Supplementary materials of "Integrating multidimensional data for clustering analysis with applications to cancer patient data"

Seyoung Park[a], Hao Xu[b], and Hongyu Zhao[b] *

[a]Department of Statistics, Sungkyunkwan University, Seoul, Korea

[b]Department of Biostatistics, Yale School of Public Health, New Haven, CT

February 5, 2020

*Contact: Hongyu Zhao, hongyu.zhao@yale.edu, Department of Biostatistics, Yale School of Public Health, New Haven, CT.

These supplementary materials include the detailed algorithm, theoretical details of the proposed method, and additional figures. In Section A, we develop the proposed algorithm. In Sections B and C, we include theoretical details of the statistical consistency of the proposed clustering method and convergence of the algorithm, respectively. In Section D, additional lemmas with proofs are included. In Sections E and F, we include brief discussion of time complexity of the algorithm and evaluation metrics used in the main paper, respectively. Detailed data processing steps and additional figures are provided in Sections G and H, respectively. Application to stomach cancer is illustrated in Section I.

A Algorithm

Let $F(\{P_m\}, \{c_m\}, \{w_{ml}\})$ be the objective function of (4) in the main paper. Although $F(\cdot)$ is not a jointly convex function, it is convex for one parameter when the other variables are fixed. Hence we iteratively solve (4) as follows: at the $(i+1)$th iterate of each update,

$$\begin{aligned} \{c_m\}_{i+1} &= \operatorname*{argmin}_{\{c_m\}} F(\{P_m\}_i, \{c_m\}, \{w_{ml}\}_i) && \text{(S1)} \\ \{w_{ml}\}_{i+1} &= \operatorname*{argmin}_{\{w_{ml}\}} F(\{P_m\}_i, \{c_m\}_{i+1}, \{w_{ml}\}) && \text{(S2)} \\ \{P_m\}_{i+1} &= \operatorname*{argmin}_{\{P_m\}} F(\{P_m\}, \{c_m\}_{i+1}, \{w_{ml}\}_{i+1}) && \text{(S3)} \end{aligned}$$

until convergence. Note that (S1)-(S3) are convex optimizations.

A.1 Update $\{c_m\}_{i+1}$

Step 1: Update $\{c_m\}_{i+1}$

For $m = 1, \cdots, M$, let

$$\widetilde{c}_m^{(i+1)} = \exp\left(\frac{1}{\rho}\left(-\epsilon\|P_m^{(i)}\|_F^2 + \langle S_m^{(i)}, P_m^{(i)}\rangle - \lambda\|P_m^{(i)}\|_1\right)\right).$$

Then, we update $\{c_m\}_{i+1}$ by

$$c_m^{(i+1)} = \frac{M\widetilde{c}_m^{(i+1)}}{\sum_m \widetilde{c}_m^{(i+1)}} \quad \text{for all} \quad m = 1, \cdots, M, \tag{S4}$$

which is the minimizer of (S1).

A.2 Update $\{w_{ml}\}_{i+1}$

Step 2: Update $\{w_{ml}\}_{i+1}$

For fixed $m \in \{1, \cdots, M\}$ and $l \in \{1, \cdots, \ell\}$,

$$w_{ml}^{(i+1)} = \frac{\exp\left(\frac{c_m^{(i+1)}}{\rho}\langle G_{ml}, P_m^{(i)}\rangle\right)}{\sum_{\widetilde{l}} \exp\left(\frac{c_m^{(i+1)}}{\rho}\langle G_{m\widetilde{l}}, P_m^{(i)}\rangle\right)},$$

which is the minimizer of (S2).

A.3 Update $\{P_m\}_{i+1}$ via ADMM

Step 3: Update $\{P_m\}_{i+1}$ via ADMM

Note that solving (S3) is equivalent to solving

$$\begin{aligned}&\min_{\{P_m\}} \ \epsilon\sum_m c_m^{(i+1)}\|P_m\|_F^2 - \sum_m c_m^{(i+1)}\langle S_m^{(i+1)}, P_m\rangle + \lambda\sum_m c_m^{(i+1)}\|P_m\|_1 + \mu\sum_{m\neq j}\|P_m - P_j\|^2\\ &\text{s.t.} \quad \mathrm{tr}(P_m) = C, \ 0 \preceq P_m \preceq I.\end{aligned} \tag{S5}$$

We solve (S5) using the general idea of ADMM. Consider the following equivalent formulation:

$$\min_{\{P_m\}} \ \epsilon \sum_m c_m^{(i+1)} \|P_m\|_F^2 - \sum_m c_m^{(i+1)} \langle S_m^{(i+1)}, P_m \rangle + \lambda \sum_m c_m^{(i+1)} \|P_m\|_1 + \mu \sum_{m \neq j} \|P_m - Q_j\|^2$$
$$\text{s.t.} \quad Q_m = P_m, \ \operatorname{tr}(Q_m) = C, \ 0 \preceq Q_m \preceq I.$$

By the augmented Lagrangian method, we solve

$$\begin{aligned}
\min_{\{P_m\},\{Q_m\},\{\Gamma_m\}} \quad & \epsilon \sum_m c_m^{(i+1)} \|P_m\|_F^2 - \sum_m c_m^{(i+1)} \langle S_m^{(i+1)}, P_m \rangle + \lambda \sum_m c_m^{(i+1)} \|P_m\|_1 \\
+ \quad & \mu \sum_{m \neq j} \|P_m - Q_j\|^2 + \sum_m \langle \Gamma_m, P_m - Q_m \rangle + \frac{\eta}{2} \sum_m \|P_m - Q_m\|^2 \\
\text{s.t.} \quad & \operatorname{tr}(Q_m) = C, \ 0 \preceq Q_m \preceq I,
\end{aligned} \tag{S6}$$

where the dual variables Γ_m's are the Lagrangian multipliers and $\eta > 0$ is the penalty parameter. We iteratively update P_m, Q_m, and Γ_m. Since the optimization (S6) is convex, this ADMM guarantees convergences of the iterates.

Let $H(\{P_m\}, \{Q_m\}, \{\Gamma_m\} \mid \{c_m\}_{i+1}, \{w_{ml}\}_{i+1})$ be the objective function in (S6). At the $(t+1)$th iterate of the ADMM, we solve for the minimizer iteratively using the following steps until it converges:

A.3.1 Update $\{P_m\}^{t+1}$

Update $\{P_m\}^{t+1} = \{P_1^{t+1}, \cdots, P_M^{t+1}\}$ by

$$\{P_m\}^{t+1} = \operatorname*{argmin}_{\{P_m\}} \ H(\{P_m\}, \{Q_m^t\}, \{\Gamma_m^t\} \mid \{c_m\}_{i+1}, \{w_{ml}\}_{i+1}).$$

This is equivalent to solving for each $m = 1, \cdots, M$,

$$\begin{aligned} P_m^{t+1} &= \underset{P}{\operatorname{argmin}}\ \epsilon c_m^{(i+1)}\|P\|_F^2 - c_m^{(i+1)}\langle S_m^{(i+1)}, P\rangle + \lambda c_m^{(i+1)}\|P\|_1 + \mu \sum_{\tilde{m}\neq m}\|P - Q_{\tilde{m}}^t\|^2 \\ &\qquad + \langle \Gamma_m^t, P - Q_m^t\rangle + \frac{\eta}{2}\|P - Q_m^t\|^2. \end{aligned}$$

Then by the KKT condition, we have for $k, j \in \{1, \cdots, n\}$,

$$\begin{aligned} 2\epsilon c_m^{(i+1)} P_{kj} - c_m^{(i+1)} S_{m,kj}^{(i+1)} + \lambda c_m^{(i+1)}\operatorname{sign}(P_{kj}) + 2\mu \sum_{\tilde{m}\neq m}(P_{kj} - Q_{\tilde{m},kj}^t) \\ + \Gamma_{m,kj}^t + \eta(P_{kj} - Q_{m,kj}^t) = 0. \end{aligned}$$

Let $t_{m,kj} = c_m^{(i+1)} S_{m,kj}^{(i+1)} - \Gamma_{m,kj}^t + \eta Q_{m,kj}^t + 2\mu \sum_{\tilde{m}\neq m} Q_{\tilde{m},kj}^t$. Then,

$$\begin{aligned} (P_m^{t+1})_{kj} &= (t_{m,kj} - \lambda c_m^{(i+1)})(2\epsilon c_m^{(i+1)} + \eta + 2\mu(M-1))^{-1} \quad \text{if} \quad t_{m,kj} > \lambda c_m^{(i+1)} \\ (P_m^{t+1})_{kj} &= (t_{m,kj} + \lambda c_m^{(i+1)})(2\epsilon c_m^{(i+1)} + \eta + 2\mu(M-1))^{-1} \quad \text{if} \quad t_{m,kj} < -\lambda c_m^{(i+1)} \\ (P_m^{t+1})_{kj} &= 0 \quad \text{if} \quad |t_{m,kj}| \leq \lambda c_m^{(i+1)}. \end{aligned}$$

A.3.2 Update $\{Q_m\}^{t+1}$

Update $\{Q_m\}^{t+1} = \{Q_1^{t+1}, \cdots, Q_M^{t+1}\}$ by solving for each $m = 1, \cdots, M$,

$$\begin{aligned} Q_m^{t+1} &= \underset{Q}{\operatorname{argmin}}\ \mu \sum_{\tilde{m}\neq m}\|P_{\tilde{m}}^{t+1} - Q\|^2 + \langle -\Gamma_{\tilde{m}}^t, Q\rangle + \frac{\eta}{2}\|P_{\tilde{m}}^{t+1} - Q\|^2 \\ &\quad \text{s.t.} \quad \operatorname{tr}(Q) = C,\ 0 \preceq Q \preceq I, \end{aligned}$$

which can be rewritten as

$$\min_Q \|Q - T_m\|_F^2 \quad \text{s.t.} \ \operatorname{tr}(Q) = C,\ 0 \preceq Q \preceq I,$$

where $T_m = (2\mu\sum_{\tilde{m}\neq m} P_{\tilde{m}}^{t+1} + \Gamma_m^t + \eta P_m^{t+1})(2\mu(M-1) + \eta)^{-1}$. Let $B_m = (T_m + T_m^T)/2$ and $B_m = U\operatorname{diag}(u)U^T$ is the spectral decomposition of the symmetric matrix B_m. By using Theorem 2 in Lu *et al.* (2016), we obtain $Q_m^{t+1} = U\,\mathrm{diag}(\lambda^*)U^T$, where λ^* is the solution to

$$\min_{\lambda} \|\lambda - u\|^2 \quad \text{s.t.} \quad 0 \le \lambda \le 1, \quad 1^T\lambda = C,$$

which can be efficiently solved as in Wang and Lu (2015).

A.3.3 Update $\{\Gamma_m\}^{t+1}$

Update $\{\Gamma_m\}^{t+1} = \{\Gamma_1^{t+1}, \cdots, \Gamma_M^{t+1}\}$ by $\Gamma_m^{t+1} = \Gamma_m^t + \eta(P_m^{t+1} - Q_m^{t+1})$.

We repeat this procedure until convergence of sequences $\{P_m\}^t$, $\{Q_m\}^t$, and $\{\Gamma_m\}^t$. The obtained $\{P_m\}^t$ at the last iterate of the above ADMM is updated as $\{P_m\}_{i+1}$ in (S3).

B Proof of statistical consistency

In this section, we prove statistical consistency of the proposed clustering method as in (4) in the main paper. We first state Theorem S1 with its proof.

Theorem S1. *Suppose the n data points $x_1, x_2, \cdots, x_n \in \mathbb{R}^p$ follow the following sub-Gaussian distribution: $x_i = \mu^{(f(i))} + z_i$, where $z_i := (z_{i1}, \cdots, z_{ip})^T \in \mathbb{R}^p$ is a random vector with independent component satisfying $E[z_{ij}] = 0$, $E[z_{ij}^2] = \sigma_z^2$, and $\|z_{ij}\|_{\psi_2} \le \sigma_z\psi$ for some positive constants ψ and σ_z. Here, $\mu^{(1)}, \cdots, \mu^{(C^*)}$ represent the underlying center of each of the C^* clusters. Suppose $\min_{k \ne \tilde{k}} \|\mu^{(\tilde{k})} - \mu^{(k)}\|_2^2 \ge 8p\sigma_z^2 + 64\sigma_z^2\psi^2 \log n/c$ for some constant $c > 0$. Assume $c_1 n \le |T_k| \le (1 - c_1)n$ for some constant $c_1 \in (0, 1/2)$. Let $\hat{P}$ be the* *solution to (2) with $\lambda = C^*/n^2$, $\epsilon = 1/n^3$, $W = D^{-1/2}SD^{-1/2}$, and $C = C^*$, where the similarity matrix $S = (s_{i,j})$ is constructed as $s_{i,j} = K_{\sigma,g}(x_i, x_j)$, where*

$$K_{\sigma,g}(x_i, x_j) = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\epsilon_{ij}^2}\right) & \textit{if } i \in N_g(j) \textit{ or } j \in N_g(i) \\ 0 & \textit{otherwise,} \end{cases} \tag{S7}$$

where

$$\epsilon_{ij} = \frac{\sigma(\mu_i + \mu_j)}{2}, \quad \mu_i = \frac{\sum_{j \in N_g(i)} \|x_i - x_j\|}{g}$$

for some positive number σ and $g < n$. Let $\hat{L}$ be the n by C^* matrix consisting of C^* eigenvectors of $\hat{P}$ corresponding to the first C^* largest eigenvalues. Then, k-means clustering to the normalized row vectors of $\hat{L}$ guarantees the exact clustering results with probability $1 - 2(C^*)^2/n$.

***Proof of Theorem S1**.* We first state the following curvature lemma (Lemma S1), where the details and the proof can be found in Vu *et al.* (2013):

Lemma S1. *Let A be a symmetric matrix and E be the projection onto the subspace spanned by the eigenvectors of A corresponding to its d largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. If $\delta_A = \lambda_d - \lambda_{d+1} > 0$, then*

$$\frac{\delta_A}{2}\|E - F\|_F^2 \leq \langle A, E - F\rangle$$

for all F satisfying $0 \preceq F \preceq I$ and $\mathrm{tr}(F) = d$.

Note that S is the similarity matrix as in Theorem S1. Let S^* be the similarity matrix by adding $1/n^2$, on some entries of S such that points x_i and x_j in the same cluster have other samples $x_{k_1}, \cdots, x_{k_q}$ in the same cluster such that $S_{i,k_1}, S_{k_1,k_2}, \cdots, S_{k_{q-1},k_q}, S_{k_q,j}$ have non-zero values, where

$q, k_1, \cdots, k_q$ depend on i and j. There are many ways to construct such S^*, but we consider the case that S^* is constructed by adding the least number of $1/n^2$ on S. We can see that this least number is less than $2n$. Let $D^* = \mathrm{diag}(d_1^*, \cdots, d_n^*)$ be the corresponding degree matrix. Let $W^* = (D^*)^{-1/2} S^* (D^*)^{-1/2}$ be the graph Laplacian and L^* be the n by C^* matrix consisting of C^* eigenvectors corresponding to the C^* largest eigenvalues of W^*. By Lemma S6, with probability at least $1 - 2(C^*)^2/n$, we have $W_{i,j}^* = 0$ if $f(i) \neq f(j)$.

Furthermore, we can see that

$$L_{i,l}^* = \begin{cases} \sqrt{d_i^*} / \sqrt{\sum_{k:\ f(k)=f(i)} d_k^*} & \text{if} \quad l = f(i) \\ 0 & \text{if} \quad l \neq f(i). \end{cases}$$

Let $P^* = L^*(L^*)^T$ be the underlying projection matrix. By Lemma S5, we have with probability $1 - 2n^{-1}$, $g \exp(-10\sigma^{-2}) \leq d_i^* \leq 2g \exp(-0.1\sigma^{-2})$ for all i. Hence, we have

$$\sqrt{\frac{\exp(-9.9\sigma^{-2})}{2(1-c_1)n}} \leq \sqrt{\frac{\exp(-9.9\sigma^{-2})}{2n_{f(i)}}} \leq L_{i,f(i)}^* \leq \sqrt{\frac{2\exp(9.9\sigma^{-2})}{n_{f(i)}}} \leq \sqrt{\frac{2\exp(9.9\sigma^{-2})}{c_1 n}}. \tag{S8}$$

Since $\hat{P}$ is the solution to (2) in the main paper, we have

$$\begin{aligned} 0 &\leq \epsilon \left(\|P^*\|_F^2 - \|\hat{P}\|_F^2 \right) + \langle W, \hat{P} - P^* \rangle + \lambda \left(\|P^*\|_1 - \|\hat{P}\|_1 \right) \\ &\leq \epsilon C^* \|P^* - \hat{P}\|_F + \langle W - W^*, \hat{P} - P^* \rangle - \langle W^*, P^* - \hat{P} \rangle + \lambda \left(\|P^*\|_1 - \|\hat{P}\|_1 \right). \end{aligned}$$

Let $T = \mathrm{support}(W) \cup \mathrm{support}(W^*)$. We have $|T| \leq 2n + 4ng \leq 6ng$. Let $\tilde{T} = \mathrm{support}(P^*)$. Since P^* is the projection matrix onto the subspace spanned by the eigenvectors of W^* corresponding to its C^* largest eigenvalues and $\lambda_{C^*}(W^*) - \lambda_{C^*+1}(W^*) = 1 - 0 = 1$, it holds that by Lemma S1,

$$\begin{aligned}\frac{1}{2}\|\hat{P} - P^*\|_F^2 &\leq \epsilon C^*\|\hat{P} - P^*\|_F + \langle W - W^*, \hat{P} - P^*\rangle + \lambda\left(\|P^*\|_1 - \|\hat{P}\|_1\right) \\ &\leq \epsilon C^*\|\hat{P} - P^*\|_F + \|(W - W^*)_T\|_{\max}\|(\hat{P} - P^*)_T\|_1 + \lambda\left(\|P^*\|_1 - \|\hat{P}\|_1\right) \\ &\leq \epsilon C^*\|\hat{P} - P^*\|_F + \lambda\left(\|\Delta_T\|_1 + \|P^*_{\tilde{T}}\|_1 - \|P^*_{\tilde{T}} + \Delta_{\tilde{T}}\|_1\right),\end{aligned}$$

where $\Delta = \hat{P} - P^*$ and the second inequality follows from $\|W - W^*\|_{\max} \leq C^*/n^2 = \lambda$. Then, we have

$$\frac{1}{2}\|\Delta\|_F^2 \leq \lambda(\|\Delta_T\|_1 + \|\Delta_{\tilde{T}}\|_1) + C^*\epsilon\|\Delta\|_F \leq 2\lambda\sqrt{ng}\|\Delta\|_F + C^*\epsilon\|\Delta\|_F.$$

Solving the above inequalities with $\epsilon = 1/n^3$, we have

$$\|\Delta\|_F \leq 4\lambda\sqrt{ng} + 2C^*\epsilon \leq 5\sqrt{g}C^*n^{-3/2}.$$

Thus, we have $\|\hat{P} - P^*\|_F \lesssim n^{-3/2}$. By using the sin$\Theta$ theorem (Davis and Kahan, 1970), we have

$$\|\hat{L}\hat{L}^T - \tilde{L}^*(\tilde{L}^*)^T\|_F \lesssim n^{-3/2}. \tag{S9}$$

Note that for i and j from the same cluster (i.e., $f(i) = f(j)$), we have

$$\begin{aligned}\frac{\langle \hat{L}_i, \hat{L}_j\rangle}{\|\hat{L}_i\|_2\|\hat{L}_j\|_2} &\gtrsim \frac{\langle L_i^*, L_j^*\rangle - n^{-3/2}}{\sqrt{\|L_i^*\|^2 + n^{-3/2}}\sqrt{\|L_j^*\|^2 + n^{-3/2}}} \\ &= 1 - \frac{n^{-1.5}(L^*_{i,f(i)} + L^*_{j,f(i)})^2}{t(t + L^*_{i,f(i)}L^*_{j,f(i)} - n^{-1.5})} \\ &\gtrsim 1 - n^{-1/2},\end{aligned}$$

where $t := \sqrt{(L^*_{i,f(i)})^2 + n^{-1.5}}\sqrt{(L^*_{j,f(i)})^2 + n^{-1.5}}$, and the first and the second inequality follows from (S9) and (S8). Thus, we have

$$\left\|\hat{L}_i/\|\hat{L}_i\|_2 - \hat{L}_j/\|\hat{L}_j\|_2\right\|_2 \lesssim n^{-1/4}.$$

For samples i and j from different clusters, we have

$$\begin{aligned}\frac{\langle \hat{L}_i, \hat{L}_j \rangle}{\|\hat{L}_i\|_2 \|\hat{L}_j\|_2} &\leq \frac{\langle L_i^*, L_j^* \rangle + n^{-3/2}}{\sqrt{\|L_i^*\|^2 - n^{-3/2}}\sqrt{\|L_j^*\|^2 - n^{-3/2}}} \\ &\leq n^{-1/2},\end{aligned}$$

where the first inequality follows from the fact that $\langle L_i^*, L_j^* \rangle = 0$ for the i and j from different clusters. Thus, we have

$$\left\| \hat{L}_i/\|\hat{L}_i\|_2 - \hat{L}_j/\|\hat{L}_j\|_2 \right\|_2 \gtrsim \sqrt{2 - 2n^{-1/2}}.$$

By Lemma S4, the k-means clustering algorithm with normalized $\hat{L}$ guarantees the exact clustering results. This completes the proof.

□

We now move to the proof of Theorem 1 of the main paper. To better understand the proof, we first present an overview of the proof with main ideas, and then move to details of the proof.

Overview of the proof of Theorem 1. First, for each similarity matrix S_{ml}, we consider a new similarity matrix S_{ml}^* by adding a small value to a few entries in S_{ml} such that samples in the same underlying cluster become connected in each graph of the new similarity matrix S_{ml}^* (see the proof of Theorem 1 for details). Here, S_{ml} represents the similarity matrix constructed by the Gaussian kernel $K_{\sigma,g}$, where σ is the mth component of $\{1, 1.25, \cdots, 2\}$ and g is the lth component of $\{10, 12, \cdots, 30\}$, respectively.

Second, for each data set (e.g. mth data set), we consider the weighted normalized similarity matrix S_m^* generated from the S_{ml}^* weighted by the obtained $\hat{w}_{ml}$ for $l = 1, \cdots, \ell$. Let P_m^* be the projection matrix onto the subspace spanned by the eigenvectors of S_m^* corresponding to its C^* largest eigenvalues. We show that P_m^* enjoys nice theoretical properties that actually include a true clustering information.

Third, we compare two matrices $\hat{P}_m$ and P_m^*, where $\hat{P}_m$ is the obtained target matrix from the proposed optimization. We prove that $\hat{P}_m$ and P_m^* are close, which implies that clustering information contained in these two projection matrices are close through the proposed method. Finally, by using the fact that P_m^* provides the true clustering information, we prove that the proposed method utilizing $\hat{P}_m$ also provides the true clustering results with high probability. □

Proof of Theorem 1. The basic idea of proof is similar to Theorem S1. Let S_{ml} be the similarity matrix constructed by the kernel function $K_{\sigma,g}$, where σ is the mth component of $\{1, 1.25, \cdots, 2\}$ and g is the lth component of $\{10, 12, \cdots, 30\}$, respectively, as in Section 2.2 of the main paper. Let $\{S_{ml}^*\}$ be the set of similarity matrices constructed by the multiple Gaussian kernels by adding sufficiently small weights $1/n^2$ on some components of $\{S_{ml}\}$ as in the proof of Theorem S1. Let $G_{ml}^* = (D_{ml}^*)^{-1/2} S_{ml}^* (D_{ml}^*)^{-1/2}$ be the corresponding graph Laplacian, where $D_{ml}^* = \text{diag}(d_1^*, \cdots, d_n^*)$ is the degree matrix and d_i^* is a degree of the sample i. Let L_m^* be the n by C^* matrix consisting of C^* eigenvectors corresponding to the first C^* largest eigenvalues of $S_m^* := \sum_l \hat{w}_{ml} G_{ml}^*$

First, by Lemma S6, with probability at least $1 - 2\ell(C^*)^2/n$, we have $(S_m^*)_{i,j} = 0$ if $f(i) \neq f(j)$. Furthermore, we have $(L_m^*)_{i,f(i)} = \frac{\sqrt{d_i^*}}{\sqrt{\sum_{j:\ f(j)=f(i)} d_j^*}}$ and $(L_m^*)_{i,j} = 0$ when $j \neq f(i)$. Let $P_m^* = L_m^*(L_m^*)^T$ be the underlying projection matrix. By Lemma S5, with probability $1 - 2n^{-1}$, we have

$$g\exp(-10) \leq g\exp(-10\sigma^{-2}) \leq d_i^* \leq 2g\exp(-0.1\sigma^{-2}) \leq 2g\exp(-0.1/4).$$

for all i. Hence, it holds that

$$\sqrt{\frac{\exp(-9.9\sigma^{-2})}{2(1-c_1)n}} \leq \sqrt{\frac{\exp(-9.9\sigma^{-2})}{2n_{f(i)}}} \leq (L_m^*)_{i,f(i)} \leq \sqrt{\frac{2\exp(9.9\sigma^{-2})}{n_{f(i)}}} \leq \sqrt{\frac{2\exp(9.9\sigma^{-2})}{c_1 n}}. \tag{S10}$$

Since $\{\hat{P}_m\}$ is the solution to (4) in the main paper, by comparing the objective function values between $\{\hat{P}_1, \cdots, \hat{P}_M\}$ and $\{\hat{P}_1, \cdots, \hat{P}_{m-1}, P_m^*, \hat{P}_{m+1}, \cdots, \hat{P}_M\}$ given $\{\hat{c}_m\}$ and $\{\hat{w}_{ml}\}$, we have for each m,

$$\begin{aligned} & \epsilon\hat{c}_m\|\hat{P}_m\|_F^2 - \hat{c}_m\langle\hat{S}_m, \hat{P}_m\rangle + \lambda\hat{c}_m\|\hat{P}_m\|_1 + \mu\sum_{j\neq m}\|\hat{P}_m - \hat{P}_j\|_F^2 \\ \leq \quad & \epsilon\hat{c}_m\|P_m^*\|_F^2 - \hat{c}_m\langle\hat{S}_m, P_m^*\rangle + \lambda\hat{c}_m\|P_m^*\|_1 + \mu\sum_{j\neq m}\|P_m^* - \hat{P}_j\|_F^2, \end{aligned}$$

where $\hat{S}_m = \sum_l \hat{w}_{ml}G_{ml}$. Hence, we have

$$\begin{aligned} 0 \quad \leq \quad & \epsilon\hat{c}_m\left(\|P_m^*\|_F^2 - \|\hat{P}_m\|_F^2\right) + \hat{c}_m\langle\hat{S}_m, \hat{P}_m - P_m^*\rangle + \lambda\hat{c}_m\left(\|P_m^*\|_1 - \|\hat{P}_m\|_1\right) \\ & +\mu\left(\sum_{j\neq m}\|P_m^* - \hat{P}_j\|_F^2 - \sum_{j\neq m}\|\hat{P}_m - \hat{P}_j\|_F^2\right) \\ \leq \quad & (\epsilon\hat{c}_m + \mu M)\left(\|P_m^*\|_F^2 - \|\hat{P}_m\|_F^2\right) + \langle\hat{c}_m\hat{S}_m + 2\mu\sum_{j\neq m}\hat{P}_j, \hat{P}_m - P_m^*\rangle \\ & +\lambda\hat{c}_m\left(\|P_m^*\|_1 - \|\hat{P}_m\|_1\right). \end{aligned}$$

Since P_m^* is the projection matrix onto the subspace spanned by the eigenvectors of S_m^* corresponding to its C^* largest eigenvalues, and $\lambda_{C^*}(S_m^*) -$

$\lambda_{C^*+1}(S_m^*) = 1 - 0 = 1$, it holds that by Lemma S1,

$$\begin{aligned}
\frac{1}{2}\|\hat{P}_m - P_m^*\|_F^2 &\leq \langle S_m^*, P_m^* - \hat{P}_m\rangle \\
&= \langle S_m^* - \hat{S}_m, P_m^* - \hat{P}_m\rangle + \langle \hat{S}_m, P_m^* - \hat{P}_m\rangle \\
&\leq \langle S_m^* - \hat{S}_m, P_m^* - \hat{P}_m\rangle + (\epsilon + \mu M/\hat{c}_m)\left(\|P_m^*\|_F^2 - \|\hat{P}_m\|_F^2\right) \\
&\quad + \langle 2\mu \sum_{j\neq m} \hat{P}_j/\hat{c}_m, \hat{P}_m - P_m^*\rangle + \lambda\left(\|P_m^*\|_1 - \|\hat{P}_m\|_1\right) \\
&\leq \|S_m^* - \hat{S}_m - 2\mu \sum_{j\neq m} \hat{P}_j/\hat{c}_m\|_{\max}\, \|P_m^* - \hat{P}_m\|_1 \\
&\quad + \lambda\left(\|P_m^*\|_1 - \|\hat{P}_m\|_1\right) + C^*\|\Delta_m\|_F(\epsilon + \mu M/\hat{c}_m) \\
&\lesssim \lambda/\hat{c}_m\left(\|\hat{P}_m - P_m^*\|_1 + \|P_m^*\|_1 - \|\hat{P}_m\|_1\right) + C^*\|\Delta_m\|_F(\epsilon + \mu M/\hat{c}_m) \\
&\leq \lambda/\hat{c}_m\left(\|\Delta_m\|_1 + \|P_m^*\|_1 - \|P_m^* + \Delta_m\|_1\right) + C^*\|\Delta_m\|_F(\epsilon + \mu M/\hat{c}_m),
\end{aligned}$$

where $\Delta_m = \hat{P}_m - P_m^*$, the third and fourth inequality uses $\|P_m^*\|_F^2, \|\hat{P}_m\|_F^2 \leq C^*$ and $\|S_m^* - \hat{S}_m\|_{\max} \leq 1/n^2 \leq \lambda$ and $\mu = \sqrt{n}\lambda/(C^*M)$. Since $\epsilon = 1/n^3$, we have

$$\begin{aligned}
\frac{1}{2}\|\Delta_m\|_F^2 &\lesssim 2\lambda\|\Delta_m\|_1/\hat{c}_m + C^*\|\Delta_m\|_F(\epsilon + \mu M/\hat{c}_m) \\
&\lesssim 2\lambda\sqrt{n}\|\Delta_m\|_F/\hat{c}_m + C^*\|\Delta_m\|_F(n^{-3} + \sqrt{n}\lambda/(\hat{c}_m C^*)) \\
&\lesssim \lambda\sqrt{n}C^*\|\Delta_m\|_F/\hat{c}_m.
\end{aligned}$$

Solving the above inequalities, we have $\|\Delta_m\|_F \lesssim \lambda\sqrt{n}C^*/\hat{c}_m$.

Since $\lambda = C^*/n^2$, we have $\|\hat{P}_m - P_m^*\|_F \lesssim 1/(\hat{c}_m n^{3/2}) \lesssim \exp(3C^*)/(Mn^{3/2})$ due to Lemma S11 with a entropy penalty function. This means that all the eigenvalues of $\hat{P}_m$ are near zero or one, i.e., $\hat{P}_m$ is very close to the following original nonlinear space:

$$\{P:\ \mathrm{tr}(P) = C^*,\ P = LL^T \text{ for some } L \text{ with } L^TL = I_{C^*}\}.$$

Hence, by using the sin Θ theorem (Davis and Kahan, 1970), we have

$$\|\hat{L}_m \hat{L}_m^T - L_m^*(L_m^*)^T\|_F \lesssim M^{-1}n^{-3/2}. \tag{S11}$$

Let $\tilde{L} = [\hat{c}_1\hat{L}_1, \cdots, \hat{c}_M\hat{L}_M]$. Then, for samples i and j from the same cluster, it holds that

$$\begin{aligned}
&\frac{\langle \tilde{L}_i, \tilde{L}_j\rangle}{\|\tilde{L}_i\|_2\|\tilde{L}_j\|_2} = \frac{\sum_{k=1}^M \hat{c}_k^2 \langle (\hat{L}_k)_i, (\hat{L}_k)_j\rangle}{\sqrt{\sum_{k=1}^M \hat{c}_k^2 \|(\hat{L}_k)_i\|^2}\sqrt{\sum_{k=1}^M \hat{c}_k^2 \|(\hat{L}_k)_j\|^2}} \\
&\geq \frac{\sum_{k=1}^M \hat{c}_k^2 \langle (L_k^*)_i, (L_k^*)_j\rangle - \sum_{k=1}^M \hat{c}_k^2 M^{-1}n^{-3/2}}{\sqrt{\sum_{k=1}^M \hat{c}_k^2 \|(L_k^*)_i\|^2 + \sum_{k=1}^M \hat{c}_k^2 M^{-1}n^{-3/2}}\sqrt{\sum_{k=1}^M \hat{c}_k^2 \|(L_k^*)_j\|^2 + \sum_{k=1}^M \hat{c}_k^2 M^{-1}n^{-3/2}}} \\
&\gtrsim 1 - n^{-1/2},
\end{aligned}$$

where we utilize (S11). Thus, we have

$$\left\| \tilde{L}_i/\|\tilde{L}_i\|_2 - \tilde{L}_j/\|\tilde{L}_j\|_2 \right\|_2 \lesssim n^{-1/4}.$$

For the samples i and j from different clusters, we have

$$\begin{aligned}
\frac{\langle \tilde{L}_i, \tilde{L}_j\rangle}{\|\tilde{L}_i\|_2\|\tilde{L}_j\|_2} &\leq \frac{\sum_{k=1}^M \hat{c}_k^2 \langle (L_k^*)_i, (L_k^*)_j\rangle + \sum_{k=1}^M \hat{c}_k^2 M^{-1}n^{-3/2}}{\sqrt{\sum_{k=1}^M \hat{c}_k^2 \|(L_k^*)_i\|^2 - \sum_{k=1}^M \hat{c}_k^2 M^{-1}n^{-3/2}}\sqrt{\sum_{k=1}^M \hat{c}_k^2 \|(L_k^*)_j\|^2 - \sum_{k=1}^M \hat{c}_k^2 M^{-1}n^{-3/2}}} \\
&\lesssim n^{-1/2},
\end{aligned}$$

where the second inequality follows from the fact that $\langle (L_k^*)_i, (L_k^*)_j\rangle = 0$ for the i and j from different clusters and (S10). Thus, we have

$$\left\| \tilde{L}_i/\|\tilde{L}_i\|_2 - \tilde{L}_j/\|\tilde{L}_j\|_2 \right\|_2 \gtrsim \sqrt{2 - 2n^{-1/2}}.$$

By Lemma S4, the k-means clustering algorithm with the normalized $\tilde{L}$ guarantees the exact clustering results. This completes the proof. □

C Proof of computational convergence

Throughout the proof, we write $a = O(b)$ or $a \lesssim b$ if $a \leq Cb$ for some positive constants C. If $a \geq Cb$ for some positive constants C, then we write $a = \Omega(b)$ or $a \gtrsim b$. We use $a \asymp b$ when $a \lesssim b$ and $b \lesssim a$. For an n by p matrix A, let $\text{vec}(A)$ be the np by 1 column vector obtained by stacking the columns of A on top of one another. We first state the following Lemma S2 and Lemma S3 with the detailed proof at the end of this section, which will be used to prove the global convergence of the proposed algorithm (Theorem S2).

Lemma S2. *Let $F(\{P_m\}, \{c_m\}, \{w_{ml}\})$ be the objective function of (4) in the main paper. Let $\{\bar{P}_m\}_i, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i$ be the iterates of* (S1)-(S3) *by assuming that $\{\bar{P}_m\}_{i+1}$ is the minimum point of* (S3) *given $\{\bar{c}_m\}_{i+1}$ and $\{\bar{w}_{ml}\}_{i+1}$ for each i. Then, the iterate $\{\bar{P}_m\}_i, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i$ converges to some stationary point $\{P^*_m\}, \{c^*_m\}, \{w^*_{ml}\}$ of $F(\cdot)$ in the sense that there exist certain $i_0 > 0$, $\theta \in (1/2, 1)$, and $C_4 > 0$ such that*

$$\sum_m \|\bar{P}_m^{(i)} - P^*_m\|_F + \sum_m \|\bar{c}_m^{(i)} - c^*_m\|_F + \sum_m \|\bar{w}_m^{(i)} - w^*_m\|_F \leq C_4 i^{-(1-\theta)/(2\theta-1)} \quad \text{(S12)}$$

for all $i \geq i_0$. Moreover, it holds that

$$\begin{aligned} & F(\{\bar{P}_m\}_{i-1}, \{\bar{c}_m\}_{i-1}, \{\bar{w}_{ml}\}_{i-1}) - F(\{\bar{P}_m\}_i, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i) \\ \geq \; & \frac{\rho}{2M}\|\bar{c}^{(i-1)} - \bar{c}^{(i)}\|^2 + \frac{\rho}{2}\sum_m \|\bar{W}_m^{(i-1)} - \bar{W}_m^{(i)}\|^2, \end{aligned}$$

where $\bar{c}^{(i)} = [\bar{c}_1^{(i)}, \cdots, \bar{c}_M^{(i)}]^T$ and $\bar{W}_m^i = [\bar{w}_{m1}^{(i)}, \cdots, \bar{w}_{m\ell}^{(i)}]^T$, that is, the objective value $F(\{\bar{P}_m\}_i, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i)$ is monotonically decreasing.

Lemma S2 shows the convergence property of the iterates $\{\bar{P}_m\}_i, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i$. Note that the sequence $\{\bar{P}_m\}_i, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i$ is oracle in the sense that it uses the optimal point of (S3) for updating $\{\bar{P}_m\}_{i+1}$ instead of updates via ADMM.

Lemma S3 shows the difference of the ADMM iterates in (S3) when different sets of $\{c_m\}_i$ and $\{w_{ml}\}_i$ are entered in (S3).

Lemma S3. *For fixed iterate number i, let $\{\hat{P}_m\}_{i,t} = \{\hat{P}_1^{i,t}, \cdots, \hat{P}_M^{i,t}\}$, $\{\hat{Q}_m\}_{i,t} = \{\hat{Q}_1^{i,t}, \cdots, \hat{Q}_M^{i,t}\}$, and $\{\hat{\Gamma}_m\}_{i,t} = \{\hat{\Gamma}_1^{i,t}, \cdots, \hat{\Gamma}_M^{i,t}\}$ be the t-th iterate of the ADMM in* (S3) *given $\{\hat{c}_m\}_i$ and $\{\hat{w}_{ml}\}_i$. Similarly, let $\{\bar{P}_m\}_{i,t} = \{\bar{P}_1^{i,t}, \cdots, \bar{P}_M^{i,t}\}$, $\{\bar{Q}_m\}_{i,t} = \{\bar{Q}_1^{i,t}, \cdots, \bar{Q}_M^{i,t}\}$, and $\{\bar{\Gamma}_m\}_{i,t} = \{\bar{\Gamma}_1^{i,t}, \cdots, \bar{\Gamma}_M^{i,t}\}$ be the t-th iterate of the ADMM in* (S3) *given $\{\bar{c}_m\}_i$ and $\{\bar{w}_{ml}\}_i$. Then, we have*

$$\begin{aligned}
\|\hat{P}_m^{i,t} - \bar{P}_m^{i,t}\|_F &\lesssim |\hat{c}_m^{(i)} - \bar{c}_m^{(i)}| + \max_l |\hat{w}_{ml}^{(i)} - \bar{w}_{ml}^{(i)}| + \|\hat{\Gamma}_m^{i,t-1} - \bar{\Gamma}_m^{i,t-1}\|_F \\
&+ \sum_m \|\hat{Q}_m^{i,t-1} - \bar{Q}_m^{i,t-1}\|_F, \\
\|\hat{Q}_m^{i,t} - \bar{Q}_m^{i,t}\|_F &\lesssim \sum_{m'} \|\hat{P}_{m'}^{i,t} - \bar{P}_{m'}^{i,t}\|_F + \|\hat{\Gamma}_m^{i,t-1} - \bar{\Gamma}_m^{i,t-1}\|_F \\
\|\hat{\Gamma}_m^{i,t} - \bar{\Gamma}_m^{i,t}\|_F &\lesssim \|\hat{\Gamma}_m^{i,t-1} - \bar{\Gamma}_m^{i,t-1}\|_F + \|\hat{P}_m^{i,t} - \bar{P}_m^{i,t}\|_F + \|\hat{Q}_m^{i,t} - \bar{Q}_m^{i,t}\|_F.
\end{aligned}$$

Theorem S2 is the main results showing the convergence of the proposed algorithm.

Theorem S2. *Let θ be the constant as in Lemma S2. Let $F(\{P_m\}, \{c_m\}, \{w_{ml}\})$ be the objective function of (4) in the main paper. Let $\{\hat{P}_m\}_i, \{\hat{c}_m\}_i, \{\hat{w}_{ml}\}_i$*

be the obtained iterate of the proposed algorithm in Section A. Then $\{\hat{P}_m\}_i = \{\hat{P}_1^{(i)}, \cdots, \hat{P}_M^{(i)}\}$, $\{\hat{c}_m\}_i = \{\hat{c}_1^{(i)}, \cdots, \hat{c}_M^{(i)}\}$, *and* $\{\hat{w}_{ml}\}_i = \{\hat{w}_{11}^{(i)}, \cdots, \hat{w}_{M\ell}^{(i)}\}$ *converges to some stationary point of* F *in the sense that for a fixed tolerance parameter* $\delta > 0$,

$$\sum_m \|\hat{P}_m^{(i^*)} - P_m^*\|_F + \sum_m |\hat{c}_m^{(i^*)} - c_m^*| + \sum_m \sum_l |\hat{w}_{ml}^{(i^*)} - w_{ml}^*| \leq C_6 \delta$$

for some absolute constant $C_6 > 0$, *with the iterate number* $i^* \asymp \delta^{-(2\theta-1)/(1-\theta)}$, *and the iterate number of ADMM for* (S3) *being* $t_{i^*} \asymp \log(\delta/i^*)/\log(\mu)$ *and* $t_{i^*-k} \asymp t_{i^*}(t_{i^*}+1)^{k-1}$ *for all* $1 \leq k \leq i^* - 1$.

We first include the proofs of Lemma S2 and Lemma S3 followed by the proof of Theorem S2.

Proof of Lemma S2. Note that the iterates $\{\bar{P}_m\}_i, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i$ are obtained via block coordinate descent method (Xu and Yin, 2013) as follows: at the ith iterate of each update,

$$\begin{aligned} \{\bar{c}_m\}_i &= \operatorname*{argmin}_{\{c_m\}} F(\{\bar{P}_m\}_{i-1}, \{c_m\}, \{\bar{w}_{ml}\}_{i-1}) && \text{(S13)} \\ \{\bar{w}_{ml}\}_i &= \operatorname*{argmin}_{\{w_{ml}\}} F(\{\bar{P}_m\}_{i-1}, \{\bar{c}_m\}_i, \{w_{ml}\}) && \text{(S14)} \\ \{\bar{P}_m\}_i &= \operatorname*{argmin}_{\{P_m\}} F(\{P_m\}, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i). && \text{(S15)} \end{aligned}$$

We rewrite $F(\{P_m\}, \{c_m\}, \{w_{ml}\})$ as follows:

$$\begin{aligned} & F(\{P_m\}, \{c_m\}, \{w_{ml}\}) \\ = \quad & \epsilon \sum_m c_m \|P_m\|_F^2 - \sum_m c_m \langle S_m, P_m \rangle + \mu \sum_{m \neq j} \|P_m - P_j\|_F^2 \\ & + \sum_m g_c(c_m) + \sum_m \sum_l g_w(w_{ml}) + \lambda \sum_m c_m \|P_m\|_1 \\ := \quad & f(\{P_m\}, \{c_m\}, \{w_{ml}\}) + \lambda \sum_m c_m \|P_m\|_1, \end{aligned}$$

where $f(\{P_m\}, \{c_m\}, \{w_{ml}\})$ is a differentiable and block multiconvex function (Xu and Yin, 2013), and $\lambda \sum_m c_m \|P_m\|_1$ is a convex block-separable function. Now to apply the results in Section 2.3 of Xu and Yin (2013), we check the following required conditions as in (A1)-(A5):

(A1) $F(\{P_m\}, \{c_m\}, \{w_{ml}\})$ is a continous function.

(A2) The $F(\{P_m\}, \{c_m\}, \{w_{ml}\})$ is lower bounded as follows:

$$F(\{P_m\}, \{c_m\}, \{w_{ml}\}) \geq \epsilon \sum_m c_m \|P_m\|_F^2 - \sum_m c_m \langle S_m, P_m \rangle \geq - \sum_m \frac{\|S_m\|_F^2}{4\epsilon c_m} > -\infty.$$

(A3) $f(\{P_m\}, \{c_m\}, \{w_{ml}\})$ is a strongly convex function of any variable given the other variables are fixed. Let $P = [\text{vec}(P_1)^T, \cdots, \text{vec}(P_M)^T]^T$, $c = [c_1, \cdots, c_M]^T$, and $w = [\text{vec}(w_1)^T, \cdots, \text{vec}(w_M)^T]^T$. Then, the following hold:

$$\begin{aligned} f(P, c, w) - f(\bar{P}, c, w) & \geq \langle \nabla_P f(\bar{P}, c, w), P - \bar{P} \rangle + \mu(M-1) \|P - \bar{P}\|^2 \\ f(P, c, w) - f(P, \bar{c}, w) & \geq \langle \nabla_c f(P, \bar{c}, w), c - \bar{c} \rangle + \frac{\rho}{2M} \|c - \bar{c}\|^2 \\ f(P, c, w) - f(P, c, \bar{w}) & \geq \langle \nabla_w f(P, c, \bar{w}), w - \bar{w} \rangle + \frac{\rho}{2} \|w - \bar{w}\|^2. \end{aligned}$$

(A4) $\nabla_P f(\bar{P}, c, w)$ is Lipschitz continuous on any bounded set because it holds that

$$\begin{aligned}\nabla_P f(P,c,w) &= 2\epsilon[c_1 P_1^T, \cdots, c_M P_M^T]^T - [c_1 \mathrm{vec}(S_1)^T, \cdots c_M \mathrm{vec}(S_M)^T]^T \\ &+ 2\mu(M-1)[P_1^T, \cdots, P_M^T]^T - [\mathrm{vec}(\sum_{j\neq 1} P_j)^T, \cdots, \mathrm{vec}(\sum_{j\neq M} P_j)^T]^T,\end{aligned}$$

thus we have

$$\|\nabla_P f(P,c,w) - \nabla_P f(\bar{P},c,w)\| \leq (2\epsilon M + 2\mu M + M^2)\|P - \bar{P}\|.$$

Similarly, we have

$$\begin{aligned}\|\nabla_c f(P,c,w) - \nabla_c f(P,\bar{c},w)\| &\leq \frac{1}{c_{\min}}\|c - \bar{c}\|, \\ \|\nabla_w f(P,c,w) - \nabla_w f(P,c,\bar{w})\| &\leq \frac{1}{w_{\min}}\|w - \bar{w}\|\end{aligned}$$

for some positive constants $c_{\min}$ and $w_{\min}$.

(A5) We note that $f(P,c,w)$ and $g(P,c) = \lambda \sum_m c_m \|P_m\|_1$ are both subanalytic and g maps bounded sets to bounded sets, thus $f+g$ is also subanalytic and satisfies the Kurdyka-Lojasiewicz (KL) inequality. See Xu and Yin (2013) for details of KL inequality.

By (A1)-(A5) and Theorem 2.9 of Xu and Yin (2013), there exist certain $i_0 > 0$, $\theta \in (1/2, 1)$, and $C > 0$ such that

$$\|\bar{P}^i - P^*\|_F + \|\bar{c}^i - c^*\|_F + \|\bar{w}^i - w^*\|_F \leq C i^{-(1-\theta)/(2\theta-1)} \tag{S16}$$

for all $i \geq i_0$. Here $P^* = [\mathrm{vec}(P_1^*)^T, \cdots, \mathrm{vec}(P_M^*)^T]^T$, $c^* = [c_1^*, \cdots, c_M^*]^T$, and $w^* = [\mathrm{vec}(w_1^*)^T, \cdots, \mathrm{vec}(w_M^*)^T]^T$ is some stationary point of (4) in the main paper. Since M is finite, this completes the proof. □

Proof of Lemma S3. Let

$$\hat{t}_m = \hat{c}_m^{(i)} \hat{S}_m^{(i)} - \hat{\Gamma}_m^{i,t-1} + \eta \hat{Q}_m^{i,t-1} + 2\mu \sum_{\widetilde{m} \neq m} \hat{Q}_{\widetilde{m}}^{i,t-1},$$
$$\bar{t}_m = \bar{c}_m^{(i)} \bar{S}_m^{(i)} - \bar{\Gamma}_m^{i,t-1} + \eta \bar{Q}_m^{i,t-1} + 2\mu \sum_{\widetilde{m} \neq m} \bar{Q}_{\widetilde{m}}^{i,t-1}.$$

Then, for fixed $k, j \in \{1, \cdots, n\}$, we have as in Section A.3.1,

$$\begin{aligned}
(\hat{P}_m^{i,t})_{kj} &= (\hat{t}_{m,kj} - \lambda)(2\epsilon \hat{c}_m^{(i)} + \eta + 2\mu(M-1))^{-1} \quad \text{if} \quad \hat{t}_{m,kj} > \lambda \\
(\hat{P}_m^{i,t})_{kj} &= (\hat{t}_{m,kj} + \lambda)(2\epsilon \hat{c}_m^{(i)} + \eta + 2\mu(M-1))^{-1} \quad \text{if} \quad \hat{t}_{m,kj} < -\lambda \\
(\hat{P}_m^{i,t})_{kj} &= 0 \quad \text{if} \quad |\hat{t}_{m,kj}| \leq \lambda.
\end{aligned}$$

Similarly, we define $(\bar{P}_m^{i,t})_{kj}$ using $\bar{t}_{m,kj}$. Since

$$\eta + 2\mu(M-1) \leq 2\epsilon \hat{c}_m^{(i+1)} + \eta + 2\mu(M-1) \leq 2\epsilon M + \eta + 2\mu(M-1),$$

it holds that for all $k, j \in \{1, \cdots, n\}$

$$|(\hat{P}_m^{i,t})_{kj} - (\bar{P}_m^{i,t})_{kj}| \leq \frac{|\hat{t}_{m,kj} - \bar{t}_{m,kj}|}{\eta + 2\mu(M-1)},$$

thus we have

$$\begin{aligned}
& \|\hat{P}_m^{i,t} - \bar{P}_m^{i,t}\|_F \lesssim \|\hat{t}_m - \bar{t}_m\|_F \qquad \text{(S17)} \\
\lesssim \ & |\hat{c}_m^{(i)} - \bar{c}_m^{(i)}| \|\hat{S}_m^{(i)}\|_F + \bar{c}_m^{(i)} \|\hat{S}_m^{(i)} - \bar{S}_m^{(i)}\|_F + \|\hat{\Gamma}_m^{i,t-1} - \bar{\Gamma}_m^{i,t-1}\|_F + \sum_m \|\hat{Q}_m^{i,t-1} - \bar{Q}_m^{i,t-1}\|_F \\
\lesssim \ & |\hat{c}_m^{(i)} - \bar{c}_m^{(i)}| + \max_l |\hat{w}_{ml}^{(i)} - \bar{w}_{ml}^{(i)}| + \|\hat{\Gamma}_m^{i,t-1} - \bar{\Gamma}_m^{i,t-1}\|_F + \sum_m \|\hat{Q}_m^{i,t-1} - \bar{Q}_m^{i,t-1}\|_F,
\end{aligned}$$

where we use $\|\hat{S}_m^{(i)}\|_F \leq 1$ and $\|\hat{S}_m^{(i)} - \bar{S}_m^{(i)}\|_F \leq 55 \max_l |\hat{w}_{ml}^{(i)} - \bar{w}_{ml}^{(i)}|$. See Lemma S7 and the proof of Lemma S8 for details of these inequalities.

For the updating $\{\hat{Q}_m\}_{i,t}$ part, we have as in Section A.3.2,

$$\hat{Q}_m^{i,t} = \operatorname*{argmin}_Q \|Q - \hat{T}_m\|_F^2 \quad \text{s.t.} \quad \operatorname{tr}(Q) = C^*, \quad 0 \preceq Q \preceq I,$$

where $\hat{T}_m = (2\mu \sum_{\widetilde{m} \neq m} \hat{P}_{\widetilde{m}}^{i,t} + \hat{\Gamma}_m^{i,t-1} + \eta \hat{P}_m^{i,t})(2\mu(M-1)+\eta)^{-1}$. Similarly, we define $\bar{T}_m$ using $\bar{P}_{\widetilde{m}}^{i,t}$ and $\bar{\Gamma}_m^{i,t-1}$. By Lemma S10, we have

$$\|\hat{Q}_m^{i,t} - \bar{Q}_m^{i,t}\|_F \le \|\hat{T}_m - \bar{T}_m\|_F \lesssim \sum_{m'} \|\hat{P}_{m'}^{i,t} - \bar{P}_{m'}^{i,t}\|_F + \|\hat{\Gamma}_m^{i,t-1} - \bar{\Gamma}_m^{i,t-1}\|_F. \tag{S18}$$

For updating $\{\hat{\Gamma}_m\}_{i,t} = \{\hat{\Gamma}_1^{i,t}, \cdots, \hat{\Gamma}_M^{i,t}\}$, as in Section A.3.3, we have $\hat{\Gamma}_m^{i,t} = \hat{\Gamma}_m^{i,t-1} + \eta(\hat{P}_m^{i,t} - \hat{Q}_m^{i,t})$. Hence

$$\|\hat{\Gamma}_m^{i,t} - \bar{\Gamma}_m^{i,t}\|_F \lesssim \|\hat{\Gamma}_m^{i,t-1} - \bar{\Gamma}_m^{i,t-1}\|_F + \|\hat{P}_m^{i,t} - \bar{P}_m^{i,t}\|_F + \|\hat{Q}_m^{i,t} - \bar{Q}_m^{i,t}\|_F. \tag{S19}$$

This completes the proof.

□

Proof of Theorem S2. Note that the sequence $\{\bar{P}_m\}_i, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i$ in Lemma S2 is oracle in the sense that it uses the optimal point of (S3) for updating $\{\bar{P}_m\}_{i+1}$. In this proof, we investigate the differences of the actual iterates $\{\hat{P}_m\}_i, \{\hat{c}_m\}_i, \{\hat{w}_{ml}\}_i$ and the oracle iterates $\{\bar{P}_m\}_i, \{\bar{c}_m\}_i, \{\bar{w}_{ml}\}_i$. Throughout the proof, we use the hat($\hat{a}$) and bar($\bar{a}$) notations for the parameter a if a is associated with the actual and oracle iterates, respectively.

For updating $\{\hat{c}_m\}_i = \{\hat{c}_1^{(i)}, \cdots, \hat{c}_M^{(i)}\}$ as in Section A.1, we have

$$\hat{c}_m^{(i)} = \frac{M\hat{d}_m^{(i)}}{\sum_m \hat{d}_m^{(i)}}, \quad \bar{c}_m^{(i)} = \frac{M\bar{d}_m^{(i)}}{\sum_m \bar{d}_m^{(i)}},$$

where

$$\hat{d}_m^{(i)} = \exp\left(\frac{1}{\rho}\left(-\epsilon\|\hat{P}_m^{(i-1)}\|_F^2 + \langle \hat{S}_m^{(i-1)}, \hat{P}_m^{(i-1)} \rangle\right)\right), \quad \hat{S}_m^{(i)} = \sum_l \hat{w}_{ml}^{(i)} G_{ml},$$

$$\bar{d}_m^{(i)} = \exp\left(\frac{1}{\rho}\left(-\epsilon\|\bar{P}_m^{(i-1)}\|_F^2 + \langle \bar{S}_m^{(i-1)}, \bar{P}_m^{(i-1)}\rangle\right)\right), \quad \bar{S}_m^{(i)} = \sum_l \bar{w}_{ml}^{(i)} G_{ml}.$$

By Lemma S8, we have

$$\begin{aligned} |\hat{c}_m^{(i)} - \bar{c}_m^{(i)}| &\leq \frac{M|\hat{d}_m^{(i)} - \bar{d}_m^{(i)}|}{\sum_m \bar{d}_m^i} + \frac{M\hat{d}_m^{(i)}|\sum_m \hat{d}_m^{(i)} - \sum_m \bar{d}_m^{(i)}|}{\sum_m \hat{d}_m^{(i)} \sum_m \bar{d}_m^{(i)}} \\ &\lesssim \|\hat{P}_m^{(i-1)} - \bar{P}_m^{(i-1)}\|_F + \|\hat{w}_m^{(i-1)} - \bar{w}_m^{(i-1)}\|_F. \end{aligned} \tag{S20}$$

For updating $\{\hat{w}_{ml}\}_i = \{\hat{w}_{11}^{(i)}, \cdots, \hat{w}_{M\ell}^{(i)}\}$ as in Section A.2, we have

$$\hat{w}_{ml}^{(i)} = \frac{\exp\left(\frac{\hat{c}_m^{(i)}}{\rho}\langle G_{ml}, \hat{P}_m^{(i-1)}\rangle\right)}{\sum_{\widetilde{m}} \exp\left(\frac{\hat{c}_{\widetilde{m}}^{(i)}}{\rho}\langle G_{\widetilde{m}l}, \hat{P}_{\widetilde{m}}^{(i-1)}\rangle\right)}, \quad \bar{w}_{ml}^{(i)} = \frac{\exp\left(\frac{\bar{c}_m^{(i)}}{\rho}\langle G_{ml}, \bar{P}_m^{(i-1)}\rangle\right)}{\sum_{\widetilde{m}} \exp\left(\frac{\bar{c}_{\widetilde{m}}^{(i)}}{\rho}\langle G_{\widetilde{m}l}, \bar{P}_{\widetilde{m}}^{(i-1)}\rangle\right)}$$

By Lemma S9, it holds that for $\hat{w}_m^{(i)} = [\hat{w}_{m1}^{(i)}, \cdots, \hat{w}_{m\ell}^{(i)}]^T$ and $\bar{w}_m^{(i)} = [\bar{w}_{m1}^{(i)}, \cdots, \bar{w}_{m\ell}^{(i)}]^T$,

$$\|\hat{w}_m^{(i)} - \bar{w}_m^{(i)}\|_F \lesssim \|\hat{P}_m^{(i-1)} - \bar{P}_m^{(i-1)}\|_F + \|\hat{c}^{(i)} - \bar{c}^{(i)}\|_F. \tag{S21}$$

For updating $\{\hat{P}_m\}_i = \{\hat{P}_1^{(i)}, \cdots, \hat{P}_M^{(i)}\}$, note that we use the ADMM as described in Section A.3. Let $\{\hat{P}_m\}_{i,t} = \{\hat{P}_1^{i,t}, \cdots, \hat{P}_M^{i,t}\}$, $\{\hat{Q}_m\}_{i,t} = \{\hat{Q}_1^{i,t}, \cdots, \hat{Q}_M^{i,t}\}$, and $\{\hat{\Gamma}_m\}_{i,t} = \{\hat{\Gamma}_1^{i,t}, \cdots, \hat{\Gamma}_M^{i,t}\}$ be the t-th iterate of the ADMM given $\{\hat{c}_m\}_i$ and $\{\hat{w}_{ml}\}_i$. Similarly, let $\{\bar{P}_m\}_{i,t} = \{\bar{P}_1^{i,t}, \cdots, \bar{P}_M^{i,t}\}$, $\{\bar{Q}_m\}_{i,t} = \{\bar{Q}_1^{i,t}, \cdots, \bar{Q}_M^{i,t}\}$, and $\{\bar{\Gamma}_m\}_{i,t} = \{\bar{\Gamma}_1^{i,t}, \cdots, \bar{\Gamma}_M^{i,t}\}$ be the t-th iterate of the ADMM given $\{\bar{c}_m\}_i$ and $\{\bar{w}_{ml}\}_i$.

For simplicity, let

$$\begin{aligned} \epsilon_c^{(i)} &:= \|\hat{c}^{(i)} - \bar{c}^{(i)}\|_F, \quad \epsilon_w^{(i)} := \|\hat{w}^{(i)} - \bar{w}^{(i)}\|_F, \quad \epsilon_P^{(i)} := \|\hat{P}^{(i)} - \bar{P}^{(i)}\|_F \\ \epsilon_P^{it} &:= \|\hat{P}^{i,t} - \bar{P}^{i,t}\|_F, \quad \epsilon_Q^{it} := \|\hat{Q}^{i,t} - \bar{Q}^{i,t}\|_F, \quad \epsilon_\Gamma^{it} := \|\hat{\Gamma}^{i,t} - \bar{\Gamma}^{i,t}\|_F, \end{aligned}$$

where $\hat{P}^{(i)} = \text{vec}(\text{vec}(\hat{P}_1^{(i)}), \cdots, \text{vec}(\hat{P}_M^{(i)}))$ is the obtained output from ADMM with the iterate number T_i, which will be defined later, and $\hat{w}^{(i)} = (\hat{w}_{11}^{(i)}, \cdots, \hat{w}_{M\ell}^{(i)})^T$

and $\hat{c}^{(i)} = (\hat{c}_1^{(i)}, \cdots, \hat{c}_M^{(i)})^T$. Similarly, $\bar{P}^{(i)} = \text{vec}(\text{vec}(\bar{P}_1^{(i)}), \cdots, \text{vec}(\bar{P}_M^{(i)}))$, $\bar{w}^{(i)} = (\bar{w}_{11}^{(i)}, \cdots, \bar{w}_{M\ell}^{(i)})^T$ and $\bar{c}^{(i)} = (\bar{c}_1^{(i)}, \cdots, \bar{c}_M^{(i)})^T$.

By Lemma S3, we have

$$\|\hat{P}_m^{i,1} - \bar{P}_m^{i,1}\|_F \lesssim \epsilon_c^{(i)} + \epsilon_w^{(i)}, \quad \|\hat{Q}_m^{i,1} - \bar{Q}_m^{i,1}\|_F \lesssim \epsilon_c^{(i)} + \epsilon_w^{(i)}, \quad \|\hat{\Gamma}_m^{i,1} - \bar{\Gamma}_m^{i,1}\|_F \lesssim 2\epsilon_c^{(i)} + 2\epsilon_w^{(i)}.$$

Using recursive formulas in Lemma S3 and the fact that M is finite, we have

$$\sum_m \|\hat{P}_m^{i,t} - \bar{P}_m^{i,t}\|_F + \sum_m \|\hat{Q}_m^{i,t} - \bar{Q}_m^{i,t}\|_F + \sum_m \|\hat{\Gamma}_m^{i,t} - \bar{\Gamma}_m^{i,t}\|_F \lesssim 5^t(\epsilon_c^{(i)} + \epsilon_w^{(i)}). \quad \text{(S22)}$$

Note that since (S5) is convex in $\{P_m\}$, it holds that for all t and some $\mu \in (0,1)$ (Lin *et al.*, 2015; Deng and Yin, 2016)

$$\|\bar{P}^{i,t} - \bar{P}^{(i)}\|_F \lesssim \mu^t. \quad \text{(S23)}$$

By (S22) and (S23), we have

$$\|\hat{P}^{i,t} - \bar{P}^{(i)}\|_F \lesssim 5^t(\epsilon_c^{(i)} + \epsilon_w^{(i)}) + \mu^t. \quad \text{(S24)}$$

Note that since $\epsilon_c^{(1)} = \epsilon_w^{(1)} = 0$, we have $\epsilon_P^{(1)} \lesssim \mu^{T_1}$. Using recursive formulas (S20), (S21), and (S24) with $\hat{P}^{(j)} = \hat{P}^{j,T_j}$ for $1 \le j \le i$, we have for each i,

$$\|\hat{P}^{(i)} - \bar{P}^{(i)}\|_F + \|\hat{c}^{(i)} - \bar{c}^{(i)}\|_F + \|\hat{w}^{(i)} - \bar{w}^{(i)}\|_F \lesssim \sum_{j=1}^{i} \mu^{T_j} 5^{\sum_{k=j+1}^{i} T_k}, \quad \text{(S25)}$$

where letting $T_{i+1} = 0$. For fixed tolerance parameter $\delta > 0$, we choose the iterate number $i^* \asymp \delta^{-(2\theta-1)/(1-\theta)}$, where θ is the constant defined in (S16) in Lemma S2.

Let $\psi \asymp -\log 5/\log\mu$. We set the iterate number of ADMM as $T_{i^*} \asymp \log(\delta/i^*)/\log(\mu)$ and $T_{i^*-k} \asymp \psi T_{i^*}(T_{i^*}+1)^{k-1}$ for all $1 \le k \le i^* - 1$. Then (S25) implies for all $1 \le i \le i^*$,

$$\|\hat{P}^{(i)} - \bar{P}^{(i)}\|_F + \|\hat{c}^{(i)} - \bar{c}^{(i)}\|_F + \|\hat{w}^{(i)} - \bar{w}^{(i)}\|_F \le C_5 \delta i/i^* \quad \text{(S26)}$$

for some absolute constant $C_5 > 0$. Note that we have by the results of Lemma S2,

$$\|\bar{P}^{(i)} - P^*\|_F + \|\bar{c}^{(i)} - c^*\|_F + \|\bar{w}^{(i)} - w^*\|_F \lesssim i^{-(1-\theta)/(2\theta-1)}. \tag{S27}$$

Combining (S26) and (S27), we have

$$\|\hat{P}^{(i^*)} - P^*\|_F + \|\hat{c}^{(i^*)} - c^*\|_F + \|\hat{w}^{(i^*)} - w^*\|_F \leq C_6\delta$$

for some absolute constant $C_6 > 0$. Since M is finite, this completes the proof. □

D Lemmas and proofs

In this section, we include additional lemmas with their proofs that are used to prove the main theorems. We write $[K]$ to denote $\{1, 2, \cdots, K\}$ for any positive integer K.

Lemma S4. *Suppose that $C_1, \cdots, C_K \subseteq [n]$ are partitions of points $y_1, \cdots, y_n$ which satisfying*

$$\max_{k\in[K]} \max_{i,j\in C_k} \|y_i - y_j\|_2 \leq a, \qquad \min_{k\neq\tilde{k}} \min_{i\in C_k,\ j\in C_{\tilde{k}}} \|y_i - y_j\|_2 \geq b$$

with $b > a\sqrt{n}$. Then, the k-means clustering algorithm gives the clusters $C_1, \cdots, C_K$.

Proof of Lemma S4. Note that for the n points $y_1, \cdots, y_n$, we have

$$\begin{aligned}\sum_{i,j} \|y_i - y_j\|_2^2 &= \sum_{i,j} \|y_i - \bar{y} + \bar{y} - y_j\|_2^2 \\ &= 2n \sum_i \|y_i - \bar{y}\|_2^2 + 2\sum_{i,j} \langle y_i - \bar{y}, \bar{y} - y_j \rangle \\ &= 2n \sum_i \|y_i - \bar{y}\|_2^2 + 2\sum_i \langle y_i - \bar{y}, n\bar{y} - \sum_j y_j \rangle \\ &= 2n \sum_i \|y_i - \bar{y}\|_2^2.\end{aligned}$$

Thus, we have

$$\operatorname*{argmin}_{\{C_1,\cdots,C_K\}} \sum_{k=1}^{K} \sum_{x \in C_k} \left\| x - \frac{\sum_{x \in C_k} x}{|C_k|} \right\|_2^2 = \operatorname*{argmin}_{\{C_1,\cdots,C_K\}} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{x,y \in C_k} \|x - y\|_2^2,$$

i.e. k-means clustering algorithm is equivalent to minimizing

$$F(\{C_1, \cdots, C_K\}) := \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{x,y \in C_k} \|x - y\|_2^2.$$

Suppose there exists a partition $C_1^*, \cdots, C_K^*$ satisfying $\max_{k\in[K]} \max_{i,j\in C_k^*} \|y_i - y_j\|_2 \le a$ and $\min_{k\in[K]} \min_{i\in C_k^*,\ j\in C_{\tilde{k}}^*} \|y_i - y_j\|_2 \ge b$ with $b > a\sqrt{n}$. Then, obviously,

$$F(\{C_1^*, \cdots, C_K^*\}) \le \sum_k a^2 \frac{|C_k^*| - 1}{2} = (n-K)a^2/2.$$

For any other partitions $C_1, \cdots, C_K$, there exists at least one pair of (y_i, y_j) in some C_k satisfying $\|y_i - y_j\| \ge b$. Let $C_k^1, \cdots, C_k^K$ be the partition of such C_k such that $C_k^i = C_i^* \cap C_k$ with at least two C_k^i are non-empty. Then,

$$F(\{C_1, \cdots, C_K\}) \ge \frac{\sum_{l \ne u} b^2 |C_k^l||C_k^u|}{|C_k|} \ge b^2 > (n-K)a^2/2 = F(\{C_1^*, \cdots, C_K^*\}),$$

which implies $\{C_1^*, \cdots, C_K^*\}$ are the outcome of k-means clustering algorithm. This completes the proof. □

Lemma S5. *Suppose conditions of Theorem S1. For fixed cluster C_k, let $d_1^*, \cdots, d_{n_k}^*$ be the degrees of n_k samples based on the matrix S^* presented in the proof of Theorem S1. Then, with probability $1-2n^{-2}$, $g\exp(-10\sigma^{-2}) \leq d_i^* \leq 2g\exp(-0.1\sigma^{-2})$ for all $i \in C_k$.*

Proof of Lemma S5. We fix sample i with $i \in C_k$. Note that since $S_{ij}^* = 0$ for $j \notin C_k$, we have $d_i^* = \sum_{j\in C_k} S_{ij}^*$. We see $S_{ij}^* = \exp\left(-\frac{\|x_i-x_j\|^2}{2\epsilon_{ij}^2}\right) = \exp\left(-\frac{\|z_i-z_j\|^2}{2\epsilon_{ij}^2}\right)$ for at most $2g$ number of $j \in C_k$, where $\epsilon_{ij} = \frac{\sigma(\mu_i+\mu_j)}{2}$ and $\mu_i = \frac{\sum_{j\in N_g(i)} \|x_i-x_j\|}{g}$.

Notice that $z^{ij} := z_i - z_j$ is a random vector with independent component z_l^{ij} which satisfying $E[z_l^{ij}] = 0$ and $\|z_l^{ij}\|_{\psi_2} \leq 2\sigma_z\psi$. Hence, by Theorem 1.1 of Rudelson and Vershynin (2013), we have for any $t \geq 0$,

$$P\left[\left|\|z^{ij}\|_2^2 - E[\|z^{ij}\|_2^2]\right| > t\right] \leq 2\exp\left[-c\ \min\left(\frac{t^2}{16\sigma_z^4\psi^4}, \frac{t}{4\sigma_z^2\psi^2}\right)\right]$$

for some absolute constant $c > 0$. Since $E[\|z^{ij}\|_2^2] = 2p\sigma_z^2$, with $t = 16\sigma_z^2\psi^2\log n/c$, we have with probability at least $1-2n^{-4}$, $|\|z^{ij}\|_2^2 - 2p\sigma_z^2| \leq 16\sigma_z^2\psi^2\log n/c$, i.e.,

$$\left|\|z^{ij}\|_2 - \sqrt{2p}\sigma_z\right| \leq 4\sigma_z\psi\sqrt{\log n/c}/(\sqrt{2p}\sigma_z) = 4\psi\sqrt{\frac{\log n}{2pc}}.$$

Combined with $4\psi\sqrt{\frac{\log n}{2pc}} \leq 0.5\sqrt{2p}\sigma_z$, we have $0.5\sqrt{2p}\sigma_z \leq \|z^{ij}\|_2 \leq 1.5\sqrt{2p}\sigma_z$. By the union bound, we have with probability $1-2n^{-3}$,

$$\left|\mu_i - \sqrt{2p}\sigma_z\right| \leq \sum_{j\in N_g(i)} \frac{1}{g}\left|\|z^{ij}\|_2 - \sqrt{2p}\sigma_z\right| \leq 4\psi\sqrt{\frac{\log n}{2pc}},$$

i.e. $0.5\sqrt{2p}\sigma_z \leq \mu_i \leq 1.5\sqrt{2p}\sigma_z$. Hence, with probability $1-2n^{-2}$, we have $0.5\sqrt{2p}\sigma_z \leq \mu_i \leq 1.5\sqrt{2p}\sigma_z$ for all i. Thus, for any i and j, we have

$0.5\sigma\sigma_z\sqrt{2p} \le \epsilon_{ij} \le 1.5\sigma\sigma_z\sqrt{2p}$. Hence, for i and j such that $i \in N_g(j)$ or $j \in N_g(i)$, with probability $1 - 2n^{-2}$,

$$\begin{aligned}\exp(-10\sigma^{-2}) &\le \exp\left(-\frac{1.5^2 2p\sigma_z^2}{0.5^2\sigma^2\sigma_z^2 2p}\right) \le S^*_{ij} = \exp\left(-\frac{\|z^i - z^j\|^2}{2\epsilon_{ij}^2}\right) \\ &\le \exp\left(-\frac{0.5^2 2p\sigma_z^2}{1.5^2\sigma^2\sigma_z^2 2p}\right) \le \exp(-0.1\sigma^{-2}),\end{aligned}$$

thus we have $g\exp(-10\sigma^{-2}) \le d_i^* \le 2g\exp(-0.1\sigma^{-2})$. This completes the proof. □

Lemma S6. *Suppose conditions of Theorem S1. Then, with probability at least $1 - 2(C^*)^2/n$, $N_g(i)$ consists of the samples from the same cluster of i, provided that $g < \min_{i\in[K]} n_i$.*

Proof of Lemma S6. We calculate the probability that for any sample x_i belonging to cluster $\tilde{k}$, its nearest g-neighborhoods also comes from the same cluster $\tilde{k}$ with high probability. Let E_i be the event that some of the nearest g-neighbors of x_i are from other clusters. Then, we have

$$\begin{aligned}P[E_i] &= P[\cup_{j\in C^c_{\tilde{k}}}\{j \in N_g(i)\}] \\ &\le \sum_{j\in C^c_{\tilde{k}}} P[j \in N_g(i)] \\ &= \sum_{k\neq\tilde{k}}\sum_{j\in C_k} P[j \in N_g(i)] \\ &= \sum_{k\neq\tilde{k}} n_k P[j \in N_g(i) \mid j \in C_k] \\ &\le \sum_{k\neq\tilde{k}} n_k n_{\tilde{k}} P\left[\|x_i - x_j\|_2 < \|x_i - x_l\|_2 \mid j \in C_k, l \in C_{\tilde{k}}\right], \quad \text{(S28)}\end{aligned}$$

where the last inequality follows from the inequality

$$\{j \in N_g(i)\} \subset \cup_{l\in C_{\tilde{k}}}\{\|x_i - x_j\|_2 < \|x_i - x_l\|_2\}$$

because of $g < n_{\tilde{k}}$, and the union bound. Note that since $x_i = \mu^{(\tilde{k})} + z_i$, $x_l = \mu^{(\tilde{k})} + z_l$, and $x_j = \mu^{(k)} + z_j$, we have

$$\begin{aligned} \|x_i - x_j\|_2 &= \|\mu^{(\tilde{k})} - \mu^{(k)} + z_i - z_j\|_2 \geq \|\mu^{(\tilde{k})} - \mu^{(k)}\|_2 - \|z_i - z_j\|_2 \\ \|x_i - x_l\|_2 &= \|z_i - z_l\|_2, \end{aligned}$$

thus we have

$$\begin{aligned} & P[\|x_i - x_j\|_2 < \|x_i - x_l\|_2 \mid j \in C_k, l \in C_{\tilde{k}}] \\ \leq \;& P[\mu^{(\tilde{k})} - \mu^{(k)}\|_2 < \|z_i - z_l\|_2 + \|z_i - z_j\|_2] \\ \leq \;& P[\mu^{(\tilde{k})} - \mu^{(k)}\|_2 < 2\|z_i - z_l\|_2] + P[\mu^{(\tilde{k})} - \mu^{(k)}\|_2 < 2\|z_i - z_j\|_2] \\ = \;& 2P[\|z_i - z_j\|_2 > \|\mu^{(\tilde{k})} - \mu^{(k)}\|_2/2]. \end{aligned}$$

Notice that $z^{ij} := z_i - z_j$ is a random vector with independent component z^{ij}_l satisfying $E[z^{ij}_l] = 0$ and $\|z^{ij}_l\|_{\psi_2} \leq 2\sigma_z\psi$. Hence, by Theorem 1.1 of Rudelson and Vershynin (2013), we have for any $t \geq 0$,

$$P\left[\left|\|z^{ij}\|_2^2 - E[\|z^{ij}\|_2^2]\right| > t\right] \leq 2\exp\left[-c \;\min\left(\frac{t^2}{16\sigma_z^4\psi^4}, \frac{t}{4\sigma_z^2\psi^2}\right)\right]$$

for some absolute constant $c > 0$. Since $E[\|z^{ij}\|_2^2] = 2p\sigma_z^2$, by letting $t = 16\sigma_z^2\psi^2\log n/c$, we have with probability at least $1 - 2n^{-4}$,

$$\|z^{ij}\|_2^2 \leq 2p\sigma_z^2 + 16\sigma_z^2\psi^2\log n/c \leq \min_{k\neq\tilde{k}} \|\mu^{(\tilde{k})} - \mu^{(k)}\|_2^2/4,$$

where we use the condition $\min_{k\neq\tilde{k}} \|\mu^{(\tilde{k})} - \mu^{(k)}\|_2^2 \geq 8p\sigma_z^2 + 64\sigma_z^2\psi^2\log n/c$. Combining with (S28), we have

$$P[E_i] \;\leq\; \sum_{k\neq\tilde{k}} n_k n_{\tilde{k}}(2n^{-4}) \leq 2(C^*)^2 n^{-2}.$$

Thus, $P[\cup_i E_i] \leq 2(C^*)^2/n$, i.e., the nearest g-neighborhoods also comes from the same cluster with probability at least $1 - 2(C^*)^2/n$. This completes the proof. □

Lemma S7. *Let $G_{ml} = D_{ml}^{-1/2} S_{ml} D_{ml}^{-1/2}$ for $m = 1, \cdots, M$ and $l = 1, \cdots, 55$ be the normalized similarity matrix as defined in Section 2.2 of the main paper. Then $\|G_{ml}\|_F \leq 1$.*

Proof. For fixed l and m, let $S_{ml} = (S_{ij})$ and $D_{ml} = \text{diag}(d_1, \cdots, d_n)$. We have

$$\|G_{ml}\|_F = \sqrt{\text{tr}(D_{ml}^{-1} S_{ml} D_{ml}^{-1} S_{ml})} = \sqrt{\sum_{ij} \frac{S_{ij}^2}{d_i d_j}}.$$

Since $|S_{ij}| \leq 1$ by the definition as in Section 2.2, we have

$$\left(\sum_{ij} \frac{S_{ij}^2}{d_i d_j}\right)\left(\sum_{ij} d_i d_j\right) \leq (\sum_{ij} S_{ij}^2)^2 \leq (\sum_{ij} S_{ij})^2 = \sum_{ij} d_i d_j,$$

where the last equality follows from $d_i = \sum_j S_{ij}$. Combining the above two inequalities, we have $\|G_{ml}\|_F \leq 1$. This completes the proof. □

Lemma S8. *For $m = 1, \cdots, M$, let*

$$\hat{d}_m^{(i)} = \exp\left(\frac{1}{\rho}\left(-\epsilon\|\hat{P}_m^{(i-1)}\|_F^2 + \langle \hat{S}_m^{(i-1)}, \hat{P}_m^{(i-1)}\rangle\right)\right), \quad \hat{S}_m^{(i)} = \sum_l \hat{w}_{ml}^{(i)} G_{ml},$$

$$\bar{d}_m^{(i)} = \exp\left(\frac{1}{\rho}\left(-\epsilon\|\bar{P}_m^{(i-1)}\|_F^2 + \langle \bar{S}_m^{(i-1)}, \bar{P}_m^{(i-1)}\rangle\right)\right), \quad \bar{S}_m^{(i)} = \sum_l \bar{w}_{ml}^{(i)} G_{ml},$$

where G_{ml} is defined in Lemma S7. Then, we have

$$\begin{aligned} \exp(-(\epsilon C^* + \sqrt{C^*})/\rho) &\leq |\hat{d}_m^{(i)}| \leq \exp(\sqrt{C^*}/\rho) \\ \exp(-(\epsilon C^* + \sqrt{C^*})/\rho) &\leq |\bar{d}_m^{(i)}| \leq \exp(\sqrt{C^*}/\rho) \\ |\bar{d}_m^{(i)} - \hat{d}_m^{(i)}| &\lesssim \|\hat{P}_m^{(i-1)} - \bar{P}_m^{(i-1)}\|_F + \|\hat{w}_m^{(i-1)} - \bar{w}_m^{(i-1)}\|_F. \end{aligned}$$

Proof. First, since $0 \preceq \bar{P}_m^{(i)} \preceq I$, we have

$$\|\bar{P}_m^{(i)}\|_F = \sqrt{\mathrm{tr}((\bar{P}_m^{(i)})^2)} \leq \sqrt{\mathrm{tr}(\bar{P}_m^{(i)})} = \sqrt{C^*}.$$

Similarly, $\|\hat{P}_m^{(i)}\|_F \leq \sqrt{C^*}$. Hence, it holds that

$$|\|\hat{P}_m^{(i)}\|_F^2 - \|\bar{P}_m^{(i)}\|_F^2| \leq (\|\hat{P}_m^{(i)}\|_F + \|\bar{P}_m^{(i)}\|_F)\|\hat{P}_m^{(i)} - \bar{P}_m^{(i)}\|_F \leq 2\sqrt{C^*}\|\hat{P}_m^{(i)} - \bar{P}_m^{(i)}\|_F.$$

By Lemma S7, we have

$$\|\hat{S}_m^{(i)}\|_F \leq \max_l \|G_{ml}\|_F \leq 1.$$

Hence, we have

$$\begin{aligned}
\|\hat{S}_m^{(i)} - \bar{S}_m^{(i)}\|_F &= \|\sum_l \hat{w}_{ml}^{(i)} G_{ml} - \sum_l \bar{w}_{ml}^{(i)} G_{ml}\|_F \\
&\leq 55 \max_l |\hat{w}_{ml}^{(i)} - \bar{w}_{ml}^{(i)}| \max_l \|G_{ml}\|_F \\
&\leq 55 \max_l |\hat{w}_{ml}^{(i)} - \bar{w}_{ml}^{(i)}|.
\end{aligned}$$

Hence it holds that

$$\begin{aligned}
|\langle \hat{S}_m^{(i)}, \hat{P}_m^{(i)} \rangle - \langle \bar{S}_m^{(i)}, \bar{P}_m^{(i)} \rangle| &\leq |\langle \hat{S}_m^{(i)}, \hat{P}_m^{(i)} - \bar{P}_m^{(i)} \rangle| + |\langle \hat{S}_m^{(i)} - \bar{S}_m^{(i)}, \bar{P}_m^{(i)} \rangle| \\
&\leq |\langle \hat{S}_m^{(i)}, \hat{P}_m^{(i)} - \bar{P}_m^{(i)} \rangle| + |\langle \hat{S}_m^{(i)} - \bar{S}_m^{(i)}, \bar{P}_m^{(i)} \rangle| \\
&\leq \|\hat{P}_m^{(i)} - \bar{P}_m^{(i)}\|_F + \sqrt{C^*}\|\hat{S}_m^{(i)} - \bar{S}_m^{(i)}\|_F \\
&\leq \|\hat{P}_m^{(i)} - \bar{P}_m^{(i)}\|_F + 55\sqrt{C^*} \max_l |\hat{w}_{ml}^{(i)} - \bar{w}_{ml}^{(i)}|.
\end{aligned}$$

Combining the above results, we have

$$\begin{aligned}
|\log(\bar{d}_m^{(i)}) - \log(\hat{d}_m^{(i)})| &\leq \frac{1}{\rho}\left(\epsilon \left|\|\hat{P}_m^{(i)}\|_F^2 - \|\bar{P}_m^{(i)}\|_F^2\right| + \left|\langle \hat{S}_m^{(i)}, \hat{P}_m^{(i)} \rangle - \langle \bar{S}_m^{(i)}, \bar{P}_m^{(i)} \rangle\right|\right) \\
&\leq \frac{\epsilon}{\rho}\left|\|\hat{P}_m^{(i)}\|_F^2 - \|\bar{P}_m^{(i)}\|_F^2\right| + \frac{1}{\rho}\left|\langle \hat{S}_m^{(i)}, \hat{P}_m^{(i)} \rangle - \langle \bar{S}_m^{(i)}, \bar{P}_m^{(i)} \rangle\right| \\
&\leq \frac{2\sqrt{C^*}\epsilon}{\rho}\|\hat{P}_m^{(i)} - \bar{P}_m^{(i)}\|_F + \frac{1}{\rho}\left(\|\hat{P}_m^{(i)} - \bar{P}_m^{(i)}\|_F + 55\sqrt{C^*}\max_l |\hat{w}_{ml}^{(i)} - \bar{w}_{ml}^{(i)}|\right) \\
&:= x.
\end{aligned}$$

Since

$$|\langle \hat{S}_m^{(i-1)}, \hat{P}_m^{(i-1)}\rangle| \le \|\hat{S}_m^{(i-1)}\|_F \|\hat{P}_m^{(i-1)}\|_F \le \sqrt{C^*}, \quad |\langle \bar{S}_m^{(i-1)}, \bar{P}_m^{(i-1)}\rangle| \le \|\bar{S}_m^{(i-1)}\|_F \|\bar{P}_m^{(i-1)}\|_F \le \sqrt{C^*},$$

it holds that

$$\begin{aligned} \exp(-(\epsilon C^* + \sqrt{C^*})/\rho) &\le |\hat{d}_m^{(i)}| \le \exp(\sqrt{C^*}/\rho) \\ \exp(-(\epsilon C^* + \sqrt{C^*})/\rho) &\le |\bar{d}_m^{(i)}| \le \exp(\sqrt{C^*}/\rho). \end{aligned}$$

Hence, by the mean value theorem, we have

$$|\bar{d}_m^{(i)} - \hat{d}_m^{(i)}| \le \exp(\sqrt{C^*}/\rho)\ |\log(\bar{d}_m^{(i)}) - \log(\hat{d}_m^{(i)})| \le x \exp(\sqrt{C^*}/\rho).$$

This completes the proof. □

Lemma S9. *For fixed $l \in \{1, \cdots, \ell\}$ and $m \in \{1, \cdots, M\}$, let*

$$\hat{w}_{ml}^{(i)} = \frac{\exp\left(\frac{\hat{c}_m^{(i)}}{\rho}\langle G_{ml}, \hat{P}_m^{(i-1)}\rangle\right)}{\sum_{\widetilde{m}} \exp\left(\frac{\hat{c}_{\widetilde{m}}^{(i)}}{\rho}\langle G_{\widetilde{m}l}, \hat{P}_{\widetilde{m}}^{(i-1)}\rangle\right)}, \quad \bar{w}_{ml}^{(i)} = \frac{\exp\left(\frac{\bar{c}_m^{(i)}}{\rho}\langle G_{ml}, \bar{P}_m^{(i-1)}\rangle\right)}{\sum_{\widetilde{m}} \exp\left(\frac{\bar{c}_{\widetilde{m}}^{(i)}}{\rho}\langle G_{\widetilde{m}l}, \bar{P}_{\widetilde{m}}^{(i-1)}\rangle\right)},$$

where $\rho, \hat{c}_m^{(i)}, \bar{c}_m^{(i)} > 0$, G_{ml} is defined in Lemma S7, and $\hat{P}_m^{(i-1)}$ and $\bar{P}_m^{(i-1)}$ satisfy the constraint (S5) *with $C = C^*$. Then, we have*

$$|\hat{w}_{ml}^{(i)} - \bar{w}_{ml}^{(i)}| \lesssim \|\hat{P}_m^{(i-1)} - \bar{P}_m^{(i-1)}\|_F + \|\hat{c}^{(i)} - \bar{c}^{(i)}\|.$$

Proof. We can easily show that

$$55 \exp\left(-\frac{\sqrt{C^*}}{\rho}\right) \le \sum_{\widetilde{m}} \exp\left(\frac{\hat{c}_{\widetilde{m}}^{(i)}}{\rho}\langle G_{\widetilde{m}l}, \hat{P}_{\widetilde{m}}^{i-1}\rangle\right) \le 55 \exp\left(\frac{\sqrt{C^*}}{\rho}\right).$$

Since $\|G_{ml}\|_F \le 1$, $|\hat{c}_{\widetilde{m}}^{(i)}| \le 1$, and $\langle G_{ml}, \bar{P}_m^{(i-1)}\rangle \le \sqrt{C^*}$, we have

$$\begin{aligned} \left|\frac{\hat{c}_m^{(i)}}{\rho}\langle G_{ml}, \hat{P}_m^{(i-1)}\rangle - \frac{\bar{c}_m^{(i)}}{\rho}\langle G_{ml}, \bar{P}_m^{(i-1)}\rangle\right| &\le \frac{1}{\rho}\left\|\hat{c}_{\widetilde{m}}^{(i)}\hat{P}_m^{(i-1)} - \bar{c}_{\widetilde{m}}^{(i)}\bar{P}_m^{(i-1)}\right\|_F \\ &\le \frac{1}{\rho}\|\hat{P}_m^{(i-1)} - \bar{P}_m^{(i-1)}\|_F + \frac{\sqrt{C^*}}{\rho}|\hat{c}_m^{(i)} - \bar{c}_m^{(i)}|. \end{aligned}$$

Since $\left|\frac{\hat{c}_{\widetilde{m}}^{(i)}}{\rho}\langle G_{\widetilde{m}l}, \hat{P}_{\widetilde{m}}^{(i-1)}\rangle\right| \le \frac{\sqrt{C^*}}{\rho}$, we have

$$\begin{aligned}
& \left|\exp\left(\frac{\hat{c}_m^{(i)}}{\rho}\langle G_{ml}, \hat{P}_m^{(i-1)}\rangle\right) - \exp\left(\frac{\bar{c}_m^{(i)}}{\rho}\langle G_{ml}, \bar{P}_m^{(i-1)}\rangle\right)\right| \\
\le \quad & 55\exp\left(\frac{\sqrt{C^*}}{\rho}\right)\left|\frac{\hat{c}_m^{(i)}}{\rho}\langle G_{ml}, \hat{P}_m^{(i-1)}\rangle - \frac{\bar{c}_m^{(i)}}{\rho}\langle G_{ml}, \bar{P}_m^{(i-1)}\rangle\right| \\
\le \quad & 55\exp\left(\frac{\sqrt{C^*}}{\rho}\right)\left(\frac{1}{\rho}\|\hat{P}_m^{(i-1)} - \bar{P}_m^{(i-1)}\|_F + \frac{\sqrt{C^*}}{\rho}|\hat{c}_m^{(i)} - \bar{c}_m^{(i)}|\right) \\
:= \quad & y.
\end{aligned}$$

Note that we have

$$\begin{aligned}
& \exp\left(\frac{\hat{c}_m^{(i)}}{\rho}\langle G_{ml}, \hat{P}_m^{(i-1)}\rangle\right)\left|\sum_{\widetilde{m}}\exp\left(\frac{\hat{c}_{\widetilde{m}}^{(i)}}{\rho}\langle G_{\widetilde{m}l}, \hat{P}_{\widetilde{m}}^{(i-1)}\rangle\right) - \sum_{\widetilde{m}}\exp\left(\frac{\bar{c}_{\widetilde{m}}^{(i)}}{\rho}\langle G_{\widetilde{m}l}, \bar{P}_{\widetilde{m}}^{(i-1)}\rangle\right)\right| \\
\le \quad & 3025\exp\left(\frac{2\sqrt{C^*}}{\rho}\right)\left(\frac{1}{\rho}\|\hat{P}_m^{(i-1)} - \bar{P}_m^{(i-1)}\|_F + \frac{\sqrt{C^*}}{\rho}\|\hat{c}^{(i)} - \bar{c}^{(i)}\|\right).
\end{aligned}$$

Combining the above inequalities, we have

$$\begin{aligned}
|\hat{w}_{ml}^{(i)} - \bar{w}_{ml}^{(i)}| \quad & \le \quad \frac{y}{3025\exp\left(-\frac{\sqrt{C^*}}{\rho}\right)} + \frac{3025\exp\left(\frac{\sqrt{C^*}}{\rho}\right)y}{3025\exp\left(-\frac{2\sqrt{C^*}}{\rho}\right)} \\
& \le \quad 2\exp\left(\frac{3\sqrt{C^*}}{\rho}\right)y.
\end{aligned}$$

This completes the proof. □

Lemma S10. *For any n by n matrices $\hat{T}_1$ and $\bar{T}_1$, let*

$$\begin{aligned}
\hat{Q}_m^{i,t} &= \operatorname*{argmin}_Q \|Q - \hat{T}_1\|_F^2 \quad \text{s.t.} \quad \operatorname{tr}(Q) = C,\ 0 \preceq Q \preceq I, \\
\bar{Q}_m^{i,t} &= \operatorname*{argmin}_Q \|Q - \bar{T}_1\|_F^2 \quad \text{s.t.} \quad \operatorname{tr}(Q) = C,\ 0 \preceq Q \preceq I.
\end{aligned}$$

Then, we have

$$\|\hat{Q}_m^{i,t} - \bar{Q}_m^{i,t}\|_F \le \|\hat{T}_1 - \bar{T}_1\|_F.$$

Proof. Let $\mathrm{mat}(\cdot)$ be the inverse of the vectorization operator such that $\mathrm{mat}(\mathrm{vec}(Q)) = Q$ for any n by n matrix Q. Define the set

$$E = \{q \in \mathbb{R}^{n^2} \mid \mathrm{tr}(\mathrm{mat}(q)) = C,\ 0 \preceq \mathrm{mat}(q) \preceq I\},$$

which is a nonempty closed convex set of $\mathbb{R}^{n^2}$. Note that $\mathrm{vec}(\hat{Q}_m^{i,t})$ and $\mathrm{vec}(\bar{Q}_m^{i,t})$ are the projections of $\mathrm{vec}(\hat{T}_1)$ and $\mathrm{vec}(\bar{T}_1)$ on the set E, respectively. Hence by the Projection theorem (Bertsekas, 2009), it holds that

$$(\mathrm{vec}(\hat{T}_1) - \mathrm{vec}(\hat{Q}_m^{i,t}))^T(\mathrm{vec}(\bar{Q}_m^{i,t}) - \mathrm{vec}(\hat{Q}_m^{i,t})) \leq 0,$$
$$(\mathrm{vec}(\bar{T}_1) - \mathrm{vec}(\bar{Q}_m^{i,t}))^T(\mathrm{vec}(\hat{Q}_m^{i,t}) - \mathrm{vec}(\bar{Q}_m^{i,t})) \leq 0.$$

This implies

$$\|\mathrm{vec}(\hat{Q}_m^{i,t}) - \mathrm{vec}(\bar{Q}_m^{i,t})\|^2 \leq (\mathrm{vec}(\bar{Q}_m^{i,t}) - \mathrm{vec}(\hat{Q}_m^{i,t}))^T(\mathrm{vec}(\bar{T}_1) - \mathrm{vec}(\hat{T}_1)),$$

thus $\|\mathrm{vec}(\hat{Q}_m^{i,t}) - \mathrm{vec}(\bar{Q}_m^{i,t})\| \leq \|\mathrm{vec}(\bar{T}_1) - \mathrm{vec}(\hat{T}_1)\|$. Hence we have

$$\|\hat{Q}_m^{i,t} - \bar{Q}_m^{i,t}\|_F \leq \|\hat{T}_1 - \bar{T}_1\|_F.$$

This completes the proof. □

Lemma S11. *Suppose conditions of Theorem 1. Then,* $\hat{c}_m \geq M\exp(-3C^*)$ *for all* $m = 1, \cdots, M$.

Proof. Note that $\{\hat{c}_m\}$ is the solution to the following optimization problem:

$$\begin{aligned} \min_{\{c_m\}} \quad & \epsilon \sum_m c_m \|\hat{P}_m\|_F^2 - \sum_m c_m \langle S_m, \hat{P}_m \rangle + \lambda \sum_m c_m \|\hat{P}_m\|_1 + \rho \sum_m c_m \log c_m \\ \text{s.t.} \quad & \sum_m c_m = 1. \end{aligned}$$

By the method of Lagrange multipliers, $\{\hat{c}_m\}$ is the solution to the following unconstrained problem:

$$\min_{\{c_m\}} \epsilon \sum_m c_m \|\hat{P}_m\|_F^2 - \sum_m c_m \langle S_m, \hat{P}_m \rangle + \lambda \sum_m c_m \|\hat{P}_m\|_1 + \rho \sum_m c_m \log c_m + \tilde{\lambda} \left(\sum_m c_m - M \right)$$

for some Lagrange multiplier $\tilde{\lambda}$. The derivative of the above function with respect to c_m equal to zero gives the solution

$$\hat{c}_m = \exp\left(\frac{-\epsilon \|\hat{P}_m\|_F^2 - \langle S_m, \hat{P}_m \rangle - \lambda \|\hat{P}_m\|_1 - \tilde{\lambda} - \rho}{\rho} \right).$$

Since it holds that $\|\hat{P}_m\|_F^2 \le C^*$, $\lambda \|\hat{P}_m\|_1 \le C^* n^{-2} n C^* = (C^*)^2/n$, and

$$|\langle S_m, \hat{P}_m \rangle| \le \|S_m\|_F \|\hat{P}_m\|_F \le \|\hat{P}_m\|_F \max_{m,l} \|G_{m,l}\|_F \le C^*$$

due to Lemma S7, we have $\hat{c}_m / \hat{c}_{m'} \ge \exp(-3C^*/\rho)$ for any $m, m' \in \{1, \cdots, M\}$. Since $\sum_m \hat{c}_m = M$, this implies $\hat{c}_m \ge M \exp(-3C^*)$ as $\rho = 1$. This completes the proof. □

E Time complexity of the algorithm

The computational complexity of the algorithm is $O(Kn^3)$, where n is the number of data points and K is the number of iterations. In the experiments, K is less than 30. Note that the traditional spectral clustering algorithm has the complexity $O(n^3)$. The proposed algorithm is still fast for the cancer data set because n is quite small compared with the number of variables (genes).

F Evaluation metrics

We use the following three performance metrics to evaluate the consistency between the obtained clustering and the true labels: Normalized Mutual Information (NMI) (Strehl and Ghosh, 2003), Purity (Wagner and Wagner, 2007), and Adjusted Rand Index (ARI) (Wagner and Wagner, 2007). Given two clustering results U and V on a set of n data points with C_U and C_V clusters, respectively, the mutual information is NMI is defined as

$$\mathrm{NMI}(U,V) = \frac{\sum_{p=1}^{C_U}\sum_{q=1}^{C_V} |U_p \cap V_q| \log \frac{n|U_p \cap V_q|}{|U_p| \times |V_q|}}{\max\left(-\sum_{p=1}^{C_U} |U_p| \log \frac{|U_p|}{n}, -\sum_{q=1}^{C_V} |V_q| \log \frac{|V_q|}{n}\right)},$$

where U_p and V_q are the index sets of the pth and qth clusters of clustering results U and V, for $p = 1, \cdots, C_U$ and $q = 1, \cdots, C_V$, respectively. Here, the numerator is the mutual information between U and V, and the denominator represents the entropy of the clustering U and V. For Purity, each identified cluster is assigned to the one which is most frequent in the cluster, and then the accuracy of this assignment is computed by counting the number of correctly assigned samples divided by the number n:

$$\mathrm{Purity}(U,V) = \frac{\sum_p \max_q |U_p \cap V_q|}{n}.$$

The value of ARI depends on the following four quantities: a_{uv}, the number of objects in a pair that are placed in the same group in U and V; a_u, the number of objects in a pair that are placed in the same group in U but in different groups in V; a_v, the number of objects in a pair that are placed in the same group in V but in different groups in U; and a, the number of objects in a pair that are placed in the different groups in U and V. ARI is defined as

$$\mathrm{ARI}(U,V)=\frac{\binom{n}{2}(a_{uv}+a)-[(a_{uv}+a_u)(a_{uv}+a_v)+(a_v+a)(a_u+a)]}{\binom{n}{2}-[(a_{uv}+a_u)(a_{uv}+a_v)+(a_v+a)(a_u+a)]}.$$

Note that the NMI and Purity take on values between 0 and 1, but ARI can yield negative values. These metrics measure the concordance of two clustering results such that higher value refers to higher concordance with true labels.

G Data

We collect 22 major cancer types with sufficient patients from the TCGA project, where three molecular profiles are used: RNA expression (RNA-seq V2), miRNA, and copy number alterations (CNA). To reduce platform differences, we only consider each molecular profile from one platform. We consider the RNA data sets measured using the IIllumina sequencing technology with the $\log_2(x+1)$ transformed RSEM (RNA-Seq by Expectation Maximization) values, the miRNA mature strand expression RNAseq data sets measured by Illumina miRNA-seq, and the CNA estimated using the GISTIC2 threshold method and compiled using data from all TCGA cohorts (Weinstein *et al.*, 2013). Specifically, CNA have values in $\{-2,-1,0,1,2\}$. They are obtained by applying both low-level and high-level thresholds to the gene copy levels of all the samples. The value 2 or -2 means that corresponding gene copy levels exceed the high-level thresholds for amplification or deletion, respectively, and those with 1 or -1 means that corresponding gene copy levels exceed the low-level thresholds for amplification or deletion but not the high-level thresholds, respectively (Mermel *et al.*, 2011).

We remove patients with missing molecular profiles with 6,976 patients included in the clustering and survival analysis as follows:

- Bladder Cancer (BLCA), n=400, C=3.
- Breast Invasive Carcinoma (BRCA), n=741, C=4.
- Cervical Squamous Cell Carcinoma (CESC), n=289, C=3.
- Colon Adenocarcinoma (COAD), n=249, C=4.
- Esophageal Cancer (ESCA), n=180, C=5.
- Head and Neck Squamous Cell carcinoma (HNSC), n=471, C=6.
- Kidney Renal Clear cell carcinoma (KIRC), n=236, C=3.
- Kidney Papillary Cell Carcinoma (KIRP), n=283, C=3.
- Lower Grade Glioma (LGG), n=505, C=4.
- Liver Cancer (LIHC), n=357, C=4.
- Lung Adenocarcinoma (LUAD), n=442, C=5.
- Lung Squamous Cell carcinoma (LUSC), n=333, C=3.
- Mesothelioma (MESO), n=87, C=3.
- Ovarian Cancer (OV), n=297, C=4.
- Pancreatic Cancer (PAAD), n=176, C=4.
- Prostate Adenocarcinoma (PRAD), n=484, C=3.

- Rectum Adenocarcinoma (READ), n=87, C=3.
- Sarcoma (SARC), n=250, C=4.
- Stomach Adenocarcinoma (STAD), n=365, C=4.
- Thyroid Cancer (THCA), n=494, C=4.
- Uterine Corpus Endometrioid Carcinoma (UCEC), n=170, C=4.
- Uveal Melanoma (UVM), n=80. C=3.

H Additional figures

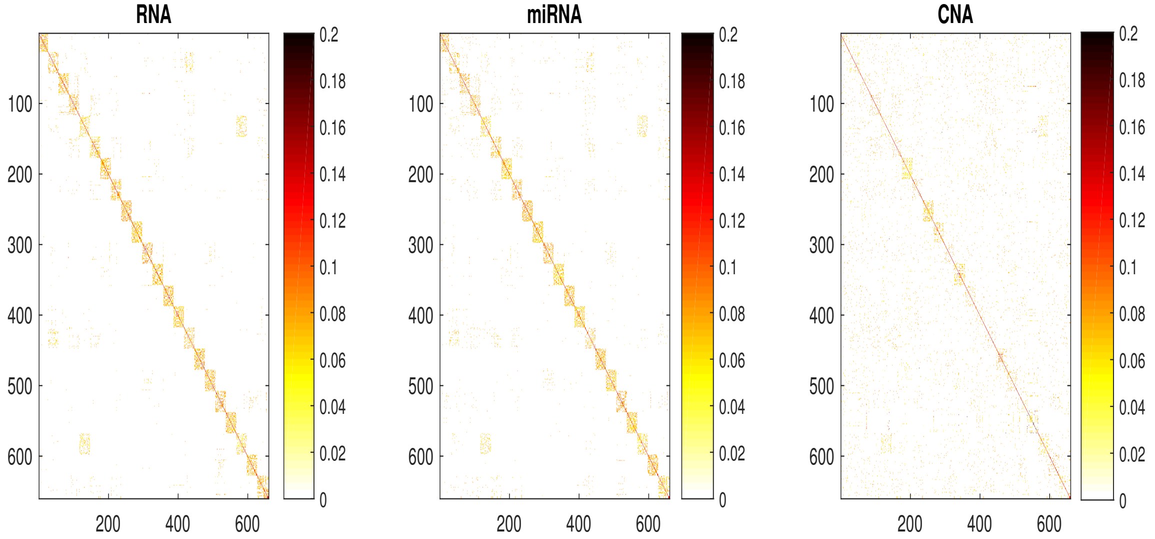

Figure S1: Heatmaps of selected similarity matrices for the three data sets. The patients are ordered based on their cancer types.

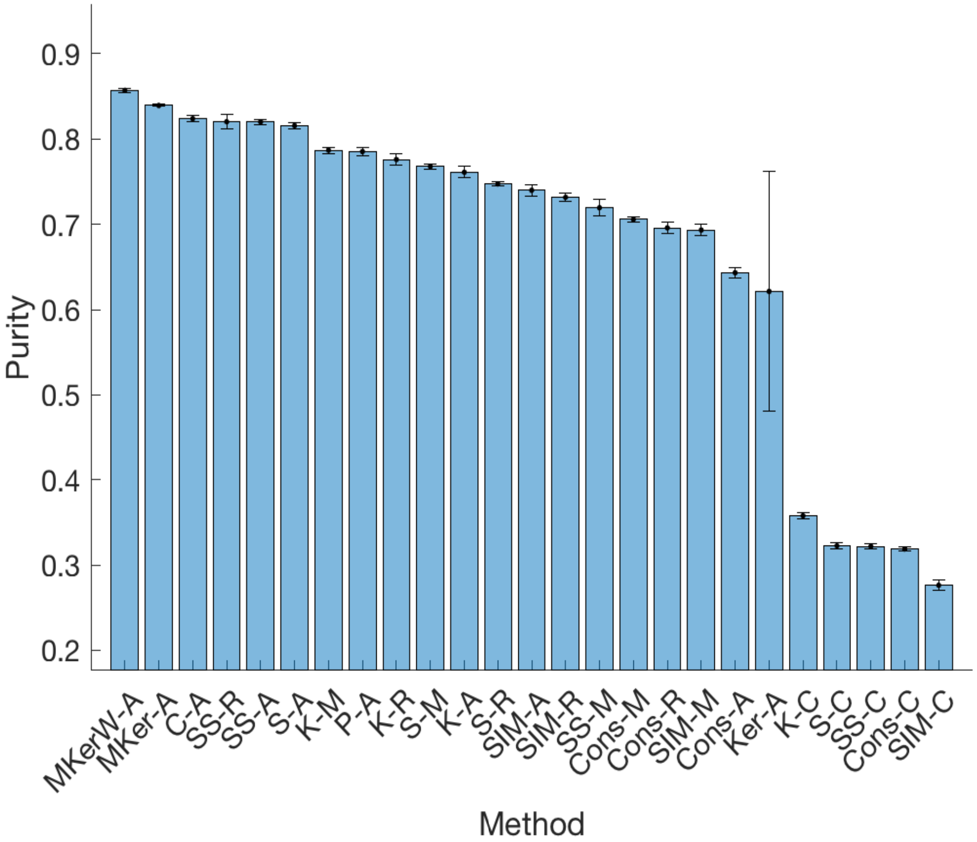

Figure S2: The average Purity and one standard deviation of 50 replicates for the 25 clustering methods when thirty patients were randomly selected from each of the 22 cancer types. The methods are ordered according to the Purity values.

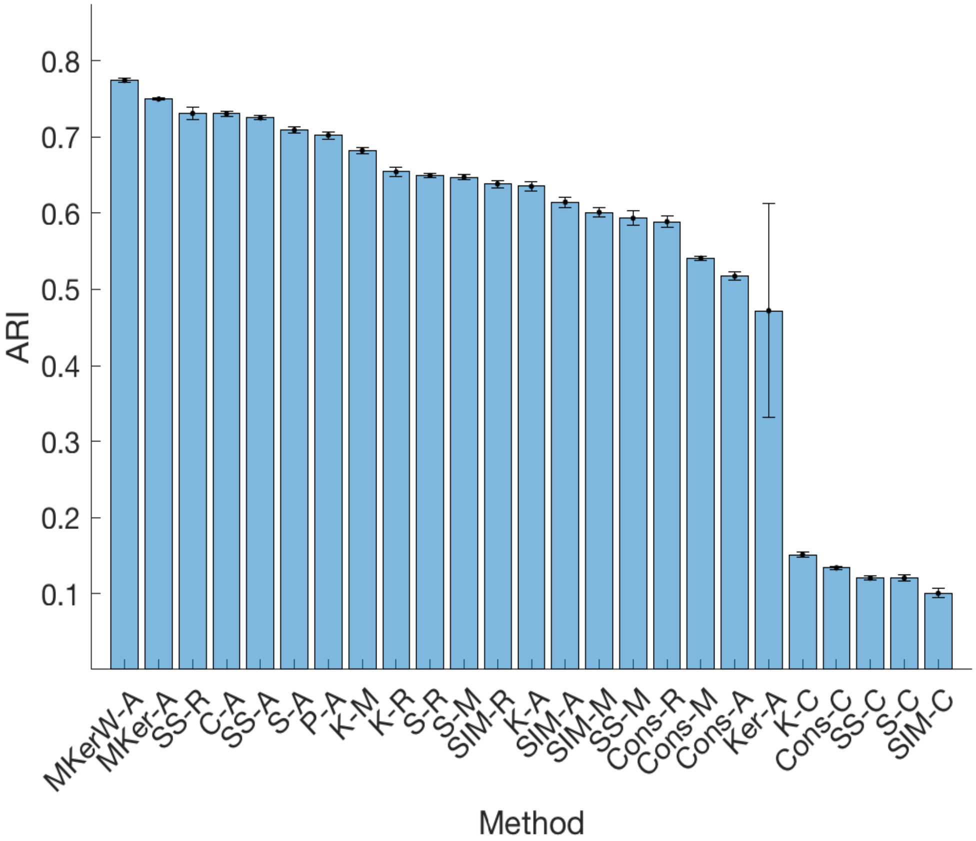

Figure S3: The average ARI and one standard deviation of 50 replicates for the 25 clustering methods when thirty patients were randomly selected from each of the 22 cancer types. The methods are ordered according to the ARI values.

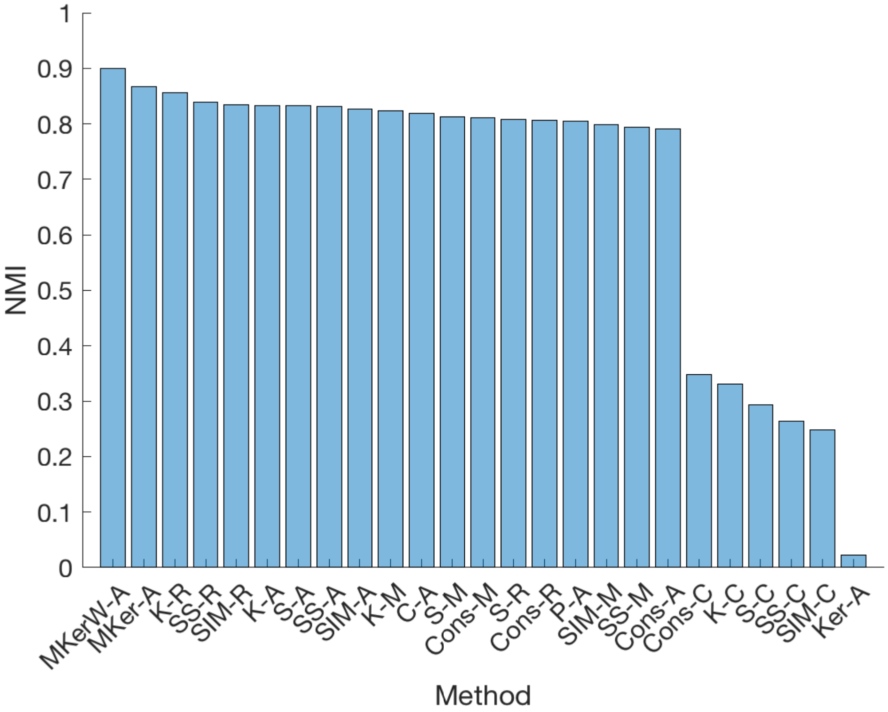

Figure S4: The NMI for the 25 clustering methods when all patients were included in the analysis. The methods are ordered according to the NMI values.

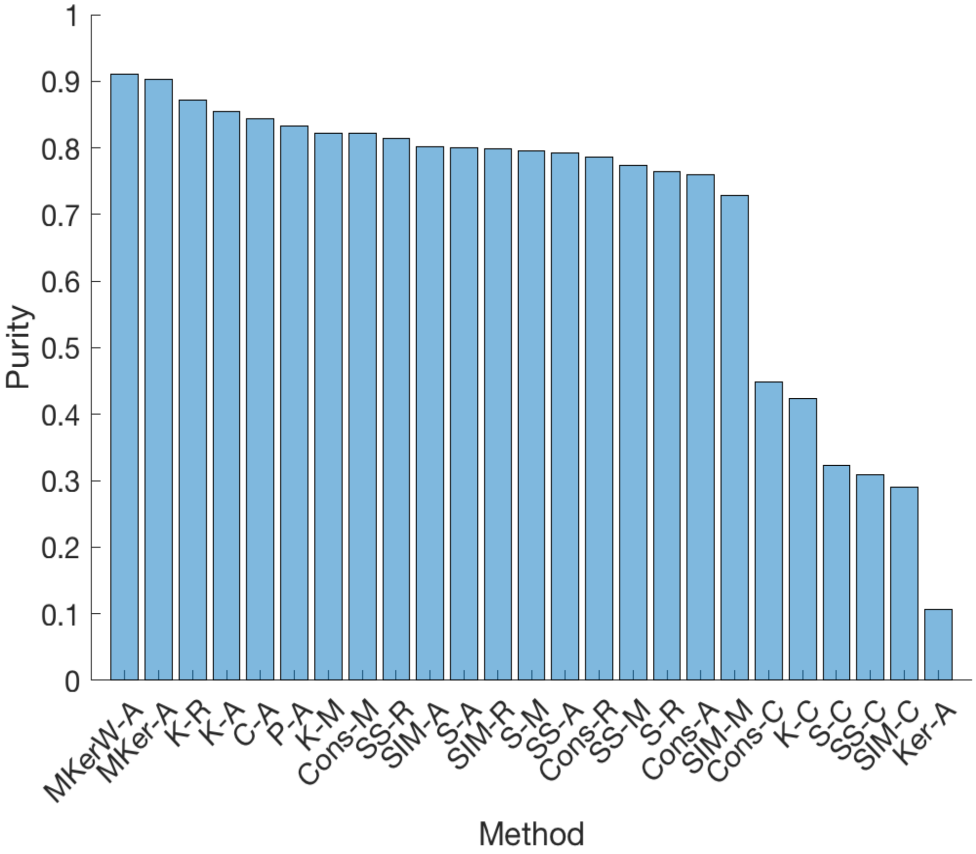

Figure S5: The Purity for the 25 clustering methods when all patients were included in the analysis. The methods are ordered according to the Purity values.

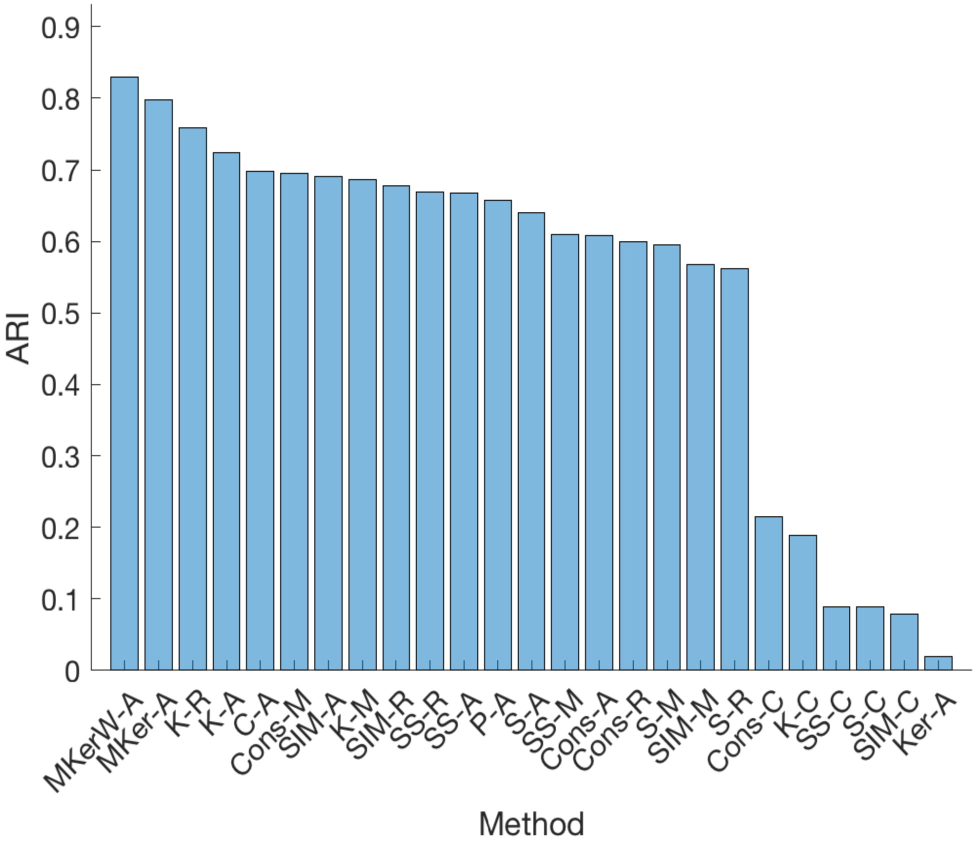

Figure S6: The ARI for the 25 clustering methods when all patients were included in the analysis. The methods are ordered according to the ARI values.

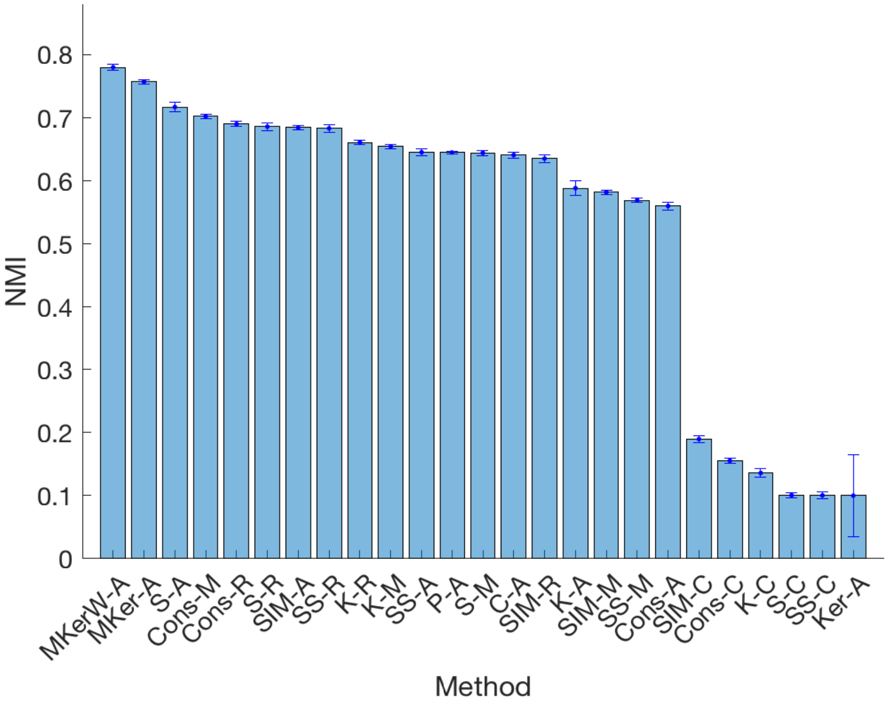

Figure S7: The average NMI and one standard deviation of 50 replicates for the 25 clustering methods when about half of patients were randomly selected from each of the 22 cancer types. The methods are ordered according to the NMI values.

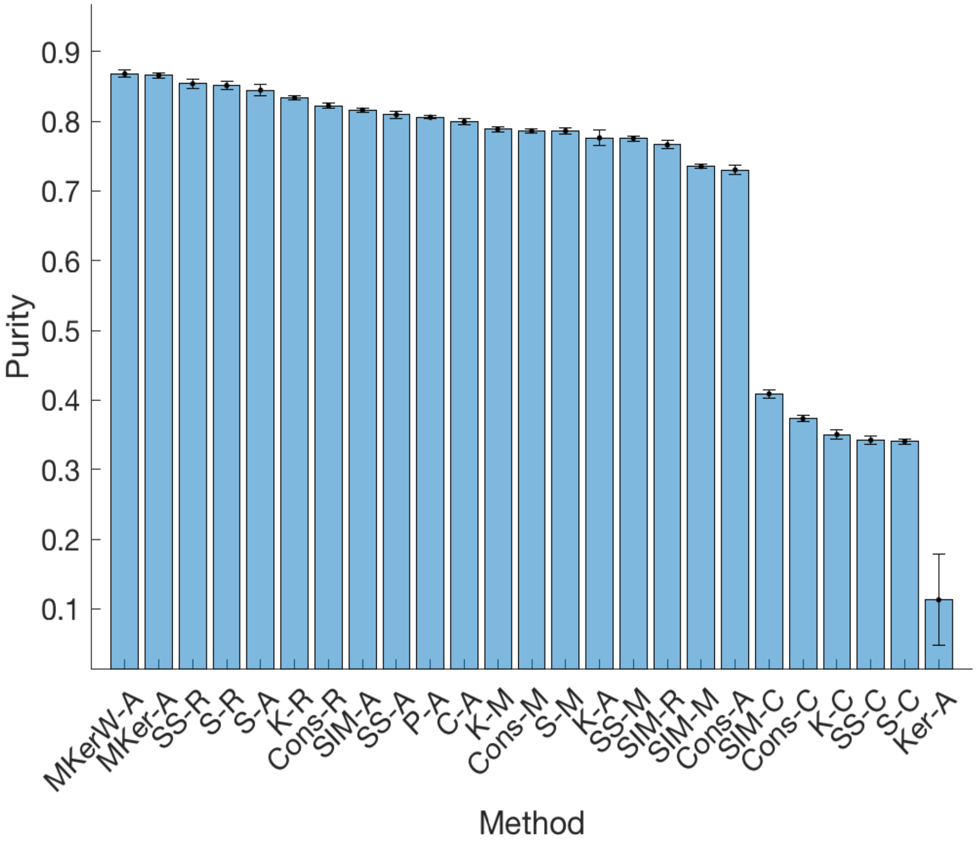

Figure S8: The average Purity and one standard deviation of 50 replicates for the 25 clustering methods when about half of patients were randomly selected from each of the 22 cancer types. The methods are ordered according to the Purity values.

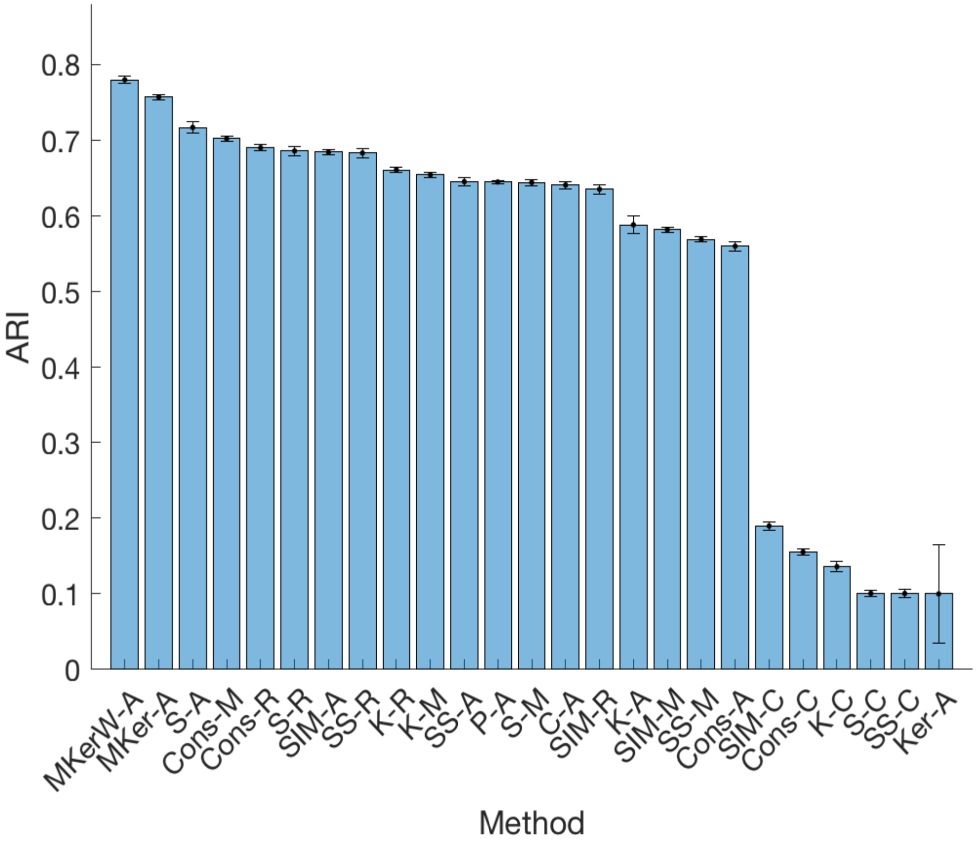

Figure S9: The average ARI and one standard deviation of 50 replicates for the 25 clustering methods when about half of patients were randomly selected from each of the 22 cancer types. The methods are ordered according to the ARI values.

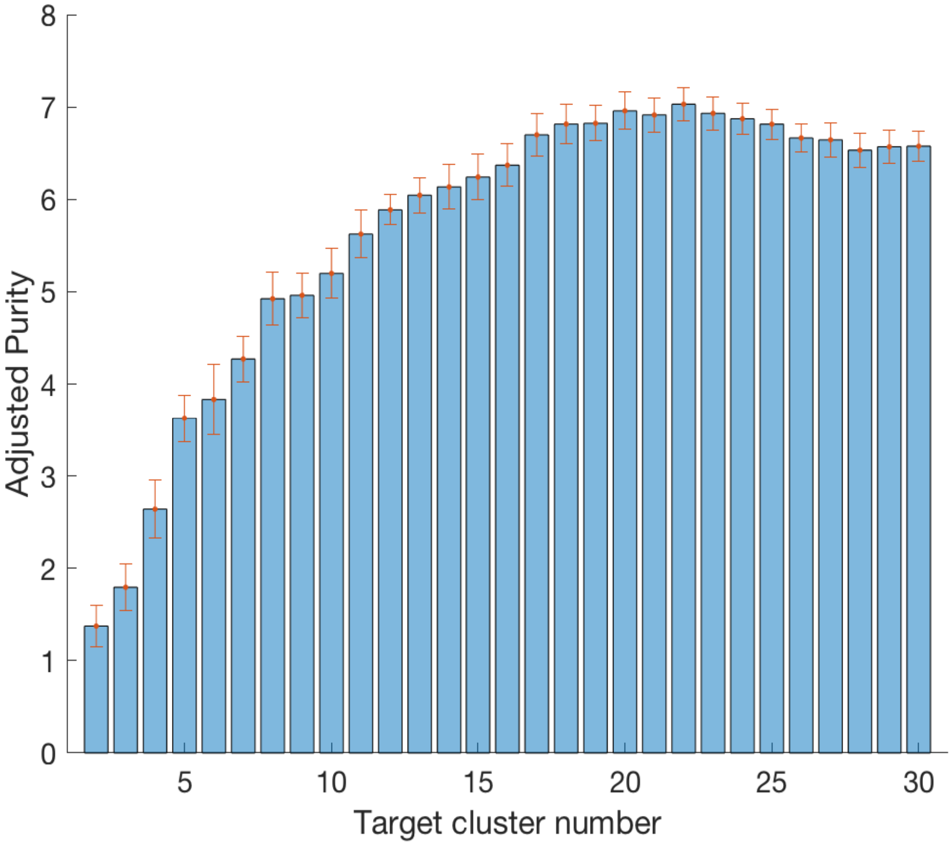

Figure S10: Robustness of clustering for additive noises when the target cluster number is varied between 2 and 30, when all patients were included in the analysis. The adjusted Purity values are averaged over 100 runs for each target cluster number. The error bars represent one standard deviation.

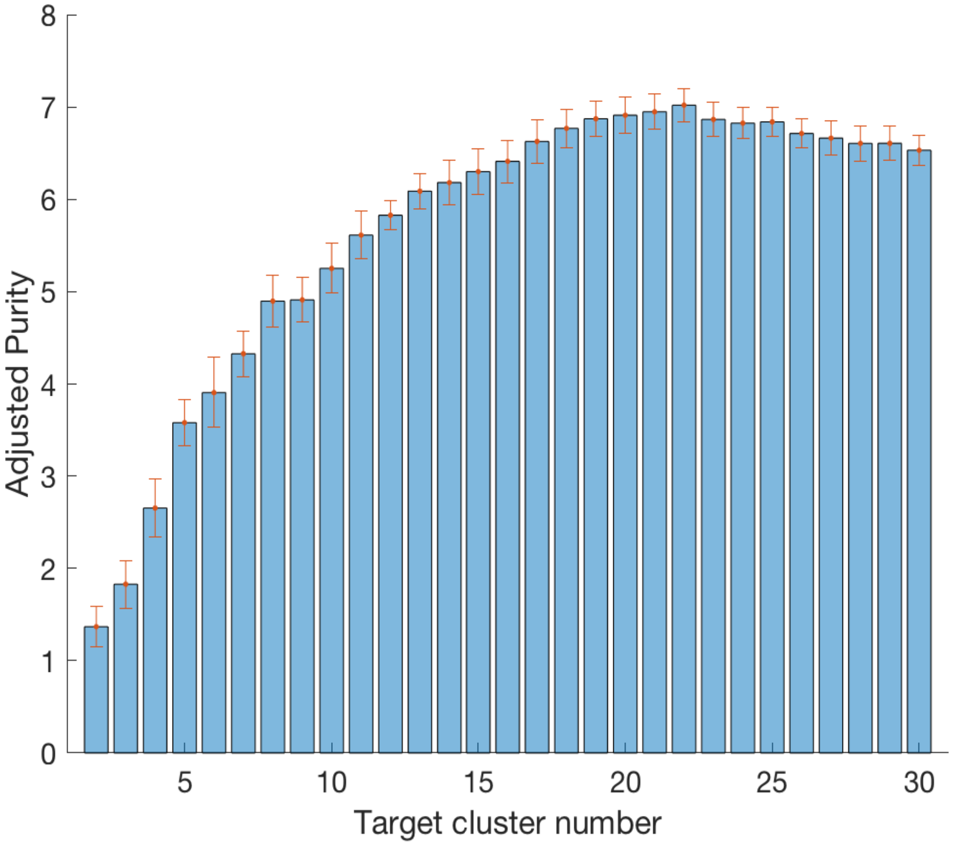

Figure S11: Robustness of clustering for additive noises when the target cluster number is varied between 2 and 30, when about half of patients were randomly selected from each of the 22 cancer types. The adjusted Purity values are averaged over 100 runs for each target cluster number. The error bars represent one standard deviation.

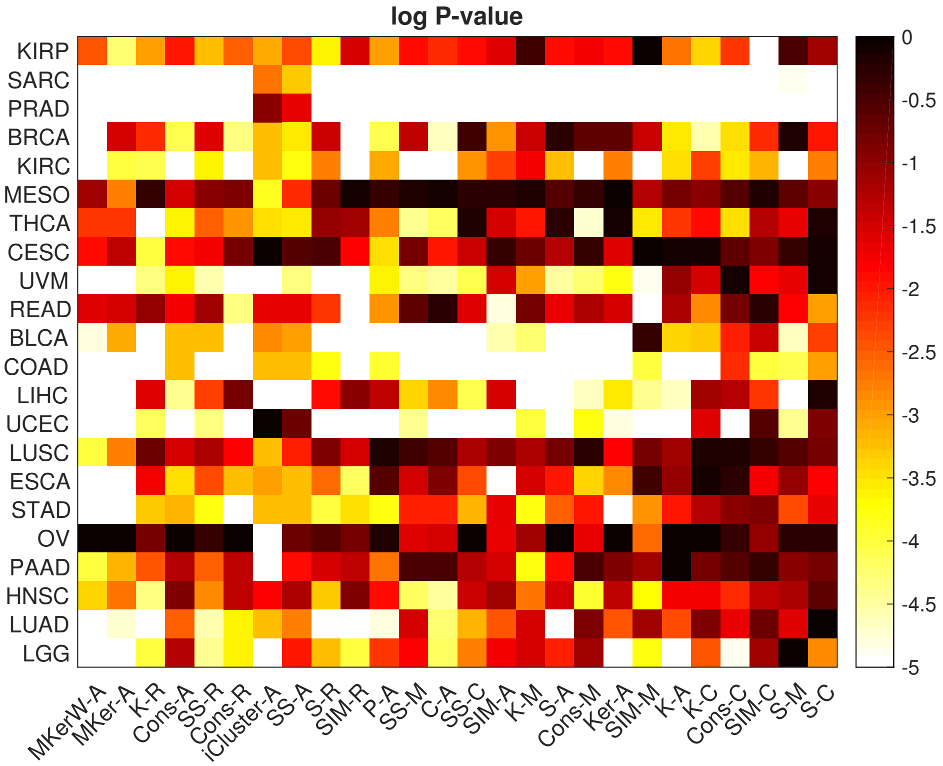

Figure S12: Heatmap of log p-values of the log-rank test.

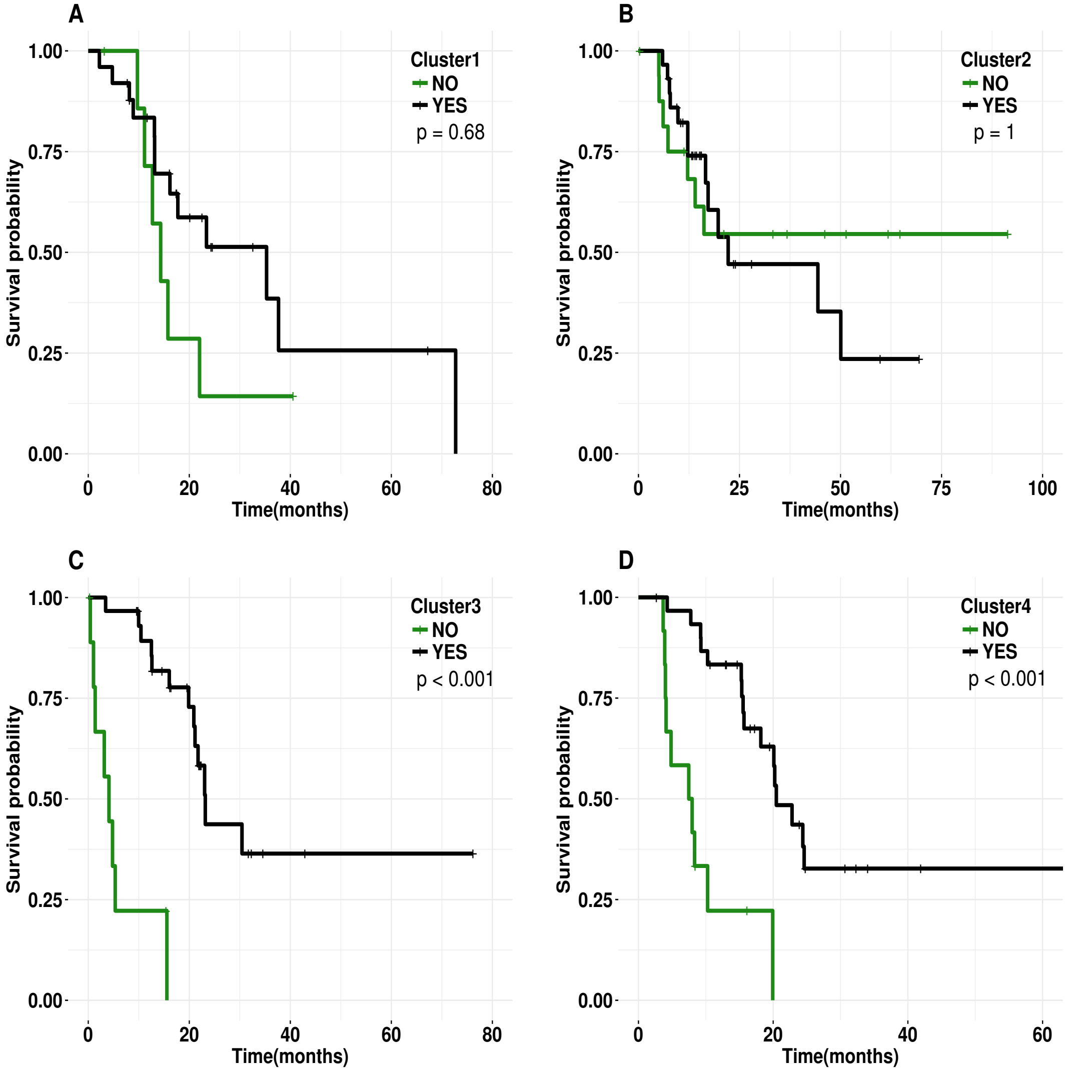

Figure S13: The survival distributions of Pancreatic cancer patients treated versus untreated for each of the four clusters identified by ‘MKerW-A’.

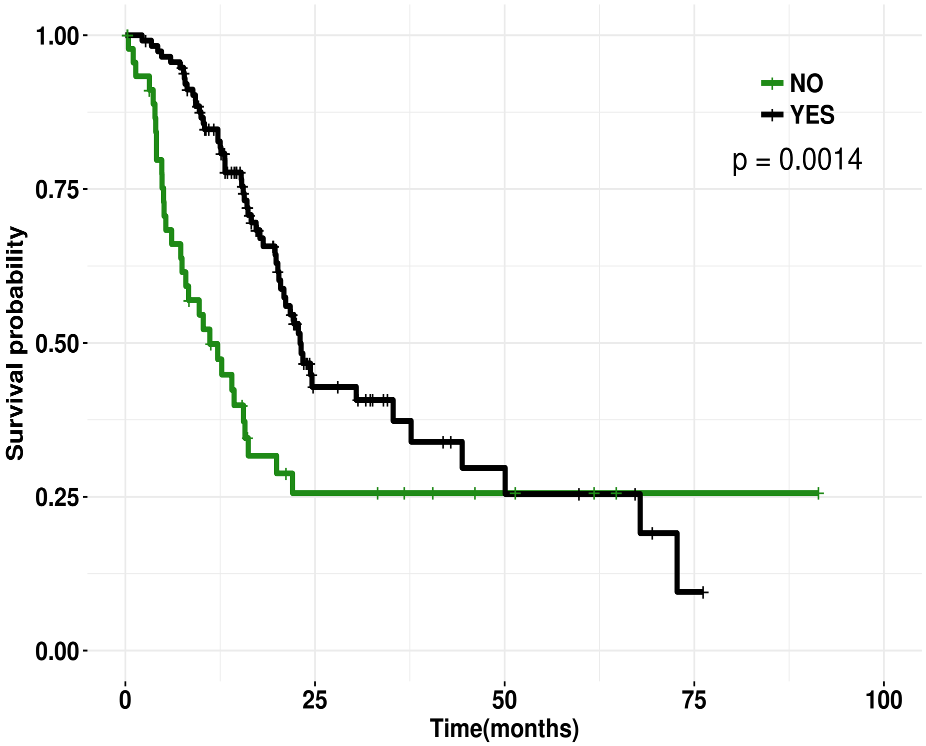

Figure S14: Difference of survival time of all pancreatic cancer patients treated versus those not treated with targeted therapy.

I Application to Stomach Adenocarcinoma

To gain further insights into the biological consequences of the identified clusters, we have investigated how patients of the individual clusters respond to different treatments. Note that four clusters are identified by our method, ‘MKerW-A’. Figure S15 shows the survival time of patients treated versus those not treated with Radiotherapy for each cluster.

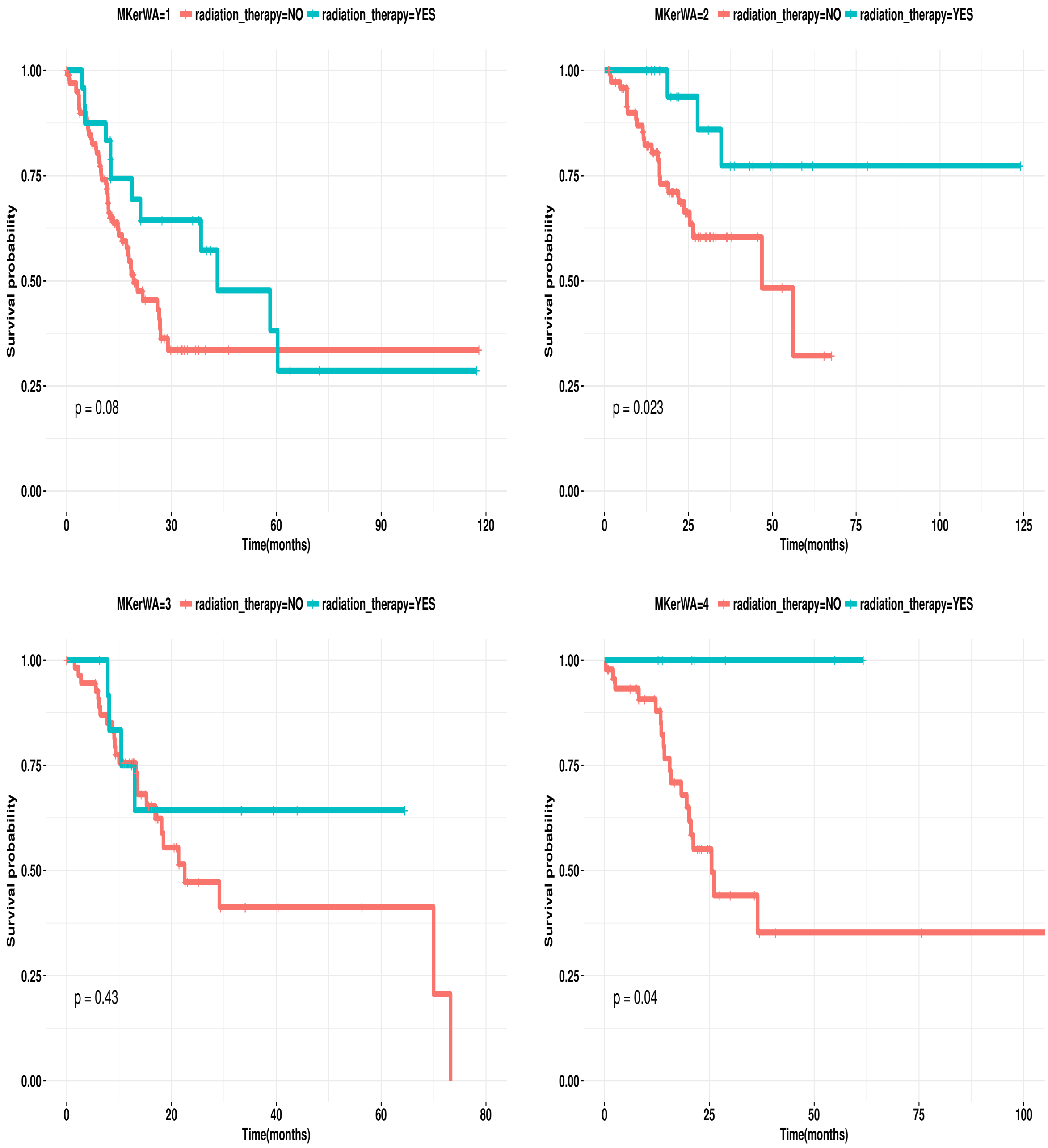

Figure S15: Survival curves of STAD patients for treatment with Radiotherapy in the different clusterings. The specified p-values are corrected for multiple testing using the Bonferroni method.

We observe that Radiotherapy is effective only in a subset of the identified groups. Patients in Clusters 2 and 4 have a significantly increased survival time when treated with Radiotherapy (log-rank test p-value < 0.05). Patients in Cluster 1 also seem to have an increased survival time with radiotherapy (log-rank test p-value $= 0.08$), but this pattern is reversed and the difference is very small after 5 years. For Cluster 3, we do not detect significant differences in survival time between treated and untreated patients.

References

Bertsekas, D. P. (2009). *Convex optimization theory*. Athena Scientific.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a pertubation. *SIAM Journal on Numerical Analysis*, **7**, 1–46.

Deng, W. and Yin, W. (2016). On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, **66**(3), 889–916.

Lin, T., Ma, S., and Zhang, S. (2015). On the global linear convergence of the admm with multi-block variables. *SIAM Journal on Optimization*, **25**(3), 1478–1497.

Lu, C. *et al.* (2016). Convex sparse spectral clustering: single-view to multi-view. *IEEE Transactions on Image Processing*, **25**(6), 2833–2843.

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, **12**.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, **18**, 1–9.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, **3**, 583–617.

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *NIPS*.

Wagner, S. and Wagner, D. (2007). Comparing clusterings- an overview.

Wang, W. and Lu, C. (2015). Projection onto the capped simplex. *arXiv preprint arXiv:1503.01002*.

Weinstein, J. N. *et al.* (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, **45**(10), 1113–1120.

Xu, Y. and Yin, W. (2013). A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, **6**(3), 1758–1789.