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

Aparajita Khan and Pradipta Maji

I. Computational Complexity

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

$$\widetilde{\mathbf{X}}_M = \begin{bmatrix} X_1 & \ldots & X_m & \ldots & X_M \end{bmatrix}. \quad (1)$$

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:

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 parallelly 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, in steps 3 and 4 can be done for all the modalities in parallel. The 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 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 and 9 are assignments operations which take $\mathcal{O}(1)$ time. For the 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 the eigenspace update. 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.

During eigenspace update, 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 (20) 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 $\left(\mathcal{O}(d_{max} + nk^2 + nk + k^2 d + nk^2) =\right) \mathcal{O}(k^2 d)$. 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 $\mathcal{O}(Mk^2 d)$. The overall computational complexity of the proposed SURE algorithm is $\left(\mathcal{O}(n^2 d_{max} + Mk^2 d) =\right) \mathcal{O}(n^2 d_{max})$, assuming $M, k << n < d_{max}$. Thus the time complexity is bounded by that of individual eigenspace construction in steps 1-5.

II. Evaluation of Individual Modality

This section describes the two modality evaluation measures, namely, relevance and concordance. While relevance assesses the quality of cluster information provided by each modality, the concordance measures the amount of cluster information shared between two modalities. Let $X_i \in \mathbb{R}^{n \times d_i}$ and $X_j \in \mathbb{R}^{n \times d_j}$ be two modalities of a multimodal data set whose rank k eigenspaces are given by

$$\Psi(X_i) = \langle \mu(X_i), U(X_i), \Sigma(X_i), V(X_i) \rangle; \quad (2)$$

$$\Psi(X_j) = \langle \mu(X_j), U(X_j), \Sigma(X_j), V(X_j) \rangle. \quad (3)$$

A. Relevance

As described in Section III-B of the main article, the relevance of a modality is defined in terms of the compactness of the cluster structure embedded in the left subspace of its eigenspace. The compactness of cluster structure of modality

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.

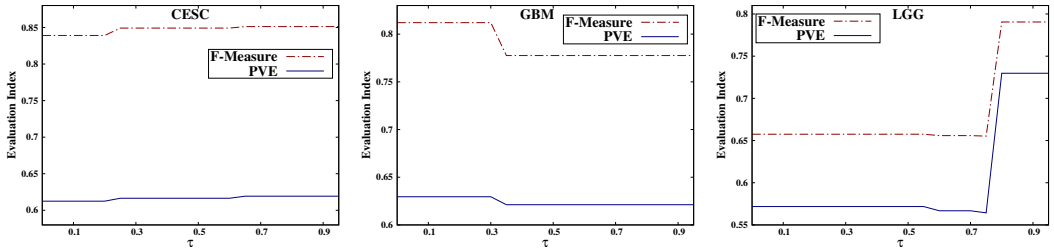

Fig. S1: Variation of PVE and F-measure for different values of threshold τ for CESC, GBM, and LGG data sets.

X_i is given by the percentage of variance explained (PVE) by a partition of its left subspace $U(X_i)$. Let the projection of the n samples in the k-dimensional left subspace $U(X_i)$ be given by $U(X_i) = \{x_1^i, \ldots, x_p^i, \ldots, x_n^i\}$, where $x_p^i \in \mathbb{R}^k$. Let $\mathcal{C}^i = \{C_1^i, \ldots, C_j^i, \ldots, C_k^i\}$ be a partition of $U(X_i)$ into k clusters. The PVE in $U(X_i)$ by the partition $\mathcal{C}^i$ is given by the ratio of between-cluster variance in $\mathcal{C}^i$ to the total variance of $U(X_i)$. The total variance is the total sum-of-squared distance of each sample from its mean, given by

$$\mathrm{T}(U(X_i)) = \sum_{p=1}^{n} ||x_p^i - \bar{x}^i||^2 \tag{4}$$

where $\bar{x}^i$ is the mean of $U(X_i)$. Since $U(X_i)$ contains principal subspace projection of data in X_i, the projection values in $U(X_i)$ have zero mean. Hence, $\bar{x}^i = 0$. Moreover, the columns of $U(X_i)$ are orthonormal to each other, therefore,

$$\begin{aligned} \mathrm{T}(U(X_i)) &= \sum_{p=1}^{n} ||x_p^i||^2 = ||U(X_i)||_F^2 \\ &= trace(U(X_i)^T U(X_i)) = trace(\mathbf{I}_k) = k, \end{aligned} \tag{5}$$

where $||A||_F^2$ denotes the Frobenius norm of matrix A. The within-cluster variance of partition $\mathcal{C}^i$ is the sum-of-squared distance of each data point from its cluster centroid, given by

$$\mathrm{W}_{\mathcal{C}^i}(U(X_i)) = \sum_{j=1}^{k} \sum_{x_p^i \in C_j^i} ||x_p^i - m_j||^2 \tag{6}$$

where m_j is the centroid of cluster C_j^i. The between-cluster variance in $\mathcal{C}^i$ is obtained by subtracting the within-cluster variance in $\mathcal{C}^i$ from the total variance of $U(X_i)$. Thus, the PVE in $U(X_i)$ by the partition $\mathcal{C}^i$ is given by

$$\mathrm{PVE}(U(X_i)) = \frac{\mathrm{T}(U(X_i)) - \mathrm{W}_{\mathcal{C}^i}(U(X_i))}{\mathrm{T}(U(X_i))}. \tag{7}$$

The relevance of a modality X_i is given by the PVE in $U(X_i)$ as follows:

$$\mathrm{Rel}(X_i) = \mathrm{PVE}(U(X_i)) = 1 - \frac{1}{k}\mathrm{W}_{\mathcal{C}^i}(U(X_i)). \tag{8}$$

The relevance measure gives a value in between 0 and 1 with higher value indicating better cluster structure.

B. Concordance

The concordance measure is based on the normalized mutual information (NMI) between the cluster assignments of two modalities. Let $\mathcal{C}^i$ and $\mathcal{C}^j$ be k-partitions of the subspaces $U(X_i)$ and $U(X_j)$, respectively. The concordance C between X_i and X_j is given by the NMI between the cluster solutions $\mathcal{C}^i$ and $\mathcal{C}^j$, given by

$$\mathsf{C}\,(X_i, X_j) = \mathrm{NMI}(\mathcal{C}^i, \mathcal{C}^j). \tag{9}$$

NMI is defined as follows:

$$\mathrm{NMI}(\mathcal{C}^i, \mathcal{C}^j) = \frac{2\ \mathbb{I}\left(\mathcal{C}^i, \mathcal{C}^j\right)}{[\mathbb{H}(\mathcal{C}^i) + \mathbb{H}(\mathcal{C}^j)]}; \tag{10}$$

where $\mathbb{H}(\mathcal{C}^i)$ is the entropy of $\mathcal{C}^i$ and $\mathbb{I}\left(\mathcal{C}^i, \mathcal{C}^j\right)$ is the mutual information between $\mathcal{C}^i$ and $\mathcal{C}^j$, which are as follows:

$$\mathbb{H}\left(\mathcal{C}^i\right) = -\sum_{p=1}^{k} Pr(C_p^i) \log Pr(C_p^i);$$

$$\mathbb{I}\left(\mathcal{C}^i, \mathcal{C}^j\right) = \sum_{p=1}^{k}\sum_{q=1}^{k} Pr(C_p^i \cap C_q^j) \log \left[\frac{Pr(C_p^i \cap C_q^j)}{Pr(C_p^i) Pr(C_q^j)}\right];$$

where $Pr(S)$ denotes the probability of the set S. The value of concordance C lies in the range $[0, 1]$, with larger value being indicative of more shared information between two modalities.

III. Experimental Results and Discussion

This section describes an unsupervised method for obtaining the optimal value of concordance threshold τ for the proposed SURE algorithm and empirically establishes the computational efficiency of SURE over principal component analysis (PCA).

A. Optimum Value of Concordance Threshold

The threshold parameter τ of the proposed SURE algorithm (given in Section III-C of the main paper) decides whether the remaining individual eigenspaces will be considered for updating the current joint eigenspace. At each iteration of joint eigenspace construction, the modality having maximum average concordance $\bar{\mathsf{C}}$, with respect to pre-selected modalities, is taken into consideration. The joint eigenspace is updated only if the value of $\bar{\mathsf{C}}$ is beyond some threshold τ. This threshold prevents modalities having low concordance or shared information with the previously updated ones from being integrated into the joint eigenspace. Given M modalities, different subsets of modalities get selected for different values of threshold τ. For each data set, the value of τ is varied in the range $[0, 0.95]$ at an interval of 0.05. For each value of threshold τ, the PVE by a k partition of the final joint subspace is evaluated, which is denoted by PVE_τ. The optimum value τ^* for each data set is chosen using the following relation:

$$\tau^* = \arg\max_{\tau}\ \{\mathrm{PVE}_\tau\}. \tag{11}$$

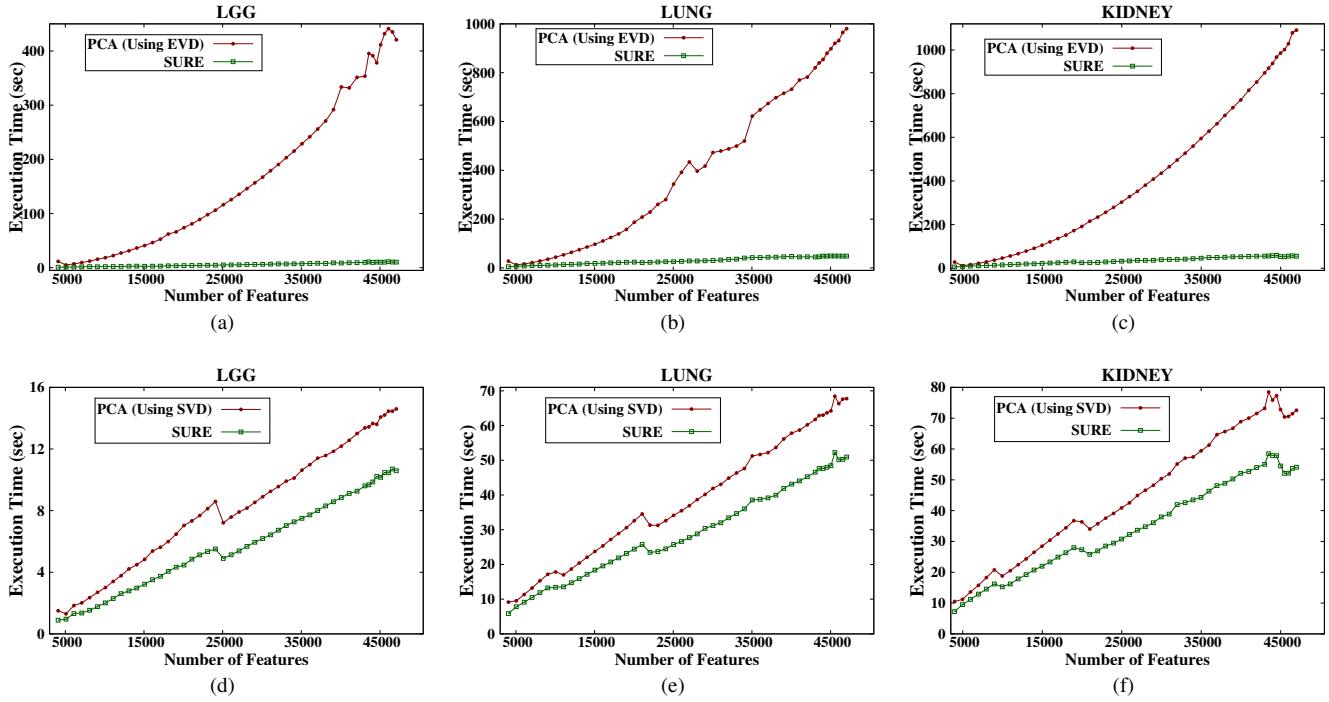

Fig. S2: Comparison of execution time for PCA computed using EVD (top row) and SVD (bottom row) and the proposed SURE approach on LGG, LUNG, and KIDNEY data sets.

It is worth noting that the upper bound for varying τ is 0.95 instead of 1.00. For $\tau = 1.00$, a candidate modality has to have full concordance or agreement in cluster structure with all the previously integrated ones. For real-life omics data sets, this is highly unlikely, and hence no candidate modality will ever get selected for updating the eigenspace. So, for $\tau = 1.00$, a unimodal solution, consisting of only the most relevant modality, will be considered always. As integration of multiple modalities can capture the biological variations across multiple genomic levels, the threshold τ is upper bounded at 0.95 in order to prefer selection of multiple modalities.

Fig. S1 shows the variation of F-measure and PVE for different values of τ for CESC, GBM, and LGG data sets, as examples. From Fig. S1, it is seen that the values of F-measure and PVE vary in a similar fashion with the change in τ. The PVE is calculated based on the generated clusters, while the F-measure is computed based on the ground truth subtype information. Since these two indices are found to vary similarly, the optimal value of τ inferred from PVE also gives the optimal value of F-measure, thus giving good clustering performance. For each data set, the best value of F-measure, obtained from all possible values of threshold τ, is compared with that obtained for optimal threshold τ^*. For all data sets, the best F-measure is exactly same with the F-measure corresponding to τ^*.

B. Execution Efficiency of SURE

One major advantage of the proposed algorithm is that it extracts the principal subspace of the integrated data matrix by iteratively updating the principal subspaces of the individual modalities, and its time complexity is $\mathcal{O}(n^2 d_{max})$. On the other hand, the time complexity of performing PCA on the integrated data matrix using eigenvalue decomposition (EVD) of the covariance matrix is $\mathcal{O}(d^3)$, while that using SVD of mean-centered data matrix is $\mathcal{O}(n^2 d)$, where $n << d_{max} << d$. This makes the proposed algorithm particularly efficient for PCA based dimensionality reduction of large multimodal data sets. Fig. S2 compares the execution time of the proposed SURE algorithm with that for extracting the principal components using EVD and SVD for LGG, LUNG, and KIDNEY data sets. The RNA and mDNA modalities have large number of features such as 20,502 and 25,978, respectively. The variation in execution time for extracting top k principal components using these three algorithms is observed by gradually increasing the number of features from RNA and mDNA modalities. The plots in Fig. S2(a)-(c) show that the execution time of PCA computed using EVD increases quadratically with respect the proposed SURE approach. This is because PCA using EVD takes $\mathcal{O}(d^3)$ time which is significantly higher compared to $\mathcal{O}(n^2 d_{max})$. Fig. S2(d)-(f) show 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 using SVD, especially for large data sets like LUNG and KIDNEY with 671 and 757 samples, respectively.

Availability and Implementation:
The R implementation of the proposed SURE algorithm, along with the description and statistical power of the multimodal data sets, and survival analysis, is available at www.isical.ac.in/~bibl/results/sure/sure.html.