso penalty parameter for iCluster and iCluster2 Algorithms

Different Datasets	Algorithm	Lasso penalty parameter (λ)				Algorithm	Lasso penalty parameter (λ)			
		DNA	Gene	miRNA	Protein		DNA	Gene	miRNA	Protein
CESC	iCluster	0.95928338	0.35667752	0.04723127	0.05048859	iCluster2	0.32736156	0.81596091	0.33713355	0.22638436
LGG		0.52280130	0.02442996	0.09283387	0.96579804		0.70846905	0.82247557	0.61074918	0.71824104
OV		0.93322475	0.26221498	0.07980456	0.41856677		0.34690553	0.63680781	0.81270358	0.20032573
BRCA		0.33387622	0.08957654	0.82899022	0.88436482		0.82573289	0.74755700	0.46416938	0.56188925

joint subspace of a multimodal data set, where k is the number of clusters in the data set. For both iCluster and iCluster2, clustering is performed in the $(k-1)$ dimensional joint subspace extracted by the corresponding algorithms. Therefore, the dimension of low-rank subspaces extracted by iCluster and iCluster2 algorithms for CESC, LGG, OV, and BRCA data sets are 2, 2, 3, and 3, respectively. For each modality, the iCluster and iCluster2 algorithms have a lasso penalty parameter (λ), which is tuned using proportion of deviance (POD) statistic [15] and reproducibility index (RI) [19], [16], respectively. For iCluster, 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. On the other hand, for iCluster2, RI is computed by repeatedly partitioning the samples into a learning and a test set and then evaluating the degree of agreement between the predicted and the fitted cluster assignment using adjusted Rand index. Values of RI close to 1 indicate perfect cluster reproducibility and values of RI close to 0 indicate poor cluster reproducibility. The penalty parameter (λ) for each modality ranges between 0 and 1, with 0 representing the null model where no features are selected and 1 representing the full model where all features are included. The uniform sampling design (UD) approach of Fang and Wang [20] is used to generate different combination of λ values that are scattered uniformly across the search domain. The optimal value λ parameter for iCluster and iCluster2 algorithms is given in Table 3 for different data sets.

- **PCA-con** [17]: In this approach the performance of k-means clustering on the k largest principal components of the integrated data is studied, where the integrated data is obtained by naively concatenating features from all the available modalities.
- **BCC** [10]: This is a consensus clustering based approach which uses Dirichlet mixture model to separately cluster the individual modalities. Then a Bayesian framework is used for simultaneous estimation of both the consensus clustering and the source-specific clusterings. The number of clusters is set to k overall and the algorithm is executed for 10,000 iterations. The adherence parameter α is given Beta(1,1) prior distribution and is fitted separately for each modality. The Dirichlet prior concentration parameter β_0 has a default value of 1 which often yielded less than k clusters. Therefore, the Dirichlet 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 value of β_0 is selected using an estimated adherence statistic $\alpha*$, as proposed by the authors.
- **MDI** [11]: This is a Bayesian method for integrative modeling of multimodal datasets. Each modality is modeled using a Dirichlet-multinomial allocation mixture model, and the dependencies between these models is captured pairwise agreement between their clusterings. The Matlab implementation of the MDI algorithm available from https://warwick.ac.uk/fac/sci/systemsbiology/research/software/ has been used with its default parameter settings in this work. The MDI algorithm allows different modalities to be modeled by different distributions. The count based Gene and miRNA modalities of CESC, LGG, and BRCA data sets are not log transformed and are modeled using Poisson distribution, while array based DNA and Protein modalities are modalities are modeled using Gaussian distribution. For the OV data set, all the four modalities contain array based real-valued data, so they are modeled using Gaussian distribution. The maximum number of clusters that may appear in the data set was set to half the

number of samples in the dataset, as recommended by the authors. Given these parameter settings, MDI automatically estimates the number of clusters in the data set. However, the number of clusters estimated by MDI ranges between 8 to 13 on the four TOGA data sets, which is much larger than the actual number of clusters in those data sets.

- **Clusternomics** [12]: This is a probabilistic clustering method which identifies groups of samples that share global behavior across heterogeneous modalities. The algorithm models clusters on the level of individual modalities, while also extracting global structure that arises from the local cluster assignments. Clusters on both the local and the global level are modeled using a hierarchical Dirichlet mixture model to identify structure on both levels. All the four modalties of each data set are modeled using the default multivariate Normal distribution with diagonal covariance matrix. The number of local clusters in each modality, as well as the number of global clusters is set to k, the actual number of clusters in the data set.

The Bayesian approaches like BCC, MDI, and clusternomics algorithms use Markov chain Monte Carlo (MCMC) simulations to obtain the consensus/global clusters. For these approaches, the MCMC algorithm is executed for 1,000 iterations and different MCMC simulations can result in different solutions. Therefore, for BCC, MDI, and clusternomics, the average performance over 10 different executions is reported here.

4 Cluster Evaluation Measures

In this work, the clustering performance of the individual modalities, the existing integrative clustering approaches, and the proposed algorithm is evaluated using five external cluster evaluation indices, namely, F-measure [21], Rand index [22], Jaccard coefficient [23], Dice coefficient [23], and purity [24]. These external cluster validity indices compare the identified clusters with the clinically established subtypes of each cancer data set. The indices are described as follows:

Let $\mathcal{T} = \{t_1, \ldots, t_i, \ldots, t_{k_\mathcal{T}}\}$ be the true partition of n samples of a data set into $k_\mathcal{T}$ clusters. Let $\mathcal{C} = \{c_1, \ldots, c_j, \ldots, c_{k_\mathcal{C}}\}$ be the $k_\mathcal{C}$ 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 next.

1) Set-matching indices: These indices are based on matching entire clusters, where similar clusters are first found either by pairing or matching, and their similarity is then measured using set matching methods. Two set-matching based indices considered in this work are as follows:
   a) F-measure [21]: The F-measure 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.

$$Precision\ P_{ij} = \frac{|c_i \cap t_j|}{|c_i|}. \tag{1}$$

$$Recall\ R_{ij} = \frac{|c_i \cap t_j|}{|t_j|}. \tag{2}$$

$$\mathcal{F}(t_j, c_i) = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}} \tag{3}$$

$$= \frac{2|c_i \cap t_j|}{|c_i| + |t_j|}. \tag{4}$$

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

$$F-measure = \frac{1}{n}\sum_{j=1}^{k_\mathcal{T}} |t_j|\ \max_i\{\mathcal{F}(t_j, c_i)\}. \tag{5}$$

   b) Purity [24]: It 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,

$$Purity = \frac{1}{n}\sum_{i=1}^{k_\mathcal{C}} \max_j\ \{|c_i \cap t_j|\}. \tag{6}$$

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

2) Pair-counting indices: Pair-counting measures count the pairs of points on which the two clusterings agree or disagree. In a n sample data set, the $\binom{n}{2}$ pairs of points can be divided into four categories. Let a represent the number of pairs that are in the same cluster both in $\mathcal{C}$ and $\mathcal{T}$, b represent the number of pairs that are in the same cluster in $\mathcal{C}$ but in different clusters in $\mathcal{T}$, c represents the number of pairs that are in different clusters in $\mathcal{C}$ but in the same cluster in $\mathcal{T}$, and d represent the number of pairs that are in different clusters both in $\mathcal{C}$ and $\mathcal{T}$.The values a and d count the agreements while b and c the disagreements. Three pair-counting based indices are considered in this work are as follows:
   a) Rand [22]: It is defined as the ratio of the total number of agreements to the total number of pairs, given by

$$Rand = \frac{a+d}{a+b+c+d} \tag{7}$$

   b) Jaccard [23]: The Jaccard similarity coefficient is defined as

$$Jaccard = \frac{a}{a+b+c} \tag{8}$$

c) Srensen-Dice coefficient [23]: This index is defined as

$$Dice = \frac{2a}{2a+b+c} \tag{9}$$

All the external cluster validation indices lie in [0,1] and a higher value indicates better clustering.

Inorder to evaluate the robustness of clusters identified by the proposed approach, the Davies-Bouldin index [25] is used. It is an internal cluster validity index which evaluates the quality of clustering based on the information intrinsic to data like compactness and separation of the identified clusters. The information of the correct partition of the data is not used during internal cluster evaluation. Let $X = \{x_1, ..., x_i, ..., x_n\}$ be the set of n samples, where $x_i \in \mathbb{R}^k$ represents the i-th sample in a k-dimensional subspace. Let the Euclidean distance between samples x_i and x_j be denoted as $d_e(x_i, x_j)$. The k clusters are represented as $\mathcal{C} = C_1, ..., C_k$, and the centroids of each of k clusters are $v_1, ..., v_k$. Let the centroid of the dataset be given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The Davies-Bouldin index estimates the compactness based on the distance from the samples in a cluster to its centroid and separation based on the distance between centroids. It is defined as

$$DB = \frac{1}{k}\sum_{C_j \in \mathcal{C}} \max_{C_l \in \mathcal{C} \setminus C_j} \left\{ \frac{S(C_j) + S(C_l)}{d_e(v_j, v_l)} \right\}, \tag{10}$$

$$\text{where } S(C_j) = \frac{1}{|C_j|}\sum_{x_i \in C_j} d_e(x_i, v_j). \tag{11}$$

REFERENCES

[1] 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.

[2] TCGA Research Network, "Integrated genomic and molecular characterization of cervical cancer," *Nature*, vol. 543, no. 7645, pp. 378–384, 2017.

[3] 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.

[4] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin*, Sep 2018.

[5] TCGA Research Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, pp. 609–615, Jun 2011.

[6] TCGA Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, Oct 2012.

[7] Z. Hu *et al.*, "The molecular portraits of breast tumors are conserved across microarray platforms," *BMC Genomics*, vol. 7, p. 96, Apr 2006.

[8] T. Sorlie *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, pp. 10869–10874, Sep 2001.

[9] K. A. Hoadley, C. Yau, *et al.*, "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," *Cell*, vol. 158, no. 4, pp. 929–944, 2014.

[10] E. F. Lock and D. B. Dunson, "Bayesian consensus clustering," *Bioinformatics*, vol. 29, pp. 2610–2616, Oct 2013.

[11] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild, "Bayesian correlated clustering to integrate multiple datasets," *Bioinformatics*, vol. 28, pp. 3290–3297, Dec 2012.

[12] E. Gabasova, J. Reid, and L. Wernisch, "Clusternomics: Integrative context-dependent clustering for heterogeneous datasets," *PLoS Comput. Biol.*, vol. 13, p. e1005781, Oct 2017.

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

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

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

[16] 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.

[17] I. Jolliffe, *Principal Component Analysis*. Springer Series in Statistics, Springer, 2002.

[18] M. D. Wilkerson and D. N. Hayes, "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking," *Bioinformatics*, vol. 26, pp. 1572–1573, Jun 2010.

[19] R. Shen, S. Wang, and Q. Mo, "Sparse integrative clustering of multiple omics data sets," *The annals of applied statistics*, vol. 7, no. 1, pp. 269–294, 2013.

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

[21] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 16–22, ACM, 1999.

[22] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[23] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, no. 2-3, pp. 107–145, 2001.

[24] E. Rendón, I. M. Abundez, *et al.*, "A comparison of internal and external cluster validation indexes," in *Proceedings of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications*, pp. 158–163, 2011.

[25] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, vol. 1, no. 2, pp. 224–227, 1979.