Supplementary Material: Multi-Manifold Optimization for Multi-View Subspace Clustering

Aparajita Khan and Pradipta Maji

The main article introduces a manifold optimization based multi-view data clustering algorithm, termed as MiMIC (Multi-Manifold Integrative Clustering). In this document, Section S1 presents the proofs of Theorems 1 and 2 of the main paper. Section S2 theoretically proves the convergence of the proposed algorithm stated in Theorem 3 of the main paper, and the asymptotic convergence bound obtained in Theorem 4. Section S3 discusses the choice of convex combination used in this work to construct the joint Laplacian. The computational complexity of the proposed algorithm is reported in Section S4. A brief description of the benchmark and multi-omics data sets used in this work is provided in Section S5. Additional results on synthetic, benchmark, and multi-omics data sets are given in Section S6. Section S7 provides the definitions of four cluster evaluation indices used in this work to compare the performance of different algorithms.

S1. Proof of Theorem 1 and Theorem 2

This section presents the proofs of Theorems 1 and 2 of the main paper. The theorems establish that at each iteration, the next iterates $U_{\text{Joint}}^{(t+1)}$ and $U_j^{(t+1)}$ obtained by the proposed MiMIC algorithm belong to their respective manifolds.

Theorem 1. $U_{\text{Joint}}^{(t+1)}$ *belongs to the* k*-means manifold.*

Proof. For $U_{\text{Joint}}^{(t+1)}$ to belong to k-means manifold, denoted by Km, it must satisfy its properties given in (5) of the main paper. So, $U_{\text{Joint}}^{(t+1)}$ must have orthonormal columns:

$$\begin{aligned}
&\left(U_{\text{Joint}}^{(t+1)}\right)^T U_{\text{Joint}}^{(t+1)} \qquad\qquad \text{(from (15) of main paper)}\\
&=\left(\exp(B)\exp(Q')U_{\text{Joint}}^{(t)}\right)^T\left(\exp(B)\exp(Q')U_{\text{Joint}}^{(t)}\right)\\
&=\left(U_{\text{Joint}}^{(t)}\right)^T\exp(Q')^T\exp(B)^T\exp(B)\exp(Q')U_{\text{Joint}}^{(t)}\\
&=\left(U_{\text{Joint}}^{(t)}\right)^T\exp(-Q')\exp(-B)\exp(B)\exp(Q')U_{\text{Joint}}^{(t)}\\
&=\left(U_{\text{Joint}}^{(t)}\right)^T U_{\text{Joint}}^{(t)}=\mathbf{I}_r.
\end{aligned}$$

It can be shown that $U_{\text{Joint}}^{(t)}\left(U_{\text{Joint}}^{(t)}\right)^T$ commutes with $\exp(Q')$ [1] (see Lemma 1 for details). Hence,

$$\exp(Q')U_{\text{Joint}}^{(t)}\left(U_{\text{Joint}}^{(t)}\right)^T = U_{\text{Joint}}^{(t)}\left(U_{\text{Joint}}^{(t)}\right)^T\exp(Q').$$

Also, $\exp(B)\mathbf{1}=\exp(-B)\mathbf{1}=\mathbf{1}$. So,

$$\begin{aligned}
&U_{\text{Joint}}^{(t+1)}\left(U_{\text{Joint}}^{(t+1)}\right)^T\mathbf{1}\\
&=\left(\exp(B)\exp(Q')U_{\text{Joint}}^{(t)}\right)\left(\exp(B)\exp(Q')U_{\text{Joint}}^{(t)}\right)^T\mathbf{1}\\
&=\exp(B)\exp(Q')U_{\text{Joint}}^{(t)}\left(U_{\text{Joint}}^{(t)}\right)^T\exp(Q')^T\exp(B)^T\mathbf{1}\\
&=\exp(B)\exp(Q')U_{\text{Joint}}^{(t)}\left(U_{\text{Joint}}^{(t)}\right)^T\exp(-Q')\exp(-B)\mathbf{1}\\
&=\exp(B)U_{\text{Joint}}^{(t)}\left(U_{\text{Joint}}^{(t)}\right)^T\exp(Q')\exp(-Q')\exp(-B)\mathbf{1}\\
&=\exp(B)U_{\text{Joint}}^{(t)}\left(U_{\text{Joint}}^{(t)}\right)^T\mathbf{1}=\exp(B)\mathbf{1}=\mathbf{1}.
\end{aligned}$$

Thus, the next iterate $U_{\text{Joint}}^{(t+1)}$ satisfies both the properties of Km, and therefore, belongs to it. □

Theorem 2. $U_j^{(t+1)}$ *belongs to the Stiefel manifold.*

Proof. For $U_j^{(t+1)}$ to belong to the Stiefel manifold, it must satisfy its properties given by (11) of the main paper, that is, it must have orthonormal columns. The matrices $E_j^{(t+1)}$ and $V_j^{(t+1)}$, given by (18) of the main paper, contain the left and right singular vectors of $\boldsymbol{Z}_j^{(t+1)}$, respectively, which have onrthonormal columns. Therefore,

$$\left(U_j^{(t+1)}\right)^T U_j^{(t+1)} = V_j^{(t+1)}(E_j^{(t+1)})^T E_j^{(t+1)}(V_j^{(t+1)})^T=\mathbf{I}_r.$$

Thus, the next iterate of U_j belongs to the Stiefel manifold. □

In Theorem 1, the commutative property of $U_{\text{Joint}}^{(t)}\left(U_{\text{Joint}}^{(t)}\right)^T$ and $\exp(Q')$ is used to prove that $U_{\text{Joint}}^{(t+1)}$ belongs to the k-means manifold. The following lemma proves the commutative property [1].

Lemma 1. $U_{\text{Joint}}^{(t)}\left(U_{\text{Joint}}^{(t)}\right)^T$ *commutes with* $\exp(Q')$.

Proof. The t-th iterate of U_{Joint} belongs to the k-means manifold. So, from the properties of k-means manifold (defined in (5) of the main paper), it satisfies that

$$\left(U_{\text{Joint}}^{(t)}\right)^T U_{\text{Joint}}^{(t)}=\mathbf{I}_r. \tag{1}$$

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.

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.

Digital Object Identifier 10.1109/TNNLS.2021.3054789

From (15) of the main paper, we have

$$Q' = U_{\text{Joint}}^{(t)}\, Q\, \left(U_{\text{Joint}}^{(t)}\right)^T \in \Re^{n\times n}, \tag{2}$$

where $Q \in \Re^{r\times r}$. The exponential of Q' is given by [2]

$$\exp(Q') = \mathbf{I}_n + Q' + \frac{Q'^2}{2!} + \frac{Q'^3}{3!} + \ldots = \sum_{j=0}^{\infty} \frac{Q'^j}{j!}.$$

Now,

$$\begin{aligned}
& Q'\, U_{\text{Joint}}^{(t)} \left(U_{\text{Joint}}^{(t)}\right)^T \\
&= U_{\text{Joint}}^{(t)} Q \left(U_{\text{Joint}}^{(t)}\right)^T U_{\text{Joint}}^{(t)} \left(U_{\text{Joint}}^{(t)}\right)^T && \text{(from (2))} \\
&= U_{\text{Joint}}^{(t)} \left(U_{\text{Joint}}^{(t)}\right)^T U_{\text{Joint}}^{(t)} Q \left(U_{\text{Joint}}^{(t)}\right)^T && \text{(from (1))} \\
&= U_{\text{Joint}}^{(t)} \left(U_{\text{Joint}}^{(t)}\right)^T Q'. && (3)
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \exp(Q')\, U_{\text{Joint}}^{(t)} \left(U_{\text{Joint}}^{(t)}\right)^T \\
&= \left(\mathbf{I}_n + Q' + \frac{Q'^2}{2!} + \frac{Q'^3}{3!} + \ldots\right) U_{\text{Joint}}^{(t)} \left(U_{\text{Joint}}^{(t)}\right)^T \\
& \quad \text{(applying (13) repetatively)} \\
&= U_{\text{Joint}}^{(t)} \left(U_{\text{Joint}}^{(t)}\right)^T \left(\mathbf{I}_n + Q' + \frac{Q'^2}{2!} + \frac{Q'^3}{3!} + \ldots\right) \\
&= U_{\text{Joint}}^{(t)} \left(U_{\text{Joint}}^{(t)}\right)^T \exp(Q').
\end{aligned}$$

Hence, $U_{\text{Joint}}^{(t)} \left(U_{\text{Joint}}^{(t)}\right)^T$ commutes with $\exp(Q')$. □

S2. Convergence Analysis

The proposed MiMIC algorithm for multi-view data clustering is provided in Algorithm 3 of the main article. To prove its convergence, certain restrictions are imposed on the descent direction and choice of step size during optimization. Before discussing the convergence result and analyzing its asymptotic behavior, the retraction operation on a manifold (stated in Section III-B of the main article) and some important definitions are briefly stated below.

Given a manifold $\mathcal{M}$, a point $y \in \mathcal{M}$, let $T_y\mathcal{M}$ denote the tangent space of the manifold rooted at point y. Given a tangent $\xi \in T_y\mathcal{M}$, the retraction operation $\mathsf{R}_y(\xi)$ denotes the combination of two steps. First, movement along ξ to get the point $y+\xi$ in the tangent space. Second, projection of the point $y+\xi$ back to the manifold $\mathcal{M}$. For minimization of a function $f(y)$ over $\mathcal{M}$, given the current iterate $y^{(t)}$ at iterarion t, the update equation for line-search [3] on $\mathcal{M}$ is given by

$$y^{(t+1)} = \mathsf{R}_{y^{(t)}}(\eta d^{(t)}),$$

where $d^{(t)}$ is a descent direction and η is the step length. For the proposed MiMIC algorithm, while optimizing the joint objective $\boldsymbol{f}$ with respect to U_{Joint} over the k-means manifold Km, the set of update equations is given by (Section III-B1 of the main paper)

$$\begin{aligned}
\boldsymbol{Q}_{\text{Joint}}^{(t)} &= -\nabla_{U_{\text{Joint}}^{(t)}} \boldsymbol{f} \\
\boldsymbol{Z}_{\text{Joint}}^{(t)} &= \mathbf{\Pi}_{T_{U_{\text{Joint}}^{(t)}} \mathsf{Km}} \left(\boldsymbol{Q}_{\text{Joint}}^{(t)}\right) \\
\boldsymbol{Z}_{\text{Joint}}^{(t+1)} &= U_{\text{Joint}}^{(t)} + \eta \boldsymbol{Z}_{\text{Joint}}^{(t)} \\
U_{\text{Joint}}^{r(t+1)} &= \mathsf{PKm}_{U_{\text{Joint}}^{(t)}} \left(\boldsymbol{Z}_{\text{Joint}}^{(t+1)}\right),
\end{aligned} \tag{4}$$

where $U_{\text{Joint}}^{(t)}$ denotes the value of U_{Joint} at iteration t. The set of equations in (4) can be coupled using the retraction operation R and written as

$$U_{\text{Joint}}^{(t+1)} = \mathsf{RKm}_{U_{\text{Joint}}^{(t)}} \left(-\eta \nabla_{U_{\text{Joint}}^{(t)}} \boldsymbol{f}\right), \tag{5}$$

where RKm denotes retraction on the k-means manifold.

Some definitions related to the choice of descent direction and step length for the proposed MiMIC algorithm are stated next.

A. Background

Definition 1 (Gradient-related sequence). *Given a cost function f on a Riemannian manifold $\mathcal{M}$, a sequence $\{\xi^{(t)}\}$, where $\xi^{(t)} \in T_{y^{(t)}}\mathcal{M}$, is gradient-related if, for any subsequence $\{y^{(t)}\}_{t\in\tau}$ of $\{y^{(t)}\}$ that converges to a non-critical point of f, the corresponding subsequence $\{\xi^{(t)}\}_{t\in\tau}$ is bounded and satisfies*

$$\limsup_{t\to\infty,\ t\in\tau} \left\langle \nabla f(y^{(t)}), \xi^{(t)} \right\rangle < 0.$$

Here, $\langle .,. \rangle$ denotes the inner product. For a function f, descent direction at a point y refers to a vector moving along which leads to a reduction of the function. A direction ξ is a descent direction if the directional derivative along ξ is negative, that is,

$$\langle \nabla f(y), \xi \rangle < 0.$$

Definition 1 implies that a sequence of directions $\{\xi^{(t)}\}$ on the tangent space of $\mathcal{M}$ is gradient related if it contains a subsequence of descent directions of f. Thus, moving along a gradient-related sequence at each iteration would lead to a reduction of the function f.

To ensure the convergence of the proposed algorithm, the Armijo condition [4] is imposed on the choice of step size during the optimization. The condition is defined as follows:

Definition 2 (Armijo criterion). *Given a cost function f on a Riemannian manifold $\mathcal{M}$ with retraction R, a point $y \in \mathcal{M}$, a tangent vector $\xi \in T_y\mathcal{M}$, and scalars $\bar{\eta} > 0$ and $\sigma \in (0,1)$, the step length $\bar{\eta}$ is said to satisfy the Armijo condition restricted to the direction ξ if the following inequality holds:*

$$f(y) - f\big(\mathsf{R}_y(\bar{\eta}\xi)\big) \geq -\sigma\bar{\eta}\big\langle \nabla f(y), \xi \big\rangle. \tag{6}$$

The Armijo condition is a popular line-search condition that states that the reduction in f, given by $f(y) - f\big(\mathsf{R}_y(\bar{\eta}\xi)\big)$, should be proportional to both the step length $\bar{\eta}$ and the directional derivative $\big\langle \nabla f(y), \xi \big\rangle$, where $\sigma \in (0,1)$ is the constant of proportionality.

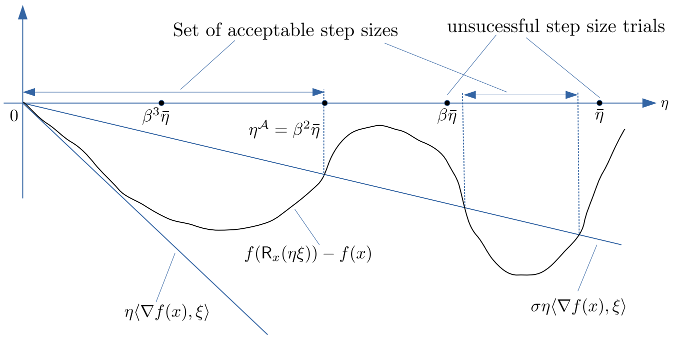

Set of acceptable step sizes
unsucessful step size trials
0
$\beta^3\bar{\eta}$
$\eta^{\mathcal{A}} = \beta^2\bar{\eta}$
$\beta\bar{\eta}$
$\bar{\eta}$
η
$f(\mathsf{R}_x(\eta\xi)) - f(x)$
$\eta\langle\nabla f(x), \xi\rangle$
$\sigma\eta\langle\nabla f(x), \xi\rangle$

Fig. S1. Armijo condition for the choice of step size.

Let $\boldsymbol{f}^{(t)}$ denote the value of the objective function $\boldsymbol{f}$ evaluated using $U^{(t)}_{\text{Joint}}$ and $U^{(t)}_j$'s, obtained at iteration t of the proposed algorithm. For the proposed algorithm, the step lengths for optimization on both the manifolds are chosen to be identical, that is, $\eta_{\mathsf{K}} = \eta_{\mathsf{S}} = \eta$. Also, the direction of movement on the tangent space is always the negative gradient $-\nabla\boldsymbol{f}$ (as in (13) and (16) of the main paper), and the retracted point from the tangent space gives the next iterate. Between two consecutive iterations, the reduction in the objective function $\boldsymbol{f}$ is given by $\boldsymbol{f}^{(t)} - \boldsymbol{f}^{(t+1)}$. Inorder to satisfy the Armijo criterion, this reduction must be proportional to the directional derivative. This is evaluated using

$$C_{\mathcal{A}} = \boldsymbol{f}^{(t)} - \boldsymbol{f}^{(t+1)} + \sigma\eta\left\langle\nabla\boldsymbol{f}, -\nabla\boldsymbol{f}\right\rangle. \tag{7}$$

Here, $C_{\mathcal{A}} \geq 0$ implies that the Armijo condition is satisfied and there has been a sufficient reduction in the value of the objective function. The proposed algorithm moves to the next iterate only when the Armijo criterion is satisfied. The value of Armijo parameter σ is set to $1e-05$ following [5].

Definition 3 (**Armijo point**). *Given a cost function f on a Riemannian manifold $\mathcal{M}$ with retraction R, a point $y \in \mathcal{M}$, a tangent vector $\xi \in T_y\mathcal{M}$, and scalars $\bar{\eta} > 0$, $\beta, \sigma \in (0, 1)$, the Armijo point is $\xi^{\mathcal{A}} = \eta^{\mathcal{A}}\xi = \beta^{\omega}\bar{\eta}\xi$, where ω is the smallest non-negative integer such that*

$$f(x) - f\big(\mathsf{R}_y(\beta^{\omega}\bar{\eta}\xi)\big) \geq -\sigma\big\langle\nabla f(y), \beta^{\omega}\bar{\eta}\xi\big\rangle.$$

The real number $\eta^{\mathcal{A}}$ is called the Armijo step size [3].

The smallest step size that satisfies the Armijo condition is called the Armijo step size $\eta^{\mathcal{A}}$. It is given by $\eta^{\mathcal{A}} = \beta^{\omega}\bar{\eta}$, such that ω is the smallest non-negative integer to achieve this for a given $\bar{\eta} > 0$ and $\beta \in (0, 1)$. Fig. S1 shows an example of the Armijo condition for choosing the step size. To choose a step size that satisfies the Armijo condition, we start with a step length $\bar{\eta} > 0$ and then check for the choices $\beta\bar{\eta}$, $\beta^2\bar{\eta}$, $\ldots$, until $\beta^{\omega}\bar{\eta}$ falls under the set of acceptable step sizes that satisfy (6). This choice of step size would give a sufficient decrease in the value of the function f.

B. Proof of Convergence

The convergence analysis of the proposed MiMIC algorithm (Algorithm 3 of the main paper) is as follows.

Theorem 3. *Every limit point of the sequence $\{U^{(t)}_{\text{Joint}}\}_{t=0,1,2,\ldots}$, generated by the proposed algorithm for a set of given U_j's for $j \in \{1, .., M\}$, is a critical point of the cost function $\boldsymbol{f}$.*

Proof. (By contradiction) Let there be a subsequence of iterations $\left\{U^{(t)}_{\text{Joint}}\right\}_{t\in\tau}$ that converges to some $U^{\star}_{\text{Joint}}$ which is not a critical point of $\boldsymbol{f}$, that is $\nabla_{U^{\star}_{\text{Joint}}}\boldsymbol{f} \neq 0$. The direction of movement at each iteration is the negative gradient along which the reduction of cost $\boldsymbol{f}$ is maximum. It follows that the whole sequence $\left\{\boldsymbol{f}\big(U^{(t)}_{\text{Joint}}\big)\right\}$ is non-increasing and converges to $\boldsymbol{f}\big(U^{\star}_{\text{Joint}}\big)$. So, the difference $\boldsymbol{f}\big(U^{(t)}_{\text{Joint}}\big) - \boldsymbol{f}\big(U^{(t+1)}_{\text{Joint}}\big)$ goes gradually to zero. The Armijo criterion $C_{\mathcal{A}}$, given by (7), is evaluated at each iteration of the proposed MiMIC algorithm. The algorithm proceeds to the next iteration only if $C_{\mathcal{A}} \geq 0$. The k-means manifold, over which U_{Joint} is optimized, is a Riemannian manifold with the inner product given by $\langle Z_1, Z_2\rangle = tr(Z_1^T Z_2)$ [1]. This relation can used to replace the trace term in (7). Furthermore, for a set of given U_j's for $j \in \{1, .., M\}$, $\boldsymbol{f}$ becomes a function of U_{Joint} only. In that case, the negative gradient becomes $-\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f} = \boldsymbol{Q}^{(t)}_{\text{Joint}}$ (see (13) of the main paper). Using (4), the Armijo criterion $C_{\mathcal{A}}$ can be written as

$$\begin{aligned} C_{\mathcal{A}} = \boldsymbol{f}\big(U^{(t)}_{\text{Joint}}\big) - \boldsymbol{f}\big(\mathsf{RKm}_{U^{(t)}_{\text{Joint}}}(\eta\boldsymbol{Q}^{(t)}_{\text{Joint}})\big) \\ + \sigma\eta^{(t)}\big\langle\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f}, \boldsymbol{Q}^{(t)}_{\text{Joint}}\big\rangle. \end{aligned} \tag{8}$$

The proposed MiMIC algorithm proceeds to the next iteration only if $C_{\mathcal{A}} \geq 0$, else it reduces the step size and checks again. Now, $C_{\mathcal{A}} \geq 0$ implies that at each iteration the proposed algorithm satisfies

$$\boldsymbol{f}\big(U^{(t)}_{\text{Joint}}\big) - \boldsymbol{f}\big(\mathsf{RKm}_{U^{(t)}_{\text{Joint}}}(\eta\boldsymbol{Q}^{(t)}_{\text{Joint}})\big) \geq -\sigma\eta^{(t)}\big\langle\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f}, \boldsymbol{Q}^{(t)}_{\text{Joint}}\big\rangle.$$

The direction of movement at each iteration is

$$\boldsymbol{Q}^{(t)}_{\text{Joint}} = -\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f}$$

which implies that

$$\big\langle\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f}, \boldsymbol{Q}^{(t)}_{\text{Joint}}\big\rangle = -\|\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f}\|_F^2 < 0, \tag{9}$$

where $\|\ .\ \|_F$ deontes the Frobenius norm of a matrix. Thus, the sequence movement directions $\{\boldsymbol{Q}^{(t)}_{\text{Joint}}\}$ is gradient related. Moreover, as $\left\{\boldsymbol{f}\big(U^{(t)}_{\text{Joint}}\big)\right\}$ is a convergent sequence, this implies that the step lengths $\{\eta^{(t)}\}_{t\in\tau} \to 0$. As the step lengths $\eta^{(t)}$'s are determined from the Armijo rule, it follows that for all t greater than some $\bar{t}$, $\eta^{(t)} = \beta^{\omega_t}\eta$, where ω_t is an integer greater than zero. Therefore, the update $\frac{\eta^{(t)}}{\beta} = \beta^{(\omega_t - 1)}\eta$ does not satisfy the Armijo condition. So,

$$\begin{aligned} \boldsymbol{f}\big(U^{(t)}_{\text{Joint}}\big) - \boldsymbol{f}\left(\mathsf{RKm}_{U^{(t)}_{\text{Joint}}}\left(\frac{\eta^{(t)}}{\beta}\boldsymbol{Q}^{(t)}_{\text{Joint}}\right)\right) \\ < -\sigma\frac{\eta^{(t)}}{\beta}\left\langle\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f}, \boldsymbol{Q}^{(t)}_{\text{Joint}}\right\rangle, \quad \forall t \in \tau, t \geq \bar{t}. \end{aligned} \tag{10}$$

Let

$$\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}} = \frac{\boldsymbol{Q}^{(t)}_{\text{Joint}}}{\|\boldsymbol{Q}^{(t)}_{\text{Joint}}\|} \quad \text{and} \quad \widehat{\eta}^{(t)} = \frac{\eta^{(t)}\|\boldsymbol{Q}^{(t)}_{\text{Joint}}\|}{\beta}.$$

For the function $\boldsymbol{f}$ over the manifold Km equipped with the retraction RKm, let $\widehat{\boldsymbol{f}} = \boldsymbol{f} \circ \mathsf{RKm}$ denote the pullback of $\boldsymbol{f}$ through RKm. For $U \in \mathsf{Km}$,

$$\widehat{\boldsymbol{f}}_U = \boldsymbol{f} \circ \mathsf{RKm}_U$$

denote the restriction of $\boldsymbol{f}$ to the tangent space $T_U\mathsf{Km}$. Denoting the zero element of tangent space $T_U\mathsf{Km}$ by $\mathbf{0}_U$, the inequality in (10) could be written as

$$\frac{\widehat{\boldsymbol{f}}_{U^{(t)}_{\text{Joint}}}\left(\mathbf{0}_{U^{(t)}_{\text{Joint}}}\right) - \widehat{\boldsymbol{f}}_{U^{(t)}_{\text{Joint}}}\left(\widehat{\eta}^{(t)}\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}\right)}{\widehat{\eta}^{(t)}} < -\sigma\left\langle\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f},\, \widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}\right\rangle, \tag{11}$$

$\forall t \in \tau$, where $t \geq \bar{t}$. The mean-value theorem is used to replace the left-hand side of (11) by the directional derivative of $\widehat{\boldsymbol{f}}$ at point $U^{(t)}_{\text{Joint}}$ in the direction of $\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}$ (see Chapter 3, [3]). So, for some $c \in [0, \widehat{\eta}^{(t)}]$, (11) can be written as

$$-\mathsf{D}\widehat{\boldsymbol{f}}_{U^{(t)}_{\text{Joint}}}\left(c\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}\right)\left[\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}\right] < -\sigma\left\langle\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f},\, \widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}\right\rangle, \tag{12}$$

$\forall t \in \tau$, where $t \geq \bar{t}$. Since $\{\eta^{(t)}\}_{t\in\tau} \to 0$ and $\boldsymbol{Q}^{(t)}_{\text{Joint}}$ is gradient-related, hence bounded, it follows that $\{\widehat{\eta}^{(t)}\}_{t\in\tau}$ also tends to 0. Moreover, as $\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}$ has unit norm, the set of unit norm vectors $\{\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}\}$ belongs to a compact set. Every sequence in a compact set converges to an element contained within the set. So, there must exist a index set $\widehat{\tau} \subset \tau$ such that $\{\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}\}_{t\in\widehat{\tau}} \to \widehat{\boldsymbol{Q}}^{\star}_{\text{Joint}}$ for some $\widehat{\boldsymbol{Q}}^{\star}_{\text{Joint}}$ having $\|\widehat{\boldsymbol{Q}}^{\star}_{\text{Joint}}\| = 1$. Taking limits in (12) over $\widehat{\tau}$, $\widehat{\eta}^{(t)} \to 0$, which implies that $c \to 0$ and $\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}} \to \widehat{\boldsymbol{Q}}^{\star}_{\text{Joint}}$. Also, $\boldsymbol{f}$ is a continuous and differentiable scalar field over the Riemannian manifold Km. Therefore, from the definition of directional derivative D (see (3.31) in Chapter 3, [3]), it satisfies that

$$\mathsf{D}\widehat{\boldsymbol{f}}_{U^{(t)}_{\text{Joint}}}(\mathbf{0})\left[\widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}\right] = \left\langle\nabla_{U^{(t)}_{\text{Joint}}}\boldsymbol{f}, \widehat{\boldsymbol{Q}}^{(t)}_{\text{Joint}}\right\rangle.$$

Taking limits, (12) becomes

$$-\left\langle\nabla_{U^{\star}_{\text{Joint}}}\boldsymbol{f}, \widehat{\boldsymbol{Q}}^{\star}_{\text{Joint}}\right\rangle < -\sigma\left\langle\nabla_{U^{\star}_{\text{Joint}}}\boldsymbol{f}, \widehat{\boldsymbol{Q}}^{\star}_{\text{Joint}}\right\rangle. \tag{13}$$

Since $0 < \sigma < 1$, it follows from (13) that

$$\left\langle\nabla_{U^{\star}_{\text{Joint}}}\boldsymbol{f}, \widehat{\boldsymbol{Q}}^{\star}_{\text{Joint}}\right\rangle > 0.$$

However, as $\{\boldsymbol{Q}^{(t)}_{\text{Joint}}\}$ is gradient related, therefore $\left\langle\nabla_{U^{\star}_{\text{Joint}}}\boldsymbol{f}, \widehat{\boldsymbol{Q}}^{\star}_{\text{Joint}}\right\rangle < 0$ (from (9)), which is a contradiction. Therefore, the subsequence of iterates $\left\{U^{(t)}_{\text{Joint}}\right\}_{t\in\tau}$ converges to some critical point of the objective function $\boldsymbol{f}$. □

C. Proof of Asymptotic Convergence Bound

The asymptotic behavior of the proposed MiMIC algorithm is studied in Theorem 4 of the main paper, which theoretically quantifies how fast the sequence of iterates generated by the proposed MiMIC algorithm converges to an optimal solution. For a sufficiently large value of t, Theorem 4 bounds the difference between the objective function $\boldsymbol{f}$ evaluated at the next iterate $(t+1)$ and at the optimal solution, in terms of the difference between $\boldsymbol{f}$ evaluated at the current iterate t and the optimal solution. This subsection proves the convergence bound obtained in Theorem 4 of the main paper.

Let $\{U^{(t)}_{\text{Joint}}\}_{t=0,1,2,\ldots}$ be an infinite sequence of iterates generated by Algorithm 3 of the main paper, for a set of given $\{U_j\}_{j=1}^{M}$. With the direction of movement being $\mathbf{Q}^{(t)}_{\text{Joint}} := -\nabla\boldsymbol{f}_{\text{Joint}^{(t)}}$, let the sequence $\{U^{(t)}_{\text{Joint}}\}_{t=0,1,\ldots}$ converge to a point $U^{\star}_{\text{Joint}}$, which is a critical point of $\boldsymbol{f}$ according to Theorem 3. Let $\mathbf{H}_{\hat{U}^{\star}_{\text{Joint}}}$ denote the Hessian matrix of $\hat{\boldsymbol{f}}$ at the converged solution $\hat{U}^{\star}_{\text{Joint}}$, and $\lambda_{\mathbf{H},\min}$ and $\lambda_{\mathbf{H},\max}$ be the smallest and largest eigenvalues of the Hessian of $\mathbf{H}_{U^{\star}_{\text{Joint}}}\boldsymbol{f}$. Assume that $\lambda_{\mathbf{H},\min} > 0$ (hence $U^{\star}_{\text{Joint}}$ is a local minimizer of $\boldsymbol{f}$). The asymptotic bound is stated as follows.

Theorem 4. *There exists an integer $t' \geq 0$ such that*

$$\boldsymbol{f}\left(U^{(t+1)}_{\text{Joint}}\right) - \boldsymbol{f}\left(U^{\star}_{\text{Joint}}\right) \leq c\left(\boldsymbol{f}\left(U^{(t)}_{\text{Joint}}\right) - \boldsymbol{f}\left(U^{\star}_{\text{Joint}}\right)\right),$$

for all $t \geq t'$, where

$$c = 1 - 2\sigma\lambda_{\mathbf{H},\min}\min\left(\eta, \frac{2\beta(1-\sigma)}{\lambda_{\mathbf{H},\max}}\right), \tag{14}$$

where η is the step length, and σ and β are Armijo criterion parameters.

Proof. Let $(\mathcal{U}, \varphi)$ be a chart of the manifold $\mathcal{M} := \mathsf{Km}(n,k)$, with $U^{\star}_{\text{Joint}} \in \mathcal{U}$. Let the negative gradient of $\boldsymbol{f}$ at any point $U \in \mathcal{M}$ be given by $\zeta_U := -\nabla\boldsymbol{f}(U)$, where ζ_U belongs to the tangent space $T_U\mathcal{M}$. Let coordinate expressions for different elements in the corresponding Euclidean space $\Re^{n\times k}$ be denoted with a hat. The following notations are used for Euclidean space representations.

$\hat{U} := \varphi(U)$ ▷ indicates that coordinate map $\hat{U}$ in $\Re^{n\times k}$ is equal to φ of U in $\mathcal{M}$,

$\hat{\mathcal{U}} := \varphi(\mathcal{U})$ ▷ similar to above notation, but for the whole set $\mathcal{U}$,

$\hat{\boldsymbol{f}}(\hat{U}) := \boldsymbol{f}(U)$ ▷ indicates that the value of $\hat{f}$ at $\hat{U} \in \Re^{n\times k}$ is equal to the value of f at $U \in \mathcal{M}$,

$\hat{\zeta}_{\hat{U}} := \mathsf{D}\varphi(U)[\zeta_U]$ ▷ $\hat{\zeta}_{\hat{U}}$ is the coordinate expression corresponding to the directional derivative in manifold $\mathcal{M}$,

$\hat{\mathsf{R}}_{\hat{U}}(\hat{\zeta}) := \varphi(\mathsf{R}_U(\zeta))$ ▷ the coordinate expression for the retracted point in $\Re^{n\times k}$ is given by the mapping φ of the retracted point $\mathsf{R}_U(\zeta)$ in $\mathcal{M}$.

Let $y_{\hat{U}}$ denote the Euclidean gradient of $\hat{\boldsymbol{f}}$ at $\hat{U}$, given by

$$y_{\hat{U}} = \begin{bmatrix} \partial_{11}\hat{\boldsymbol{f}}(\hat{U}) & \ldots & \partial_{1k}\hat{\boldsymbol{f}}(\hat{U}) \\ \partial_{21}\hat{\boldsymbol{f}}(\hat{U}) & \ldots & \partial_{2k}\hat{\boldsymbol{f}}(\hat{U}) \\ & \ldots & \\ \partial_{n1}\hat{\boldsymbol{f}}(\hat{U}) & \ldots & \partial_{nk}\hat{\boldsymbol{f}}(\hat{U}) \end{bmatrix}_{n\times k} \tag{15}$$

Let $G_{\hat{U}}$ denote the matrix representation of the Riemannian metric g of $\mathcal{M}$, in the coordinate space. Without loss of generality, we assume that the coordinate map of the critical point is $\hat{U}^\star_{\text{Joint}} = \mathbf{0}$ (the zero vector) and $G_{\hat{U}^\star_{\text{Joint}}} = \mathbf{I}_n$.

The main aim is to obtain, at a current iterate U, a suitable upper bound on $\boldsymbol{f}(\mathsf{R}_U(t^A\zeta_U))$, where t^A is the Armijo step and $t^A\zeta_U$ is the Armijo point in tangent space $\mathcal{T}_U\mathcal{M}$. The Armijo condition implies that

$$\begin{aligned} \boldsymbol{f}(U) - \boldsymbol{f}(\mathsf{R}_U(t^A\zeta_U)) &\geq -\sigma \left\langle \nabla\boldsymbol{f}(U), t^A\zeta_U \right\rangle, \\ \Rightarrow \boldsymbol{f}(\mathsf{R}_U(t^A\zeta_U)) &\leq \boldsymbol{f}(U) - \sigma\left\langle \zeta_U, t^A\zeta_U\right\rangle \\ &\leq \boldsymbol{f}(U) - \sigma t^A \left\langle \zeta_U,\zeta_U\right\rangle. \end{aligned} \tag{16}$$

First a lower bound is obtained on $\langle\zeta_U,\zeta_U\rangle$ in terms of $\boldsymbol{f}(U)$. Given a smooth scalar field $\boldsymbol{f}$ on Riemannian manifold $\mathcal{M}$, ζ_U denotes the negative gradient of $\boldsymbol{f}$ at U, given by $\zeta_U := -\nabla\boldsymbol{f}(U)$. The coordinate expression for ζ_U in $\Re^{n\times k}$ is given in terms of the Euclidean gradient $y_{\hat{U}}$ and the matrix representation of Riemannian metric G as follows (Section 3.6 in [3]):

$$\hat{\zeta}_{\hat{U}} = G^{-1}_{\hat{U}}(-y_{\hat{U}}).$$

Also, from (3.29) in [3],

$$\begin{aligned}\langle\zeta_U,\zeta_U\rangle &= \hat\zeta_{\hat U} G_{\hat U}\hat\zeta_{\hat U} = y_{\hat U} G^{-1}_{\hat U} y_{\hat U}\\ &=\| y_{\hat U}\|^2 \left(1+\mathcal{O}(\hat U)\right),\end{aligned} \tag{17}$$

as $G_{\hat U}$ is assumed to be the identity matrix at the critical point $\hat U^\star_{\text{Joint}}$. From Taylor expansion of the Euclidean gradient $y_{\hat U}$, we have

$$\begin{aligned}&\nabla\hat{\boldsymbol{f}}(\hat U^\star_{\text{Joint}}+\hat U) = \nabla\hat{\boldsymbol{f}}(\hat U^\star_{\text{Joint}}) + \mathbf{H}_{\hat U^\star_{\text{Joint}}}\,\hat U + \mathcal{O}(\|\hat U\|^2),\\ &\Rightarrow y_{\hat U} = \nabla\hat{\boldsymbol{f}}(\hat U) = \mathbf{H}_{\mathbf{0}}\,\hat U + \mathcal{O}(\|\hat U\|^2) \qquad (18)\\ &\qquad (\text{as } \hat U^\star_{\text{Joint}} = \mathbf{0}, \text{ so } \nabla\hat{\boldsymbol{f}}(\hat U^\star_{\text{Joint}}) = 0, \text{ and from (15)})\end{aligned}$$

On the other hand, from the Taylor expansion of $\hat{\boldsymbol{f}}$, we have

$$\begin{aligned}\hat{\boldsymbol{f}}(\hat U^\star_{\text{Joint}}+\hat U) &= \hat{\boldsymbol{f}}(\hat U^\star_{\text{Joint}}) + \left(\nabla\hat{\boldsymbol{f}}(\hat U^\star_{\text{Joint}})\right)^T\hat U\\ &\quad+\frac{1}{2}\hat U^T\,\mathbf{H}_{\hat U^\star_{\text{Joint}}}\,\hat U + \mathcal{O}(\|\hat U\|^3),\\ \Rightarrow \hat{\boldsymbol{f}}(\hat U) &= \hat{\boldsymbol{f}}(\mathbf{0}) + \frac{1}{2}\hat U^T\,\mathbf{H}_{\mathbf{0}}\,\hat U + \mathcal{O}(\|\hat U\|^3). \qquad (19)\\ &\qquad (\text{applying } \hat U^\star_{\text{Joint}} = \mathbf{0} \text{ and } \nabla\hat{\boldsymbol{f}}(\hat U^\star_{\text{Joint}}) = 0)\end{aligned}$$

It follows from (18) and (19) that

$$\begin{aligned}\hat{\boldsymbol{f}}(\hat U) - \hat{\boldsymbol{f}}(\mathbf{0}) &= \frac{1}{2}y^T_{\hat U}\,\mathbf{H}^{-1}_{\mathbf{0}}\,y_{\hat U} + \mathcal{O}(\|\hat U\|^3)\\ &\leq \frac{1}{2}\frac{1}{\lambda_{\mathbf{H},\min}}\,\|y_{\hat U}\|,\end{aligned} \tag{20}$$

holds for all $\hat U$ sufficiently close to $\hat U^\star_{\text{Joint}}$. This is because, in (20) above, $\lambda_{\mathbf{H},\min}$ denotes the minimum eigenvalue of the Hessian of $\hat{\boldsymbol{f}}$ at $\hat U^\star_{\text{Joint}}$, that is $\mathbf{H}_{\mathbf{0}}$, and from the properties of eigenvalues, we have that, for any vector v, $v^T\mathbf{H}^{-1}_{\mathbf{0}}v \leq (\lambda_{\mathbf{H},\min})^{-1}$. Therefore, from (17) and (20), it can be concluded that

$$\begin{aligned}&\boldsymbol{f}(U) - \boldsymbol{f}(U^\star_{\text{Joint}}) \leq \frac{1}{2}\frac{1}{\lambda_{H,\min}}\left\langle\zeta_U,\zeta_U\right\rangle,\\ &\Rightarrow 2\lambda_{\mathbf{H},\min}\left(\boldsymbol{f}(U) - \boldsymbol{f}(U^\star_{\text{Joint}})\right) \leq \left\langle\zeta_U,\zeta_U\right\rangle.\end{aligned} \tag{21}$$

Thus, (21) gives the desired lower bound on $\langle\zeta_U,\zeta_U\rangle$. Using the bound (21) in the Armijo condition (16) gives us that

$$\begin{aligned}&\boldsymbol{f}(\mathsf{R}_U(t^A\zeta_U)) \leq \boldsymbol{f}(U) - \sigma t^A 2\lambda_{\mathbf{H},\min}\left(\boldsymbol{f}(U) - \boldsymbol{f}(U^\star_{\text{Joint}})\right),\\ &\Rightarrow f(\mathsf{R}_x(t^A\zeta_U)) - \boldsymbol{f}(U^\star_{\text{Joint}})\\ &\qquad\qquad \leq \left(1 - 2\lambda_{\mathbf{H},\min}\sigma t^A\right)\left(\boldsymbol{f}(U) - \boldsymbol{f}(U^\star_{\text{Joint}})\right).\end{aligned} \tag{22}$$

Next a lower bound is obtained on the Armijo step size t^A to substitute in (22). Using the retraction operator R and the of negative gradient ζ_U, we can define a smooth curve on the manifold, from $\Re$ to $\mathcal{M}$, given by $t \to \mathsf{R}_U(t\zeta_U)$. This mapping can be further used to define a smooth function h on $\mathcal{M}$ from $\Re$ to $\Re$ with a well-defined classical derivative, given by

$$h_U(t) = \boldsymbol{f}\left(\mathsf{R}_U(t\zeta_U)\right). \tag{23}$$

The derivative of h_U is given by (Sections 3.5.1, 3.5.2, and 3.6 of [3])

$$\begin{aligned}\dot h_U(t=0) &= \frac{\mathsf{d}}{\mathsf{d}t}\boldsymbol{f}\left(\mathsf{R}_U(t\zeta_U)\right)\Big|_{t=0} = \mathsf{D}\boldsymbol{f}(U)[-\zeta_U]\\ &= \langle\nabla\boldsymbol{f}(U),\zeta_U\rangle = -\langle\zeta_U,\zeta_U\rangle.\end{aligned} \tag{24}$$

Using (23) and (24) the Armijo condition (16) reads

$$h_U(t^A) \leq h_U(0) + \sigma t^A\dot h_U(0). \tag{25}$$

The Taylor expansion of h_U gives us that

$$h_U(t) = h_U(0) + t\dot h_U(0) + t^2\frac{\ddot h_U(0)}{2}.$$

The t at which the left- and right-hand sides of (25) are equal is given by

$$\begin{aligned}h_U(0) + t\dot h_U(0) + t^2\frac{\ddot h_U(0)}{2} &= h_U(0) + \sigma t\dot h_U(0),\\ \Rightarrow t\frac{\ddot h_U(0)}{2} &= -\dot h_U(0) + \sigma\dot h_U(0),\\ \Rightarrow t &= \frac{-2(1-\sigma)\dot h_U(0)}{\ddot h_U(0)}.\end{aligned} \tag{26}$$

Using t in (26) and the definition of Armijo point (Definition 3 and Section 4.2 of [3]), the step size t^A that satisfies (22) has the following lower bound

$$t^A \geq \min\left(\eta, \frac{-2\beta(1-\sigma)\dot h_U(0)}{\ddot h_U(0)}\right), \tag{27}$$

where $\bar\eta$ and β are Armijo step size parameters. The second derivative $\ddot h_u$ is given by

$$\begin{aligned}\ddot h_U(t=0) &= \frac{\mathsf{d}^2}{\mathsf{d}t^2}f\left(\mathsf{R}_U(t\zeta_U)\right)\Big|_{t=0} = \mathsf{D}^2 f(x)[-\zeta_U]\\ &= (-\zeta_U)^T\mathbf{H}_0(-\zeta_U) = \mathbf{H}_{\mathbf{0}}\,\|\zeta_U\|^2.\end{aligned} \tag{28}$$

From properties of eigenvalues, we have that for any vector v, $v^T\mathbf{H}_{\mathbf{0}}v \leq \lambda_{\mathbf{H},\max}$. Therefore, using (24) and (28) in (27) gives that

$$t^A \geq \min\left(\eta, \frac{2\beta(1-\sigma)}{\lambda_{\mathbf{H},\max}}\right), \tag{29}$$

for all U sufficiently close to $U^\star_{\text{Joint}}$. Using the lower bound (29) in (22) gives

$$\boldsymbol{f}(\mathsf{R}_U(t^A\zeta_U)) - \boldsymbol{f}(U^\star_{\text{Joint}}) \leq c\left(\boldsymbol{f}(U) - f(U^\star_{\text{Joint}})\right) \tag{30}$$

where

$$c = 1 - 2\sigma\lambda_{\mathbf{H},\min} \min\left(\eta, \frac{2\beta(1-\sigma)}{\lambda_{\mathbf{H},\max}}\right). \tag{31}$$

In (30), t^A is the Armio step size corresponding to the Armijo point. When the next iterate $U_{\text{Joint}}^{(t+1)}$ is the Armio point, then the decrease in the value of the objective function from $U_{\text{Joint}}^{(t)}$ to $U_{\text{Joint}}^{(t+1)}$ is σ times the directional derivative at $U_{\text{Joint}}^{(t)}$. In Algorithm 3, the next iterate is

$$U_{\text{Joint}}^{(t+1)} = \mathsf{R}_{U_{\text{Joint}}^{(t)}}\left(t\zeta_{U_{\text{Joint}}^{(t+1)}}\right), \tag{32}$$

where t satisfies the Armijo condition, that is, with step length t, the decrease in the value of the objective function is greater than or equal to σ times the directional derivative at $U_{\text{Joint}}^{(t)}$. Hence using (32) in (30), we get

$$\boldsymbol{f}(U_{\text{Joint}}^{(t+1)}) - \boldsymbol{f}(U_{\text{Joint}}^{\star}) \leq c\left(\boldsymbol{f}(U_{\text{Joint}}^{(t)}) - \boldsymbol{f}(U_{\text{Joint}}^{\star})\right),$$

where c is given by (31). □

S3. Choice of Convex Combination

The Fiedler value of a graph G, denoted by λ_2, refers to the second largest eigenvalue of its Laplacian L defined according to (2) of the main paper. It indicates the separability of the graph into two component subgraphs. Higher Fiedler value is indicative of easier separability. The corresponding eigenvector, say u_2, called the Fiedler vector, can be used to partition the vertices of G [6]. One popular way is to apply the 2-means algorithm on u_2 to obtain 2-way partition of graph G. The Silhouette index [7] can then internally assess the quality of this partition. Let $\mathcal{S}(u_2)$ denote the value of the Silhouette index evaluated on a 2-partition of the Fiedler vector u_2. A higher value of this index indicates a better partition. A view with good cluster information is expected to have a higher Fiedler value as well as higher Silhouette index on the Fiedler vector. The "relevance" of a view X_m is defined by [8]

$$\boldsymbol{\chi}_m = \frac{1}{4}\lambda_2^m\left[\mathcal{S}(u_2^m) + 1\right], \tag{33}$$

where λ_2^m is the second largest eigenvalue of its Laplacian L_m and u_2^m is the corresponding eigenvector. The value of $\boldsymbol{\chi}$ lies within $[0,1]$ and a higher value of $\boldsymbol{\chi}_m$ implies better cluster structure. Thus, a linear ordering of the views can be obtained based on the relevance. Let $X_{(1)}, \ldots, X_{(m)}, \ldots, X_{(M)}$ be the ordering of $X_1, \ldots, X_m, \ldots, X_M$ based on decreasing value of relevance $\boldsymbol{\chi}$.

In the convex combination vector $\boldsymbol{\alpha}$ of (7) of the main paper, the component $\alpha_{(m)}$ represents the weighting factor of view $X_{(m)}$ and is given by [8]

$$\alpha_{(m)} = \boldsymbol{\chi}_{(m)}\Delta^{-m}, \text{ where } \Delta \geq 1. \tag{34}$$

This implies that based on the index of $X_{(m)}$ in the ordering $X_{(1)}, \ldots, X_{(M)}$, the relevance value of $X_{(m)}$ is damped by a factor of Δ^m and then used as its contribution in the convex combination $\boldsymbol{\alpha}$. Thus, in $\boldsymbol{\alpha}$, the most relevant view has contribution of $\frac{\boldsymbol{\chi}_{(1)}}{\Delta}$, while the second most relevant one contributes $\frac{\boldsymbol{\chi}_{(2)}}{\Delta^2}$, and so on. This assignment of $\boldsymbol{\alpha}$ upweights views with better cluster structure, while dampens the effect of those having poorer structure. In this work, the value of Δ is empirically set to 1 for the benchmark and synthetic data sets, and 2 for the multi-omics data sets.

S4. Computational Complexity

Let $X_1, \ldots, X_m, \ldots, X_M$, where $X_m \in \Re^{n \times d_m}$, be M different views of a multi-view data set, all measured on the same set of n samples. The number of clusters in the data set is assumed to be known and is denoted by k, and let r be the rank of joint and individual subspaces, U_{Joint} and U_js, which is given as input to the proposed Algorithm 3 of the main paper. Given the similarity matrix W_m for modality X_m, its graph Laplacian L_m is computed in step 2 in $\mathcal{O}(n^2)$ time. Then, the eigen-decomposition of L_m is computed in step 3 which takes $\mathcal{O}(n^3)$ time for the $(n \times n)$ matrix. The computation of relevance $\boldsymbol{\chi}_m$ in step 5 involves computation of Silhouette index which has pair-wise distance computation and takes $\mathcal{O}(n^2)$ time. For M views, the total complexity of steps 1−6 is bounded by $\mathcal{O}(Mn^3)$. The computation of joint Laplacian and its eigen-decomposition in steps 7 and 8, respectively, takes atmost $\mathcal{O}(n^3)$ time. Steps 9−10 are initializations, which take constant time. For a fixed j, optimization of U_j over Stiefel manifold takes $\mathcal{O}(n^2r)$ time. The loop for j in step 12 runs once for each of the M views, which contributes to a total complexity of $\mathcal{O}(Mn^2r)$ for steps 12−14. The optimization of U_{Joint} over k-means manifold in step 15 has $\mathcal{O}(n^3)$ time complexity due to the matrix exponential based retraction operation. The computation of the joint objective in step 16 takes $\mathcal{O}(Mn^2r)$ time. The evaluation of convergence criteria and variable updation in steps 17−21 takes $\mathcal{O}(1)$ time. Assuming that the algorithm takes t iterations to converge, the overall complexity of steps 11−22 is bounded by $\mathcal{O}(t\max\{n^3, Mn^2r\})$. The clustering on the final solution $U_{\text{Joint}}^{\star}$ in step 24 takes $\mathcal{O}(t_{km}nk^2)$ time, where t_{km} is the maximum number of iterations k-means clustering executes.

Hence, the overall computational complexity of the proposed MiMIC algorithm, to extract the subspace $U_{\text{Joint}}^{\star}$ and perform clustering, is $(\mathcal{O}(Mn^3 + t\max\{n^3, Mn^2r\} + t_{km}nk^2) =)\mathcal{O}(tn^3)$, assuming $M, r, k << n$.

S5. Description of Data Sets

This section presents the description of five benchmark data sets and four multi-omics cancer data sets, which are used in this work.

A. Benchmark Data Sets

Five benchmark data sets from different application domains like information retrieval, handwritten digits identification, and object detection are considered in this work. The data sets are briefly described as follows.

i. **Digits**: This data set consists of features of handwritten numerals ('0'-'9') extracted from a collection of Dutch utility maps with 200 patterns per class (for a total of 2,000 patterns), digitized in the form of binary images. The data set is publicly available at https://archive.ics.uci.edu/ml/datasets/Multiple+Features. The samples are represented in terms of the following six feature sets:

a) mfeat-fou: 76 Fourier coefficients corresponding to the character shapes.
b) mfeat-fac: 216 profile correlations.
c) mfeat-kar: 64 Karhunen-Love coefficients.
d) mfeat-pix: 240 pixel averages in 2 x 3 windows.
e) mfeat-zer: 47 Zernike moments.
f) mfeat-mor: 6 morphological features.

ii. **3Sources**: This is a multi-view text data set available at http://mlg.ucd.ie/datasets/3sources.html. It consists of 169 news articles collected from three well-known online news sources: BBC, Reuters, and The Guardian, from the period February to April 2009. Each news article story was manually annotated with one or more of the six topical labels: business, entertainment, health, politics, sport, and technology. The labels roughly correspond to the primary section headings used across the three news sources. The data set has three views, one corresponding of each of the three news sources.

iii. **BBC**: This is also a multi-view news article clustering data set constructed from the single-view BBC news corpora http://mlg.ucd.ie/datasets/segment.html. It consists of 685 news documents. Each raw document was split into four segments by separating the documents into paragraphs, and merging sequences of consecutive paragraphs. The segments for each document were then randomly assigned to views. Each document is annotated with one of the five topical labels: business, entertainment, politics, sport, and technology. The data set has four views corresponding to the four segments.

iv. **100Leaves**: It is a one-hundred plant species leaves data set https://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set. The data set consists of 1,600 samples, with sixteen samples of each type of leaf for each of the one-hundred plant species. Each sample is represented by three sets of image features: shape descriptors, fine scale margin, and texture histogram.

v. **ALOI**: This is the Amsterdam Library of Object Image data set http://elki.dbs.ifi.lmu.de/wiki/DataSets/MultiView. The data set is from the work of [9]. The data set consists of 11,025 images of 100 small objects. Each image is represented with four types of features, that is, RGB color histogram, HSV color histogram, color similiarity and Haralick features.

B. Multi-Omics Data Sets

Four real-life multi-omics cancer data sets, from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/), are used in this study, namely, lower grade glioma (LGG), stomach adenocarcinome (STAD), breast invasive carcinoma (BRCA), and lung carcinoma (LUNG). Four genomic views considered for these data sets are DNA methylation (mDNA), gene expression (RNA), miRNA expression (miRNA), and reverse phase protein array expression (RPPA). The cancer data sets are downloaded from https://portal.gdc.cancer.gov/. The four TCGA data sets used in this study are described as follows:

i. 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 [10]. 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.

ii. Stomach adenocarcinoma (**STAD**): Stomach/gastric cancer was the worlds third leading cause of cancer mortality in 2012, responsible for 723,000 deaths [11]. TCGA research network has proposed a molecular classification dividing gastric cancer into four subtypes [12]. The STAD data set has 242 samples which consists of 54 samples from microsatellite unstable tumours, which show elevated mutation rates, 21 samples of tumours showing positivity for EpsteinBarr virus, 119 samples of tumours having chromosomal instability, and 48 samples of genomically stable tumors.

iii. 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 [13]. During the last 15 years, four intrinsic molecular subtypes of breast cancer, namely, Luminal A, Luminal B, HER2-enriched, and Basal-like subtypes have been identified and intensively studied [13], [14], [15]. 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.

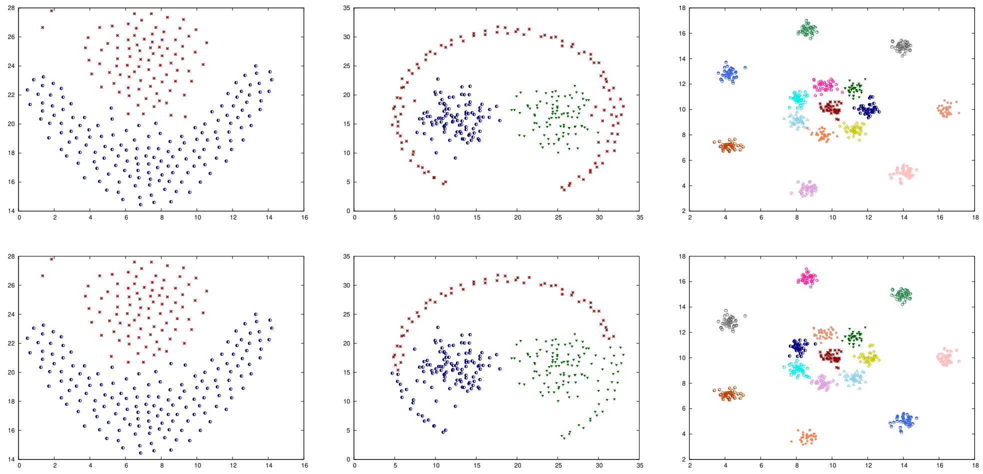
iv. Lung Carcinoma (**LUNG**): Based on the primary site of origin, lung cancer set can be categorized in two subtypes, namely, adenocarcinoma and squamous cell carcinoma. These were also the two major subtypes of lung cancer in 2015 WHO classification [16]. The LUNG data set consists of 671 samples with 360 samples of lung adenocarcinoma and 311 samples of lung squamous cell carcinoma.

S6. Experimental Results and Discussion

For the existing algorithms on benchmark data sets, results reported in their original papers and in [17] (which also compares the performance of the existing algorithms on the same data sets) are used. For the omics data sets, original implementations of the existing approaches were obtained from the authors and executed at their default parameter settings. The experimental setup for the existing algorithms on multi-omics cancer data sets is provided in the supplementary material of [8]. This section empirically establishes the significance of the asymptotic convergence bound obtained in Section IV of the main paper. Experimental results that demonstrate the importance of multi-view integration over uni-view analysis, along with the results on four additional benchmark data sets, are also provided.

TABLE S1
PERFORMANCE ANALYSIS OF PROPOSED ALGORITHMS ON SYNTHETIC CLUSTERING DATA SETS

Data Sets→	**Aggregation**	**Compound**	**Pathbased**	**Spiral**	**Jain**	**Flame**	**R15**	**D31**
No. of Samples	788	399	300	312	373	240	600	3100
No. of Clusters	7	6	3	3	2	2	15	31
Accuracy	0.99619	0.89473	0.83000	1.00	1.00	0.98750	0.99500	0.82032
NMI	0.98839	0.89671	0.63926	1.00	1.00	0.89905	0.99135	0.91780
ARI	0.99198	0.92926	0.58486	1.00	1.00	0.95014	0.98921	0.68092
F-measure	0.99622	0.91264	0.81500	1.00	1.00	0.98748	0.99497	0.85341



(a) Flame (b) Pathbased (c) R15

Fig. S2. Two-dimensional scatter plots of three synthetic shape data sets: ground truth clustering (top row) and MiMIC clustering (bottom row). Clustering accuracy:- (a) Flame: 1.00, (b) Pathbased: 0.83, (c) R15: 0.995.

A. Results on Synthetic Data Sets

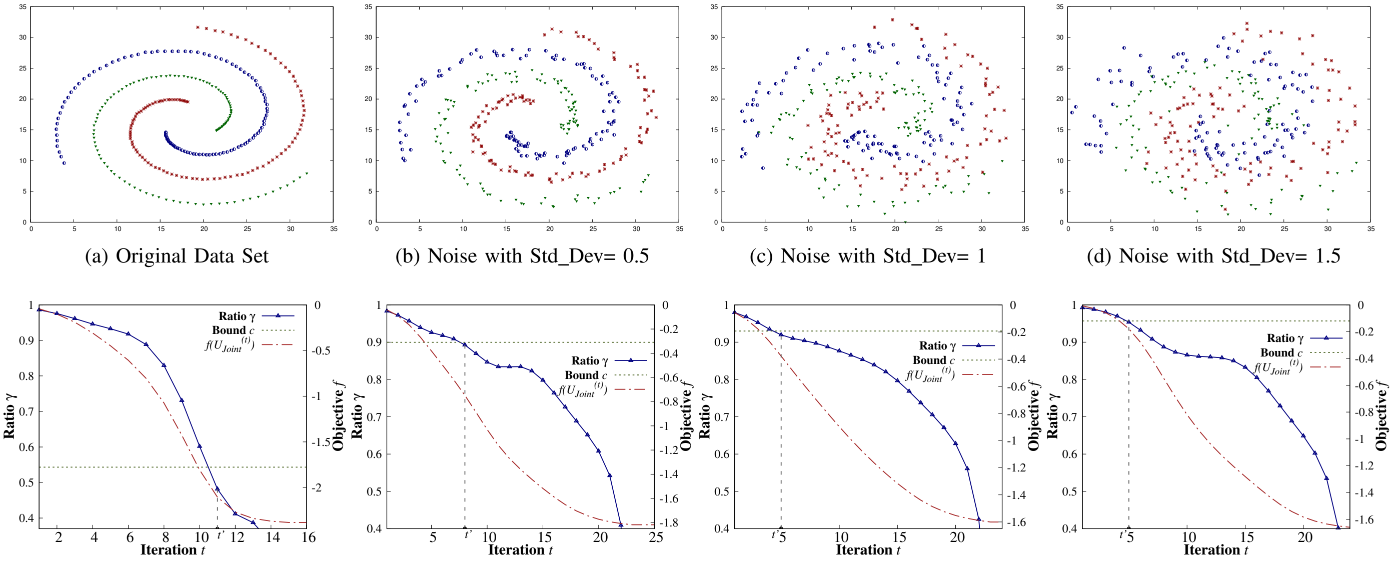
The results on five two-dimensional synthetic shape data sets are reported in Section V-B and Fig. 2 of the main paper. The quantitative results on those data sets, in terms of accuracy, adjusted rand index (ARI), normalized mutual information (NMI), and F-measure are reported in Table S1. Apart from the five synthetic data sets used in main paper, Table S1 also reports the results on three additional shape data sets, namely, Flame, Pathbased, and R15. These data sets are also available at http://cs.joensuu.fi/sipu/datasets/, along with the five other data sets used in the main paper. The scatter plots for the additional three data sets are provided in Fig. S2 for visual analysis. The results in Table S1 show that the proposed algorithm achieves perfect or nearly perfect clustering on five data sets, namely, Spiral, Jain, Aggregation, R15, and Flame. For other three data sets, the clustering performance is quiet high, with clustering accuracy above 0.8. The scatter plots in Fig. S2(b) show that for the Pathbased data set, the surrounding cluster marked in red has become mixed up with the interior green and blue clusters through the boundary points, resulting in reduced performance. However, for Flame and R15 data sets, Figs. S2(a) and S2(c) show that the proposed algorithm has achieved nearly perfect clustering. Several shape data sets studied in this work lack linearly separable cluster patterns. The results in Table S1, Fig. S2, and Fig. 2 of the main paper demonstrate that the proposed algorithm can efficiently identify non-linearly separable clusters.

B. Significance of Asymptotic Convergence Bound

The asymptotic convergence bound obtained in Theorem 4 of the main paper indicates how fast the sequence of iterates generated by the proposed algorithm converges to an optimal solution of a given data set. For a sufficiently large value of iteration number t, Theorem 4 bounds the difference between the cost function $\boldsymbol{f}$ evaluated at $U_{\text{Joint}}^{(t+1)}$ and at the optimal solution $U_{\text{Joint}}^{\star}$ in terms of the difference between that evaluated at $U_{\text{Joint}}^{(t)}$ and $U_{\text{Joint}}^{\star}$. Let γ_t be given by the ratio

$$\gamma_t = \frac{\boldsymbol{f}\left(U_{\text{Joint}}^{(t+1)}\right) - \boldsymbol{f}\left(U_{\text{Joint}}^{\star}\right)}{\boldsymbol{f}\left(U_{\text{Joint}}^{(t)}\right) - \boldsymbol{f}\left(U_{\text{Joint}}^{\star}\right)}. \tag{35}$$

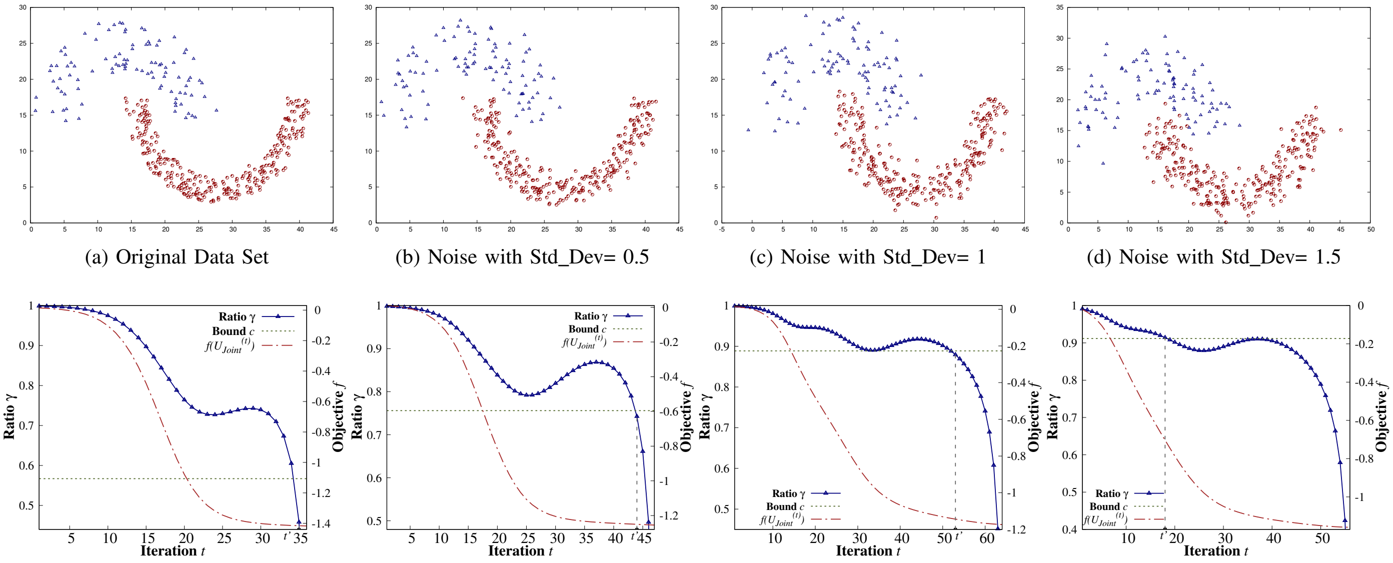
Theorem 4 states that for all t greater or equal to some t', $\gamma_t \leq c$, where c is given by (14). The convergence factor c can be used to make inference about the underlying cluster structure of the data set. As discussed in Section IV of the main paper, a value of c close to 1 indicates poor separation between the clusters present in the data set, while a value much lower than 1 indicates well-separated clusters. To experimentally establish this, multiple noisy data sets are generated from the synthetic shape data sets used in this work, by adding Gaussian noise of mean 0 and standard deviations 0.5, 1, and 1.5. Experiments are performed on noise-free and noisy variations of four shape data sets from http://cs.joensuu.fi/sipu/datasets/, namely, Spiral, Jain, R15, and Compound. The scatter plots for the noise-free and noisy variants of Spiral, Jain, R15, and



(a) Original Data Set (b) Noise with Std_Dev= 0.5 (c) Noise with Std_Dev= 1 (d) Noise with Std_Dev= 1.5

(a) c= 0.5432315 (b) c= 0.8995148 (c) c= 0.9298075 (d) c= 0.9569399

Fig. S3. Asymptotic convergence analysis for Spiral data set: scatter plot of data with varying Gaussian noise (top row) and variation of convergence ratio and objective function with increase in iteration number t (bottom row).



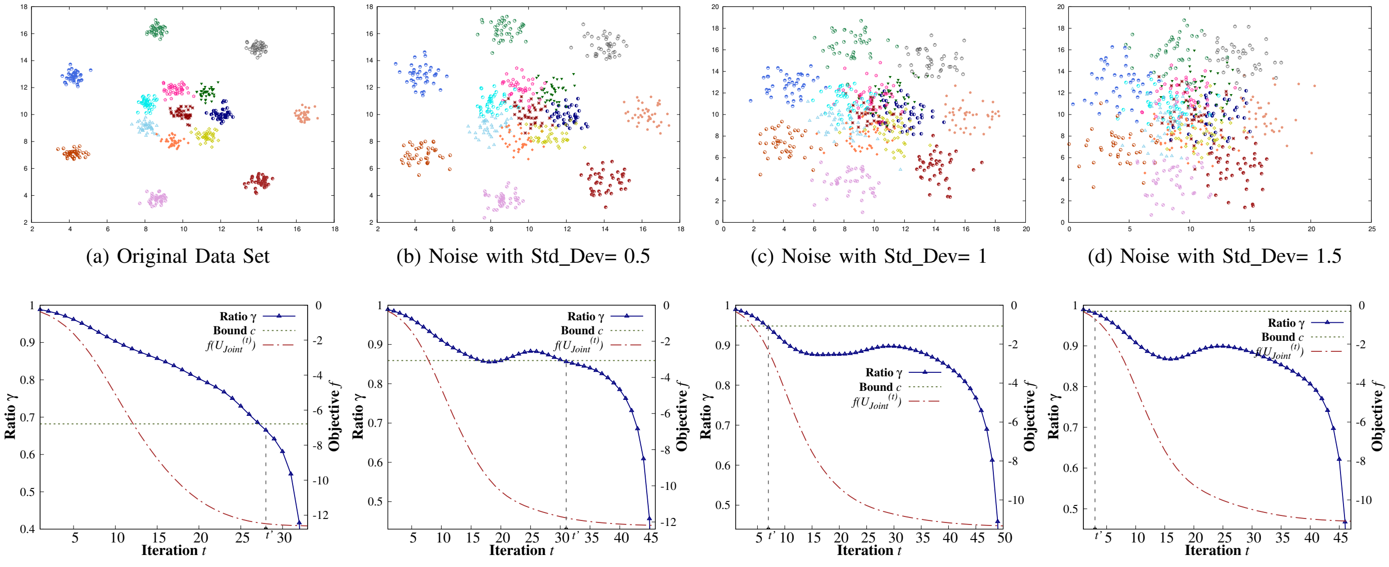
(a) Original Data Set (b) Noise with Std_Dev= 0.5 (c) Noise with Std_Dev= 1 (d) Noise with Std_Dev= 1.5

(a) c= 0.5669898 (b) c= 0.7558949 (c) c= 0.8887072 (d) c= 0.9117535

Fig. S4. Asymptotic convergence analysis for Jain data set: scatter plot of data with varying Gaussian noise (top row) and variation of convergence ratio and objective function with increase in iteration number t (bottom row).

Compound data sets are provided in the top rows of Fig. S3, S4, S5, and S6, respectively. As stated in Section V-A1 of the main paper, for each variant of each data set two views are generated using k-nearest neighbors and Gaussian kernel. Starting from a random initial iterate, the variation of γ_t and the cost function $\boldsymbol{f}\left(U_{\text{Joint}}^{(t)}\right)$ is observed for different values of $t = 1, 2, 3, \ldots$, until convergence. The variation of γ_t and $\boldsymbol{f}\left(U_{\text{Joint}}^{(t)}\right)$ along with the corresponding value of convergence factor c is provided in the bottom rows of Fig. S3, S4, S5, and S6 for Spiral, Jain, R15, and Compound data sets, respectively. The value of the bound c is marked by a horizontal dashed green line in these figures.

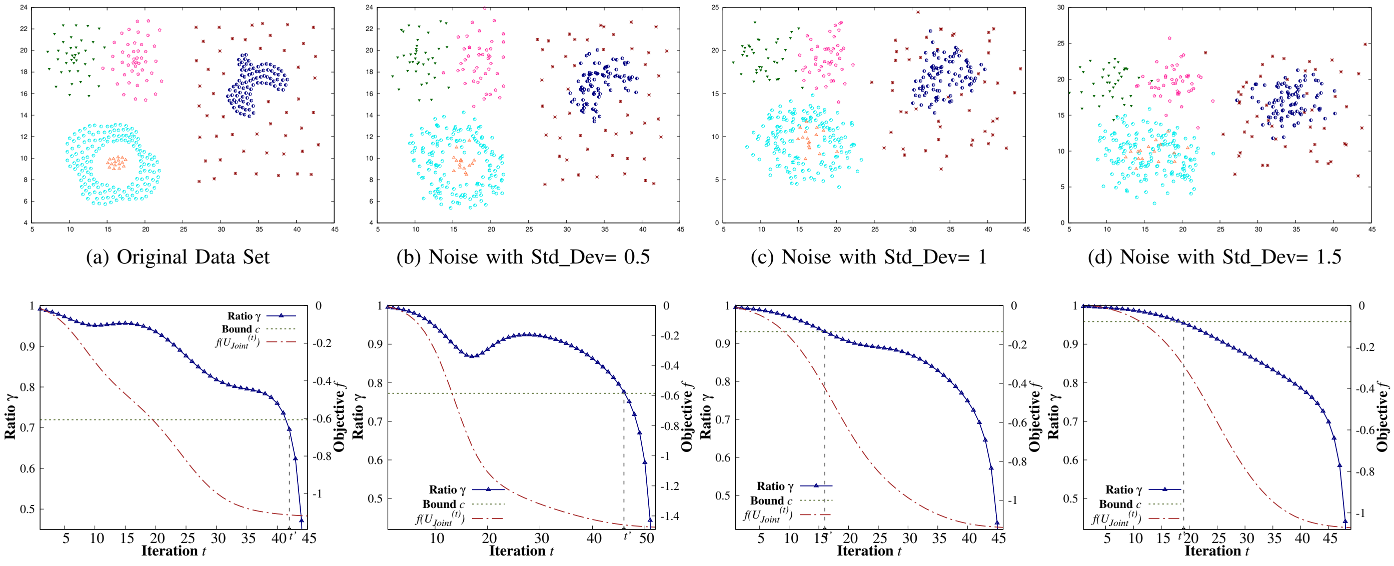
For all the data sets, the top rows of Fig. S3, S4, S5, and S6 show that the cluster structure and their separability degrades with the increase in noise, as expected. The bottom rows of these figures in turn show that with increase in noise in the data sets, the value of the convergence factor c increases and goes close to 1. For instance, for the Spiral data set, the value of c for the noise-free original data set in Fig. S3(a) is 0.5432315, while that for the three increasingly noisy variants in Figs. S4(b), S4(c), and S4(d) are 0.8995148, 0.9298075, and 0.9569399, respectively. Similar observations can be made for Jain, R15, and Compound data sets as well from the bottom rows of Figs. S4, S5, and S6, respectively. Although the results



(a) Original Data Set (b) Noise with Std_Dev= 0.5 (c) Noise with Std_Dev= 1 (d) Noise with Std_Dev= 1.5

(a) c= 0.6819649 (b) c= 0.8589394 (c) c= 0.9477831 (d) c= 0.9849866

Fig. S5. Asymptotic convergence analysis for R15 data set: scatter plot of data with varying Gaussian noise (top row) and variation of convergence ratio and objective function with increase in iteration number t (bottom row).



(a) Original Data Set (b) Noise with Std_Dev= 0.5 (c) Noise with Std_Dev= 1 (d) Noise with Std_Dev= 1.5

(a) c= 0.7194825 (b) c= 0.7725400 (c) c= 0.9311562 (d) c= 0.9583102

Fig. S6. Asymptotic convergence analysis for Compound data set: scatter plot of data with varying Gaussian noise (top row) and variation of convergence ratio and objective function with increase in iteration number t (bottom row).

are sensitive to the added noise and the choice of the random initial iterate, in general, it can be observed that lower values of c implies faster convergence. For instance, the bottom rows of Figs. S3, S4, S5, and S6 show that for all four data sets, the proposed algorithm converges in lesser number of iterations in the noise-free case compared to the noisy ones. The value of the iteration threshold t', above which the asymptotic bound is satisfied by all the iterations until convergence, is marked by a dashed vertical line in the figures. In general, it can be observed from Figs. S3, S4, S5, and S6 that for all data sets, as noise increases, the value of t' decreases implying a longer path until convergence. In brief, the results show that the convergence bound c can be used to make inference about the quality of the clusters and the speed of convergence of the proposed algorithm, for a given data set.

C. Importance of Data Integration

The proposed algorithm integrates information by optimizing a joint clustering objective while reducing the disagreement between the joint and individual subspaces. To study the importance of information integration, the performance of the proposed algorithm is compared with that of spectral clustering on the individual views. The comparative results are reported in Tables S2 and S3 for the benchmark data sets, and

TABLE S2
PERFORMANCE ANALYSIS OF INDIVIDUAL VIEWS AND MIMIC ALGORITHM ON DIGITS DATA SET

	Views→	Fac	Fou	Kar	Mor	Pix	Zer	MiMIC
Digits	Accuracy	0.5614(9.66e-4)	0.7096(7.74e-4)	0.6638(4.83e-4)	0.5109(2.10e-4)	0.6520(0.00)	0.5350(0.00)	**0.9207**(4.21e-4)
	NMI	0.6192(1.25e-3)	0.6443(3.96e-4)	0.6407(5.78e-4)	0.5361(1.80e-4)	0.6385(0.00)	0.4766(0.00)	**0.8597**(4.88e-4)
	ARI	0.4731(1.23e-3)	0.5416(9.25e-4)	0.5383(9.10e-4)	0.3723(2.84e-4)	0.5216(0.00)	0.3286(0.00)	**0.8352**(8.18e-4)
	F-measure	0.6453(9.44e-4)	0.7206(6.96e-4)	0.7023(3.96e-4)	0.5650(2.82e-4)	0.6829(0.00)	0.5544(0.00)	**0.9209**(4.15e-4)

TABLE S3
PERFORMANCE ANALYSIS OF INDIVIDUAL VIEWS AND PROPOSED MIMIC ALGORITHM FOR BENCHMARK DATA SETS

Views→		Segment1	Segment2	Segment3	Segment4	MiMIC		Shape	Texture	Margin	MiMIC
Accuracy	**BBC**	0.6202(2.1e-3)	0.6202(3.6e-2)	0.6102(3.6e-2)	0.5550(3.0e-3)	**0.8715**(0.0)	**100Leaves**	0.3095(9.1e-3)	0.4777(1.4e-2)	0.5786(1.1e-2)	**0.8185**(1.5e-2)
NMI		0.4312(1.7e-3)	0.4459(5.7e-2)	0.4097(1.3e-3)	0.4033(8.2e-3)	**0.7182**(→0)		0.6479(6.7e-3)	0.7327(5.6e-3)	0.7940(4.4e-3)	**0.9302**(4.1e-3)
ARI		0.3405(6.6e-2)	0.3895(8.9e-2)	0.3429(7.0e-3)	0.2518(1.1e-2)	**0.7273**(0.0)		0.1820(5.8e-3)	0.3265(1.4e-2)	0.4478(9.8e-3)	**0.7431**(2.5e-2)
F-measure		0.6514(1.2e-2)	0.6363(3.9e-2)	0.6435(1.2e-2)	0.6205(3.0e-3)	**0.8613**(0.0)		0.3525(7.4e-3)	0.5139(1.1e-2)	0.6113(9.7e-3)	**0.8492**(1.1e-2)
Views→		RGB	HSV	Haralick	ColorSimilarity	MiMIC		BBC	Guardian	Reuters	MiMIC
Accuracy	**ALOI**	0.4215(1.1e-2)	0.4433(7.0e-3)	0.1001(2.3e-3)	0.5191(1.1e-2)	**0.5742**(7.4e-3)	**3Sources**	0.7159(0.0)	0.6508(0.0)	0.5562(0.0)	**0.7360**(5.9e-2)
NMI		0.7179(3.9e-3)	0.7093(5.1e-3)	0.3659(4.1e-3)	0.7683(4.9e-3)	**0.7805**(2.3e-3)		0.6390(0.0)	0.5270(0.0)	0.5347(0.0)	**0.6433**(3.5e-2)
ARI		0.2915(1.4e-2)	0.2979(1.9e-2)	0.0550(6.8e-4)	0.3745(2.2e-2)	**0.4233**(6.6e-3)		**0.6082**(0.0)	0.4119(0.0)	0.41434(0.0)	0.5957(6.6e-2)
F-measure		0.4789(1.0e-2)	0.5136(7.9e-3)	0.1209(1.5e-3)	0.5843(1.1e-2)	**0.6221**(4.9e-3)		**0.7656**(0.0)	0.7036(0.0)	0.6482(0.0)	0.7581(5.0e-2)

TABLE S4
PERFORMANCE ANALYSIS OF INDIVIDUAL VIEWS AND PROPOSED MIMIC ALGORITHM FOR MULTI-OMICS DATA SETS

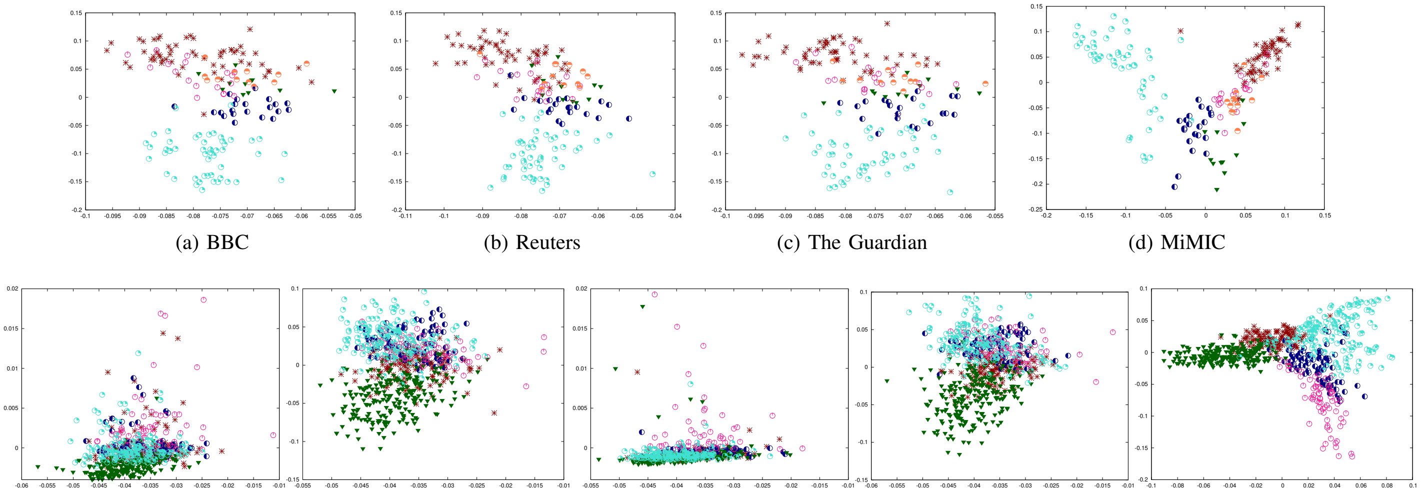
Views→		mDNA	RNA	miRNA	RPPA	MiMIC		mDNA	RNA	miRNA	RPPA	MiMIC
Accuracy	**LGG**	0.8352060	0.5917603	0.4307116	0.3970037	**0.9625468**	**STAD**	0.5413223	0.4793388	0.3719008	0.4173554	**0.7727273**
NMI		0.5734568	0.2176187	0.0498676	0.0254500	**0.8543905**		0.2282198	0.1779419	0.0771419	0.0831100	**0.5220123**
ARI		0.5567870	0.1801875	0.0510240	0.0238319	**0.8790253**		0.1927570	0.1047749	0.0514998	0.0460928	**0.4650334**
F-measure		0.8269248	0.5875701	0.4717221	0.4326018	**0.9623406**		0.5469686	0.4781377	0.3998266	0.4469459	**0.7830757**
Accuracy	**BRCA**	0.5804020	0.7688442	0.4623116	0.4798995	**0.7964824**	**LUNG**	0.8107303	0.9359165	0.8241431	0.5037258	**0.9463487**
NMI		0.3408150	0.5277072	0.1947561	0.3140984	**0.5553836**		0.2980508	0.6631276	0.3575188	0.0001449	**0.7173075**
ARI		0.3047769	0.5130244	0.1663564	0.2359641	**0.5474472**		0.3852741	0.7597207	0.4193820	-0.001743	**0.7965891**
F-measure		0.5982526	0.7690661	0.5105008	0.5630781	**0.7997020**		0.8104506	0.9357307	0.8237679	0.5630053	**0.9461134**

in Table S4 for the multi-omics cancer data sets. The results in Tables S2 and S3, clearly show that for four benchmark data sets, namely, Digits, BBC, ALOI, and 100Leaves, there is significant improvement in performance of the proposed MiMIC algorithm considering multiple views over any single view clustering. For the 3Sources data set, there is smaller improvement in terms of NMI and accuracy, and the single view BBC news source gives the best performance in terms of ARI and F-measure. In case of the multi-omics data sets, Table S4 shows that for all four data sets, the proposed algorithm achieves the best clustering performance across all four evaluation indices. The performance gain is most evident for LGG and STAD data sets. Gene or RNA expression is the most relevant view for BRCA and LUNG data sets, while for LGG and STAD data sets it is DNA-methylation. For BRCA and LUNG data sets, the clustering performance of RNA expression is very close to that of the proposed multi-view algorithm. Evidently, most of the initial works of cancer subtype identification were based on gene expression study [15], [18].

The scatter plots of the first two dimensions of the subspaces extracted by the individual views and the proposed algorithm are given in Fig. S7 for two benchmark data sets: 3Sources and BCC, and in Fig. S8 for two multi-omics data sets: LGG and STAD, as examples. The objects in these figures are colored according to the ground truth or previously established TCGA cancer subtypes. The scatter plots for the individual views in Fig. S7 and S8 demonstrate the diversity of cluster structures exhibited by the views. The scatter plots for the proposed algorithm in Figs. S7(d) (top row), S7(e) (bottom row), and S8(e) (top row) demonstrate significantly higher cluster separability compared to any of their individual views for 3Sources, BBC, and LGG data sets, respectively. The distinct omic views may exhibit disparate cluster structures, but Tables S2 - S4, and Figs. S7 and S8 indicate that proper integration gives much better idea about the overall cluster structure of the data set.

D. Results on Additional Benchmark Data Sets

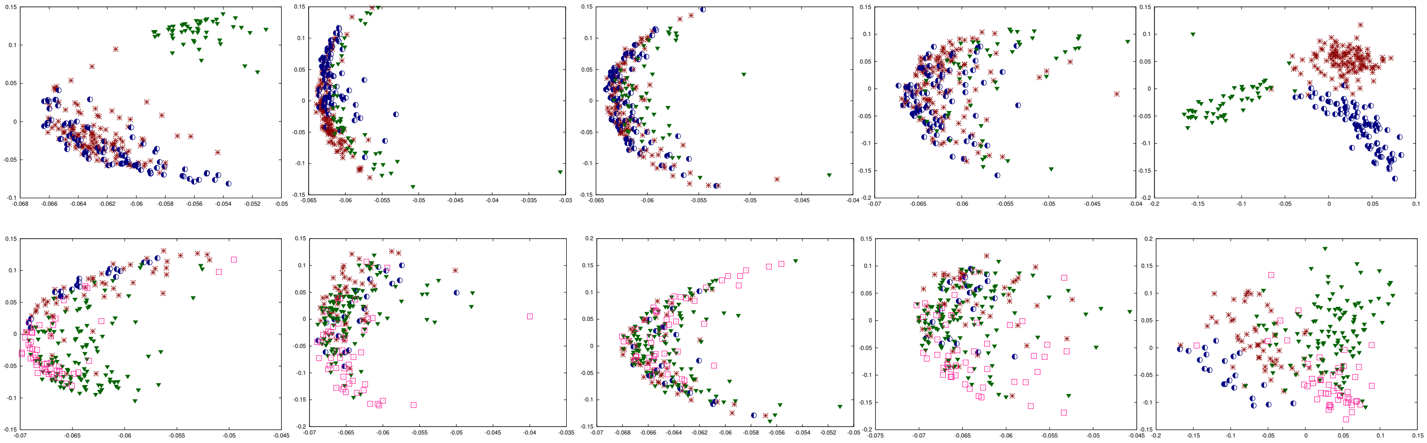
Apart from the results on various data sets reported in the main paper, experiments are also carried out on four benchmark image and social networking data sets. Among them, Football, Olympics, and Politics-IE (http://mlg.ucd.ie/aggregation/) are three benchmark multi-view Twitter data sets, each of which consists of a heterogeneous collection of nine network and content-based views [19]. Football, Olympics, and Politics-IE consists of 248, 464, and 348 samples, respectively, and their number of clusters are 20,



(a) BBC (b) Reuters (c) The Guardian (d) MiMIC

(a) Segment1 (b) Segment2 (c) Segment3 (d) Segment4 (e) MiMIC

Fig. S7. Two-dimensional scatter plots of individual views and proposed algorithm for benchmark data sets: 3Sources (top row) and BBC (bottom row).



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

Fig. S8. Two-dimensional scatter plots of individual views and proposed algorithm for multi-omics cancer data sets: LGG (top row) and STAD (bottom row).

28, and 7, respectively. The other data set Flower-17 is a 17 category flower image data set with 80 images for each class, adding up to 1360 samples. Seven dissimilarity matrices, based on chi-squared distance, are publicly available for this data set at http://www.robots.ox.ac.uk/vgg/data/flowers/17/. The performance of the proposed MiMIC algorithm on these four data sets is compared with that of their two best individual views according to clustering accuracy, the two individual manifolds, namely, k-means and Stiefel manifolds, and with the performance at rank $r = k$, the number of clusters in the data set.

The comparative results are provided in Table S5. The results in Table S5 show that the proposed algorithm significantly outperforms the best performing individual views as well as both k-means and Stiefel manifolds, for all three Twitter data sets, namely, Football, Olympics, and Politics-IE. For the Flower-17 image data set, the k-means manifold marginally outperforms the proposed algorithm, which could also be due to the randomized k-means clustering step in both the cases. The comparative performance of rank k and optimal rank $r^\star$, in case of the proposed algorithm, in Table S5 shows that rank $r^\star$ performance dominates over rank k performance for all data sets. This indicates that the optimal rank $r^\star$ preserves better cluster structure compared to rank k.

E. Choice of Damping Factor in Joint Laplacian

The joint Laplacian $\mathbf{L}^r_{\text{Joint}}$, defined in (7) of the main paper, is a convex combination of the individual approximate graph Laplacians. The convex combination is set according to Section S3. In the convex combination, the Laplacians are weighted according to the relevance of the cluster information provided by the corresponding views. The relevance measure $\boldsymbol{\chi}$ in (33) gives a linear ordering of the views based on the quality of their underlying cluster structure. Based on this ordering, the relevance values are damped by powers of Δ and then used in the convex combination. This damping strategy upweights the contribution of views with better cluster structure, while damping the effect of those having poorer structure.

TABLE S5
COMPARATIVE PERFORMANCE ANALYSIS OF PROPOSED AND EXISTING ALGORITHMS ON ADDITIONAL BENCHMARK DATA SETS

Data Set	Algorithm→	Best View	2nd Best View	Rank k	k-Means Manifold	Stiefel Manifold	MiMIC
Flower-17	Accuracy	0.3529(1.44e-2)	0.3495(3.78e-3)	0.5367(8.94e-3)	**0.5681**(1.35e-3)	0.3441(4.24e-3)	*0.5625*(1.32e-2)
	NMI	0.4015(1.02e-2)	0.3696(4.13e-3)	0.5364(6.20e-3)	**0.5840**(4.44e-3)	0.3915(2.60e-3)	*0.5821*(5.59e-3)
	ARI	0.1905(1.30e-2)	0.1789(2.97e-3)	0.3459(7.38e-3)	**0.4110**(7.62e-3)	0.1805(4.27e-3)	*0.4002*(6.50e-3)
	F-measure	0.3982(1.77e-2)	0.3968(2.94e-3)	0.5532(8.04e-3)	**0.6056**(8.52e-3)	0.3749(3.50e-3)	*0.6003*(8.74e-3)
Football	Accuracy	0.7419(1.86e-2)	0.6737(1.25e-2)	0.7915(4.76e-2)	0.8673(1.14e-2)	0.7366(1.09e-2)	**0.8846**(2.27e-2)
	NMI	0.7910(1.03e-2)	0.7432(1.05e-2)	0.8408(3.41e-2)	0.8804(9.42e-3)	0.7742(6.64e-3)	**0.8958**(1.24e-2)
	ARI	0.5814(3.46e-2)	0.4531(4.89e-2)	0.6699(8.24e-2)	0.7566(2.51e-2)	0.5584(1.63e-2)	**0.7841**(4.61e-2)
	F-measure	0.7757(1.49e-2)	0.6887(1.21e-2)	0.8159(3.97e-2)	0.8792(9.90e-2)	0.7610(9.99e-3)	**0.8941**(1.74e-2)
Olympics	Accuracy	0.8187(1.62e-2)	0.7793(8.92e-3)	0.8625(2.14e-2)	0.8228(2.88e-2)	0.7390(2.04e-2)	**0.8844**(2.60e-2)
	NMI	0.8763(5.10e-3)	0.8249(8.50e-3)	0.9223(8.63e-3)	0.9141(1.08e-2)	0.8075(1.05e-2)	**0.9394**(9.10e-3)
	ARI	0.7504(3.39e-2)	0.6591(2.59e-2)	0.8267(3.27e-2)	0.7890(6.58e-2)	0.5474(2.64e-2)	**0.8699**(3.52e-2)
	F-measure	0.8367(1.67e-2)	0.8037(9.98e-3)	0.8813(1.62e-2)	0.8520(2.87e-2)	0.7699(1.82e-2)	**0.9006**(2.35e-2)
Politics-IE	Accuracy	0.9080(0.0)	0.8508(4.16e-3)	0.8810(1.27e-2)	0.8764(0.0)	0.8048(3.30e-2)	**0.9436**(1.45e-2)
	NMI	0.8537(0.0)	0.7337(1.59e-3)	0.8137(1.46e-2)	0.8246(→0)	0.6884(2.63e-2)	**0.8573**(1.88e-2)
	ARI	**0.9162**(0.0)	0.6703(6.08e-3)	0.7978(2.80e-2)	0.7408(0.0)	0.7096(3.17e-2)	0.8693(2.81e-2)
	F-measure	0.9079(0.0)	0.8293(4.77e-3)	0.8651(1.50e-2)	0.8662(0.0)	0.7988(3.18e-2)	**0.9447**(1.21e-2)

TABLE S6
PERFORMANCE OF THE MIMIC ALGORITHM FOR DIFFERENT VALUES OF DAMPING FACTOR Δ ON BENCHMARK AND MULTI-OMICS DATA SETS

	Measure		$\Delta = 1$	$\Delta = 2$		$\Delta = 1$	$\Delta = 2$		$\Delta = 1$	$\Delta = 2$		$\Delta = 1$	$\Delta = 2$
Benchmark	Rank	**Digits**	12	42	**3Sources**	21	26	**BBC**	21	5	**100Leaves**	180	50
	Accuracy		**0.9207**(4.21e-4)	0.7860(0.0)		**0.7360**(5.92e-2)	0.6520(3.74e-3)		**0.8715**(0.0)	0.7976(3.04e-2)		**0.8185**(1.55e-2)	0.6765(1.80e-2)
	NMI		**0.8597**(4.88e-4)	0.8275(→0)		**0.6433**(3.59e-2)	0.6224(8.33e-3)		**0.7182**(→0)	0.6658(4.01e-2)		**0.9302**(4.12e-3)	0.8499(6.61e-3)
	ARI		**0.8352**(8.18e-4)	0.7367(0.0)		**0.5957**(6.69e-2)	0.5225(1.34e-2)		**0.7273**(0.0)	0.7027(6.04e-2)		**0.7431**(2.53e-2)	0.5715(2.10e-2)
	F-measure		**0.9209**(4.15e-4)	0.8428(0.0)		**0.7581**(5.04e-2)	0.6941(3.91e-3)		**0.8613**(0.0)	0.8127(3.42e-2)		**0.8492**(1.13e-2)	0.7067(1.46e-2)
Multi-Omics	Rank	**BRCA**	40	4	**LGG**	45	43	**STAD**	25	16	**LUNG**	4	3
	Accuracy		0.6683(0.0)	**0.7964**(0.0)		0.9700(0.0)	**0.9625**(0.0)		0.7727(0.0)	**0.7727**(0.0)		0.9388(0.0)	**0.9463**(0.0)
	NMI		0.4503(→0)	**0.5553**(→0)		0.8646(→0)	**0.8543**(→0)		0.5183(→0)	**0.5220**(→0)		0.6920(0.0)	**0.7173**(0.0)
	ARI		0.3894(0.0)	**0.5474**(0.0)		0.9097(0.0)	**0.8790**(0.0)		0.4658(0.0)	**0.4650**(0.0)		0.7701(0.0)	**0.7965**(0.0)
	F-measure		0.6800(0.0)	**0.7997**(0.0)		0.9700(0.0)	**0.9623**(0.0)		0.7791(0.0)	**0.7830**(0.0)		0.9385(0.0)	**0.9461**(0.0)

With damping factor $\Delta = 1$, the individual contributions are relatively close to each other depending upon their Fiedler values and Fiedler vectors. On the other hand, with $\Delta = 2$, the contributions of the views in decreasing order of relevance are $\frac{\mathcal{X}_{(1)}}{2}$, $\frac{\mathcal{X}_{(2)}}{4}$, $\frac{\mathcal{X}_{(3)}}{8}$, and so on. This indicates heavier damping and higher difference between the individual contributions. The effect of the two damping factors is studied in Table S6 for different data sets.

Table S6 shows that for four benchmark data sets, namely, Digits, 3Sources, BBC, and 100Leaves, lower damping ($\Delta = 1$) gives better performance compared to higher damping ($\Delta = 2$). The individual views of the benchmark data sets are relatively similar to each other, for instance, different segments of the same news article for the BBC data set, and RGB and HSV colour histograms of same image for ALOI data set. As a result, lower damping works better for the benchmark data sets. For the multi-oimcs data sets, however, Table S6 shows that heavier damping with $\Delta = 2$ gives better performance. Table S4 shows that there is a significant difference between the clustering performance of the most and the second most relevant views of LGG, BRCA, and LUNG data sets. Hence, significantly upweighting the most relevant view with $\Delta = 2$ gives better performance for the multi-oimcs data sets. Therefore, in this work, the damping factor Δ is chosen to be 2 for the multi-omics data sets, and 1 for the benchmark data sets.

S7. CLUSTER EVALUATION MEASURES

Four external cluster evaluation measures are used to compare the performance different approaches, namely, accuracy, adjusted rand index (ARI), normalized mutual information (NMI), and F-measure. Since there are different definitions of some of the measures, like accuracy and NMI, in clustering, the definitions used in this work is are described next. A higher value indicates a better performance for each metric. Let $\mathcal{T} = \{t_1, \ldots, t_j, \ldots, t_k\}$ be the true partition of n samples of a data set into k clusters. Let $\mathcal{C} = \{c_1, \ldots, c_i, \ldots, c_k\}$ be the k clusters returned by a clustering algorithm. Let the number of samples in the data set be denoted by n. The external evaluation indices measure how close is the clustering $\mathcal{C}$ with respect to true partition $\mathcal{T}$. The four external evaluation indices are as follows.

1) **Accuracy** [20]: Given a sample x_p , let its cluster and class labels be denoted by c_p and t_p, respectively. The clustering accuracy is given by

$$\text{Accuracy} = \frac{1}{n}\sum_{p=1}^{n} \delta(t_p, map(c_p)),$$

where $\delta(a,b) = 1$ when $a = b$, otherwise $\delta(a,b) = 0$. The function $map(c_p)$ is the permutation map function, which maps the cluster labels into class labels. The best map can be obtained by the Kuhn-Munkres algorithm [21].

2) **NMI** [22] measures the concordance of cluster assignments in $\mathcal{T}$ and $\mathcal{C}$. NMI is defined as follows:

$$\text{NMI}(\mathcal{T},\mathcal{C}) = \frac{2\,\mathbb{I}\,(\mathcal{T},\mathcal{C})}{[\mathbb{H}(\mathcal{T}) + \mathbb{H}(\mathcal{C})]}; \tag{36}$$

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

$$\mathbb{H}\,(\mathcal{C}) = -\sum_{i=1}^{k} Pr(c_i)\log Pr(c_i);$$

$$\mathbb{I}\,(\mathcal{T},\mathcal{C}) = \sum_{i=1}^{k}\sum_{j=1}^{k} Pr(c_i \cap t_j)\log\left[\frac{Pr(c_i \cap t_j)}{Pr(c_i)Pr(t_j)}\right];$$

where $Pr(S)$ denotes the probability of the set S.

3) **ARI** [23] is an adjustment of the rand index, given by,

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

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

4) **F-measure** [24] of a cluster c_i with respect to a class t_j evaluates how well cluster cluster c_i describes class t_j and is given by the harmonic mean of precision and recall.

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

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

$$\begin{aligned}\text{F-measure }\ \mathcal{F}(t_j,c_i) &= \frac{2P_{ij}R_{ij}}{P_{ij}+R_{ij}}\\ &= \frac{2|c_i \cap t_j|}{|c_i| + |t_j|}.\end{aligned}$$

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

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

where n_j denotes the number of points in class t_j.

References

[1] T. Carson, D. G. Mixon, and S. Villar, "Manifold optimization for k-means clustering," in *2017 International Conference on Sampling Theory and Applications (SampTA)*, July 2017, pp. 73–77.

[2] C. Moler and C. Loan, "Nineteen dubious ways to compute the exponential of a matrix," *SIAMREV*, vol. 20, pp. 801–836, 10 1978.

[3] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton University Press, 2008.

[4] L. Armijo, "Minimization of functions having lipschitz continuous first partial derivatives." *Pacific J. Math.*, vol. 16, no. 1, pp. 1–3, 1966.

[5] N. Boumal, "Optimization and estimation on manifolds," Ph.D. dissertation, Université catholique de Louvain, 2014.

[6] D. A. Spielman and S.-H. Teng, "Spectral partitioning works: Planar graphs and finite element meshes," *Linear Algebra and its Applications*, vol. 421, no. 2, pp. 284 – 305, 2007.

[7] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[8] A. Khan and P. Maji, "Approximate graph laplacians for multimodal data clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, doi: 10.1109/TPAMI.2019.2945574.

[9] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, ser. COLT' 98. New York, NY, USA: Association for Computing Machinery, 1998, p. 92100. [Online]. Available: https://doi.org/10.1145/279943.279962

[10] 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.

[11] 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, no. 5, pp. E359–386, Mar 2015.

[12] TCGA Research Network, "Comprehensive molecular characterization of gastric adenocarcinoma," *Nature*, vol. 513, no. 7517, pp. 202–209, 2014.

[13] TCGA Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct 2012.

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

[15] T. Sørlie *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, no. 19, pp. 10 869–10 874, Sep 2001.

[16] W. D. Travis, E. Brambilla, A. P. Burke, A. Marx, and A. G. Nicholson, "Introduction to The 2015 World Health Organization Classification of Tumors of the Lung, Pleura, Thymus, and Heart," *J Thorac Oncol*, vol. 10, no. 9, pp. 1240–1242, Sep 2015.

[17] H. Wang, Y. Yang, and B. Liu, "Gmc: Graph-based multi-view clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 6, pp. 1116–1129, 2020.

[18] M. E. Garber *et al.*, "Diversity of gene expression in adenocarcinoma of the lung," *Proceedings of the National Academy of Sciences*, vol. 98, no. 24, pp. 13 784–13 789, 2001. [Online]. Available: https://www.pnas.org/content/98/24/13784

[19] D. Greene and P. Cunningham, "Producing a unified graph representation from multiple social network views," in *5th Annual ACM Web Science Conference*, 2013, pp. 118–121.

[20] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, and D. Xu, "Generalized latent multi-view subspace clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 86–99, 2020.

[21] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 12, pp. 83–97, 1955. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109

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

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

[24] 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, 1999, pp. 16–22.