On hyperparameter tuning in general clustering problems

Supplementary Materials

This supplement contains detailed proofs of theoretical results in the main paper "On hyperparameter tuning in general clustering problems", additional theoretical results, and detailed description of the experimental parameter settings. We present proofs for MATR and MATR-CV in Sections A and Sections B respectively. Sections A.2 and Proposition S8 also contain additional theoretical results on the role of the hyperparameter in merging clusters in SDP-1 and SDP-2 respectively. Finally, Section D contains detailed parameter settings for the numerical experiments in the main paper and additional results from the experiments.

A. Additional theoretical results and proofs of results in Section 3

A.1. Proof of Theorem 1

Proof. If for tuning parameter λ, we have $\langle \hat{S}, \hat{X}_\lambda \rangle \geq \langle S, X_0 \rangle - \epsilon$, then

$$\langle S, \hat{X}_\lambda \rangle \geq \langle S, X_0 \rangle - |\langle \hat{S} - S, \hat{X}_\lambda \rangle| - \epsilon. \tag{S1.1}$$

First we will prove that this immediately gives an upper bound on $\|\hat{X}_\lambda - X_0\|_F$. We will remove the subscript λ for ease of exposition. Denote $\omega_k = \langle X_0, \hat{X}_{C_k,C_k} \rangle$, $\alpha_{ij} = \frac{\langle E_{i,j}, \hat{X} \rangle}{m_k(1-\omega_k)}$, when $\omega_k < 1$ and 0 otherwise, and off-diagonal set for kth cluster C_k^c as $\{(i,j)|i \in C_k, j \notin C_k\}$. Then we have

$$\begin{aligned}
\langle S, \hat{X} \rangle &= \sum_{k=1}^{r_0} a_{kk} \langle E_{C_k,C_k}, \hat{X} \rangle + \sum_{k=1}^{r_0} \sum_{(i,j) \in C_k^c} a_{ij} \langle E_{i,j}, \hat{X} \rangle \\
&= \sum_{k=1}^{r_0} a_{kk} m_k \omega_k + \sum_{k=1}^{r_0} m_k (1-\omega_k) \sum_{(i,j) \in C_k^c} a_{ij} \alpha_{ij} \\
&= \sum_{k=1}^{r_0} m_k \omega_k (a_{kk} - \sum_{(i,j) \in C_k^c} a_{ij} \alpha_{ij}) + \sum_{k=1}^{r_0} m_k \sum_{(i,j) \in C_k^c} a_{ij} \alpha_{ij}
\end{aligned} \tag{S1.2}$$

Since $\langle S, X_0 \rangle = \sum_k m_k a_{kk}$, by (S1.1), $\langle S, \hat{X} \rangle \geq \sum_k m_k a_{kk} - |\langle R, \hat{X} \rangle| - \epsilon$, we have

$$\sum_k m_k \omega_k (a_{kk} - \sum_{(i,j) \in C_k^c} a_{ij} \alpha_{ij}) + \sum_k m_k \sum_{(i,j) \in C_k^c} a_{ij} \alpha_{ij} \geq \sum_k m_k a_{kk} - |\langle R, \hat{X} \rangle| - \epsilon.$$

Note that, since S is weakly assortative, $a_{kk} - \sum_{(i,j) \in C_k^c} a_{ij} \alpha_{ij}$ is always positive because $\sum_{(i,j) \in C_k^c} \alpha_{ij} \leq 1$.

Denote $\epsilon' = |\langle R, \hat{X}\rangle| + \epsilon$, $\beta_k = \frac{m_k(a_{kk} - \sum_{C_k^c} \alpha_{ij} a_{ij})}{\sum_k m_k(a_{kk} - \sum_{C_k^c} \alpha_{ij} a_{ij})}$,

$$\sum_k m_k \omega_k (a_{kk} - \sum_{(i,j)\in C_k^c} a_{ij}\alpha_{ij}) \geq \sum_k m_k (a_{kk} - \sum_{(i,j)\in C_k^c} \alpha_{ij} a_{ij}) - \epsilon'$$

$$\sum_k \beta_k \omega_k \geq 1 - \frac{\epsilon'}{\sum_k m_k(a_{kk} - \sum_{C_k^c} \alpha_{ij} a_{ij})}$$

$$\sum_k \beta_k (1 - \omega_k) \leq \frac{\epsilon'}{\sum_k m_k(a_{kk} - \sum_{C_k^c} \alpha_{ij} a_{ij})}.$$

$$\sum_k (1 - \omega_k) \leq \sum_k \frac{\beta_k}{\beta_{\min}}(1 - \omega_k) \leq \frac{\epsilon'}{\beta_{\min} \sum_k m_k(a_{kk} - \sum_{C_k^c} \alpha_{ij} a_{ij})},$$

where $\beta_{\min} = \min_k \beta_k$. Since trace$(\hat{X})$ = trace(X_0),

$$\begin{aligned}
\left\|\hat{X} - X_0\right\|_F^2 &= \text{trace}((\hat{X} - X_0)^T(\hat{X} - X_0)) \\
&= \text{trace}(\hat{X} + X_0 - 2\hat{X}X_0) \\
&= 2\text{trace}(X_0) - 2\sum_k \langle X_0, \hat{X}_{C_k, C_k}\rangle \\
&= 2\sum_k (1 - \omega_k) \leq \frac{2\epsilon'}{\min_k m_k(a_{kk} - \sum_{C_k^c} \alpha_{ij} a_{ij})} \\
&\leq \frac{2\epsilon'}{n\pi_{\min} \min_k (a_{kk} - \max_{C_k^c} a_{ij})} = \frac{2\epsilon'}{\tau}.
\end{aligned}$$

Now consider the λ_* returned by MATR,

$$\langle \hat{S}, \hat{X}_{\lambda_*}\rangle \geq \langle \hat{S}, \hat{X}_{\lambda}\rangle \geq \langle S, X_0\rangle - \epsilon.$$

Then, following the above argument and from the condition from the theorem,

$$\|X_{\lambda_*} - X_0\|_F^2 \leq \frac{2\epsilon'}{n\pi_{\min} \min_k (a_{kk} - \max_{C_k^c} a_{ij})} \leq \frac{2}{\tau}(\epsilon + \sup_{X \in \mathcal{X}_{r_0}} |\langle X, R\rangle|).$$

□

A.2. Range of λ for merging clusters in SDP-1

Proposition S1. *Let $\tilde{X}$ be the optimal solution of SDP-1 for $A \sim SBM(B, Z_0)$ with λ satisfying*

$$\max_{k\neq\ell} B^*_{k,\ell} + \Omega(\sqrt{\frac{\rho \log n}{n\pi_{min}}}) \leq \lambda \leq \min_k B^*_{kk} - \max_{k,\ell=r-1,r} \frac{m_\ell}{n_k}(B_{\ell,\ell} - B_{r,r-1}) + O(\sqrt{\frac{\rho \log n}{n\pi^2_{\max}}}),$$

then $\tilde{X} = X^*$ with probability at least $1 - \frac{1}{n}$, where X^* is the unnormalized clustering matrix which merges the last two clusters, B^* is the corresponding $(r-1) \times (r-1)$ block probability matrix.

Remark: The proposition implies if the first $r-2$ clusters are more connected within each cluster than the last two clusters and the connection between first $r-2$ clusters and last two clusters are weak, we can find a range for λ that leads to merging the last two clusters with high probability. The results can be generalized to merging several clusters at one time. The result above highlights the importance of selecting λ as it affects the performance of SDP-1 significantly.

Proof. We develop sufficient conditions with a contruction of the dual certificate which guarantees X^* to be the optimal solution. The KKT conditions can be written as below:

First order stationary:

$$-A - \Lambda + \lambda E_n - \text{diag}(\beta) - \Gamma = 0$$

Primal feasibility:

$$X \succeq 0, X \geq 0, X_{ii} = 1 \quad \forall i = 1 \cdots, n$$

Dual feasibility:

$$\Gamma \geq 0, \Lambda \succeq 0$$

Complementary slackness

$$\langle \Lambda, X \rangle = 0, \Gamma \circ X = 0.$$

Consider the following construction: denote $T_k = C_k, n_k = m_k$, for $k < r-1$, $T_{r-1} = C_{r-1} \bigcup C_r, n_{r-1} = m_{r-1} + m_r$.

$$X_{T_k} = E_{n_k}$$

$$X_{T_k T_l} = 0, \text{for } k \neq l \leq r-1$$

$$\Lambda_{T_k} = -A_{T_k} + \lambda E_{n_k} - \lambda n_k I_{n_k} + \text{diag}(A_{T_k} \mathbf{1}_{n_k})$$

$$\Lambda_{T_k T_l} = -A_{T_k, T_l} + \frac{1}{n_l} A_{T_k, T_l} E_{n_l} + \frac{1}{n_k} E_{n_k} A_{T_k T_l} - \frac{1}{n_l n_k} E_{n_k} A_{T_k, T_l} E_{n_l}$$

$$\Gamma_{T_k} = 0$$

$$\Gamma_{T_k, T_l} = \lambda E_{n_k, n_l} - \frac{1}{n_l} A_{T_k, T_l} E_{n_l} - \frac{1}{n_k} E_{n_k} A_{T_k T_l} + \frac{1}{n_l n_k} E_{n_k} A_{T_k, T_l} E_{n_l}$$

$$\beta = \text{diag}(-A - \Lambda + \lambda E_n - \Gamma)$$

All the KKT conditions are satisfied by construction except for positive semidefiniteness of Λ and positiveness of Γ. Now, we show it one by one.

Positive semidefiniteness of Λ Since span$(1_{T_k}) \subset \ker(\Lambda)$, it suffices to show that for any $u \in span(1_{T_k})^{\perp}$, $u^T \Lambda u \geq 0$. Consider $u = \sum_k u_{T_k}$, where $u_{T_k} := u \circ 1_{T_k}$, then $u_{T_k} \perp 1_{T_k}$.

$$\begin{aligned} u^T \Lambda u &= -\sum_k u_{T_k}^T A_{T_k} u_{T_k} - \lambda \sum_k n_k u_{T_k}^T u_{T_k} + \sum_k u_{T_k}^T \text{diag}(A_{T_k} \mathbf{1}_{n_k}) u_{T_k} - \sum_{k \neq l} u_{T_k}^T A_{T_k T_l} u_{T_l} \\ &= -u^T (A - P) u - u^T P u - \lambda \sum_k n_k u_{T_k}^T u_{T_k} + \sum_k u_{T_k}^T \text{diag}(A_{T_k} \mathbf{1}_{n_k}) u_{T_k} \\ &= -u^T (A - P) u - u_{T_{k-1}}^T P_{T_{k-1} T_{k-1}} u_{T_{k-1}} - \lambda \sum_k n_k u_{T_k}^T u_{T_k} + \sum_k u_{T_k}^T \text{diag}(A_{T_k} \mathbf{1}_{n_k}) u_{T_k} \end{aligned} \tag{S1.3}$$

For the first term, we know

$$u^T (A - P) u \leq \|A - P\|_2 \|u\|_2^2 \leq O(\sqrt{n\rho}) \|u\|_2^2$$

with high probability.

For the second term, and note that $T_{r-1} = C_{r-1} \bigcup C_r$, and

$$P_{T_{r-1} T_{r-1}} = \begin{bmatrix} B_{r-1,r-1} E_{m_{r-1} m_{r-1}}, B_{r-1,r} E_{m_{r-1} m_r} \\ B_{r,r-1} E_{m_r m_{r-1}}, B_{r,r} E_{m_r m_r} \end{bmatrix}$$

Since $u_{T_{r-1}} \perp 1_{n_{r-1}}$,

$$u_{T_{r-1}}^T \begin{bmatrix} B_{r-1,r} E_{m_{r-1} m_{r-1}}, B_{r-1,r} E_{m_{r-1} m_r} \\ B_{r,r-1} E_{m_r m_{r-1}}, B_{r,r-1} E_{m_r m_r} \end{bmatrix} u_{T_{r-1}} = 0,$$

therefore

$$\begin{aligned} u_{T_{r-1}}^T P_{T_{r-1}T_{r-1}} u_{T_{r-1}} &= u_{T_{r-1}}^T \begin{bmatrix} (B_{r-1,r-1} - B_{r-1,r})E_{m_{r-1}m_{r-1}}, 0 \\ 0, (B_{r,r} - B_{r-1,r})E_{m_r m_r} \end{bmatrix} u_{T_{r-1}} \\ &\leq \max\{m_{r-1}(B_{r-1,r-1} - B_{r-1,r}), m_r(B_{r,r} - B_{r,r-1})\} \|u\|_2^2 \end{aligned} \tag{S1.4}$$

Consider the last term $\sum_k u_{T_k}^T \text{diag}(A_{T_k}\mathbf{1}_{n_k})u_{T_k}$. Using Chernoff, we know

$$||\text{diag}(A_{T_k}\mathbf{1}_{n_k})||_2 \geq B_{k,k}^* n_k - \sqrt{6\rho n_k \log n_k}$$

with high probability, where for $k, l < r-1$,

$$B_{kl}^* = B_{kl},$$

$$B_{k,r-1}^* = \frac{m_{r-1}B_{k,r-1} + m_r B_{k,r}}{m_{r-1} + m_r},$$

$$B_{r-1,r-1}^* = \frac{(m_{r-1}^2 B_{r-1,r-1} + 2 * m_r m_{r-1} B_{r-1,r} + m_r^2 B_{r,r})}{(m_{r-1} + m_r)^2}.$$

Therefore, :

$$-\lambda \sum_k n_k u_{T_k}^T u_{T_k} + \sum_k u_{T_k}^T \text{diag}(A_{T_k}\mathbf{1}_{n_k})u_{T_k} \geq \min_k (B_{k,k}^* n_k - \Omega(\sqrt{\rho n_k \log n}) - \lambda n_k) \|u\|_2^2 .$$

So with equation S1.3, a sufficient condition for positive semidefiniteness of Λ is

$$\min_k (B_{k,k}^* n_k - \Omega(\sqrt{\rho n_k \log n}) - \lambda n_k) \geq O(\sqrt{n\rho}) + \max\{m_{r-1}(B_{r-1,r-1} - B_{r-1,r}), m_r(B_{r,r} - B_{r,r-1})\}$$

which implies,

$$\lambda \leq \min_k B_{kk}^* - \max_k \max\{\frac{m_{r-1}}{n_k}(B_{r-1,r-1} - B_{r-1,r}), \frac{m_r}{n_k}(B_{r,r} - B_{r,r-1})\} + O(\sqrt{\rho \log n / n\pi_{\max}^2})$$

Positiveness of Γ For $i \in T_k, j \in T_l$, we have

$$\Gamma_{i,j} = \lambda - \frac{\sum_{m \in T_l} A_{i,m}}{n_l} - \frac{\sum_{m \in T_k} A_{m,j}}{n_k} + \frac{1}{n_k n_l} \sum_{m \in T_k, o \in T_l} A_{mo}.$$

Therefore, block-wise mean of Γ will be

$$\mathbb{E}[\Gamma_{T_k,T_l}] = (\lambda - B_{k,l}^*)E_{n_k,n_l},$$

and the variance for each entry belonging to cluster k and l will be in order of $O(\rho/(n_k n_l))$.

Using Chernoff bound, we have

$$p(|\Gamma_{i,j} - (\lambda - B_{k,l}^*)| > \lambda - B_{k,l}^*) \leq 2\exp\left[-\frac{n_k n_l}{2\rho}(\lambda - B_{k,l}^*)^2\right].$$

Therefore, as long as $\lambda \geq \max_{k \neq l} B_{k,l}^* + \Omega(\sqrt{\rho \log n / n\pi_{\min}})$, we have

$$p(\Gamma_{i,j} < 0) \leq 2\exp\left[-\frac{n\pi_{min}\log n}{2}\right]$$

We then applying the union bound and conclude that $\Gamma_{T_k T_l} > 0$ with a high probability when $\lambda \geq \max_{k \neq l} B_{k,l}^* + \Omega(\sqrt{\rho \log n / n\pi_{\min}})$.

□

Proposition S2. *As long as* $\max_{k\neq l} B_{k,l} + \Omega(\sqrt{\rho \log n/n\pi_{min}}) \leq \lambda \leq \min_k B_{kk} + O(\sqrt{\rho \log n/n\pi^2_{\max}})$*, SDP-1 exactly recovers* X_0 *with high probability.*

Proof. We follow the same primal-dual construction as Proposition S1 without merging the last two clusters. Consider the following construction: denote $T_k = C_k, n_k = m_k$, for $k = 1, ..., r$. We show the positive semidefiniteness and Positiveness of Λ and Γ respectively.

Positve semidefiniteness of Λ Since $\text{span}(1_{T_k}) \subset \ker(\Lambda)$, it suffices to show that for any $u \in \text{span}(1_{T_k})^\perp$, $u^T\Lambda u \geq 0$. Consider $u = \sum_k u_{T_k}$, where $u_{T_k} := u \circ 1_{T_k}$, and $u_{T_k} \perp 1_{T_k}$, we have

$$\begin{aligned} u^T\Lambda u &= -\sum_k u_{T_k}^T A_{T_k} u_{T_k} - \lambda \sum_k n_k u_{T_k}^T u_{T_k} + \sum_k u_{T_k}^T \text{diag}(A_{T_k}\mathbf{1}_{n_k}) u_{T_k} - \sum_{k\neq l} u_{T_k}^T A_{T_k T_l} u_{T_l} \\ &= -u^T(A-P)u - u^T P u - \lambda \sum_k n_k u_{T_k}^T u_{T_k} + \sum_k u_{T_k}^T \text{diag}(A_{T_k}\mathbf{1}_{n_k}) u_{T_k} \\ &= -u^T(A-P)u - \lambda \sum_k n_k u_{T_k}^T u_{T_k} + \sum_k u_{T_k}^T \text{diag}(A_{T_k}\mathbf{1}_{n_k}) u_{T_k} \end{aligned} \tag{S1.5}$$

For the first term, we know

$$u^T(A-P)u \leq \|A-P\|_2 \|u\|_2^2 \leq O(\sqrt{n\rho}) \|u\|_2^2$$

with high probability, and using Chernoff, we have

$$||\text{diag}(A_{T_k}\mathbf{1}_{n_k})||_2 \geq B_{k,k} n_k - \sqrt{6\rho n_k \log n_k}$$

with high probability. Therefore,

$$-\lambda \sum_k n_k u_{T_k}^T u_{T_k} + \sum_k u_{T_k}^T \text{diag}(A_{T_k}\mathbf{1}_{n_k}) u_{T_k} \geq \min_k (B_{k,k} n_k - \Omega(\sqrt{\rho n_k \log n}) - \lambda n_k) \|u\|_2^2,$$

which implies a sufficient condition for positive semidefiniteness of Λ is

$$\lambda \leq \min_k B_{kk} + O(\sqrt{\rho \log n/n\pi^2_{\max}}),$$

and the lower bound can be obtained exactly the same way as Proposition S1. Using Chernoff bound, $\Gamma_{T_k T_l} > 0$ with high probability as long as $\lambda \geq \max_{k\neq l} B_{k,l} + \Omega(\sqrt{\rho \log n/n\pi_{\min}})$.

□

A.3. Proof of Corollary 2

Proof. This result comes directly from Theorem 1. We have $S = \tilde{P}$, $R = (A-P) + (P-\tilde{P})$. For λ_0,

$$\langle \hat{X}_{\lambda_0}, A\rangle \geq \langle X_0, \tilde{P}\rangle - O(r\rho) - \epsilon,$$

where $r\rho = o(\tau)$ since $r\sqrt{n\rho} = o(\tau)$, and for any $\hat{X} \in \mathcal{X}_r$,

$$|\langle A - \tilde{P}, \hat{X}\rangle| \leq ||A-P||_{op}\text{trace}(\hat{X}) + O(r\rho) = O_P(r\sqrt{n\rho}).$$

The last inequality follows by (Lei & Rinaldo, 2015) and $n\rho \geq c \log n$. □

A.4. Proof of Corollary 4

Proof. Using the proof of Theorem 1 in Yan & Sarkar (2016), we have $\sup|\hat{S}_{ij} - S_{ij}| \leq O(\sqrt{\log n/d})$ with probability at least $1 - 1/n$. Therefore, $|\langle R, \hat{X}\rangle| = |\langle \hat{S} - S, \hat{X}\rangle| \leq O(n\sqrt{\log n/d})$ w.h.p. The result comes directly from Theorem 1. □

B. Additional theoretical results and proofs of results in Section 4

B.1. Proof of Theorem 6

Proof. With probability greater than $1 - \delta_{\text{est}} - \delta_{\text{over}} - \delta_{\text{under}}$, the following three inequalities hold.

For $r_T \geq r_t > r$:

$$\begin{aligned}\langle \hat{S}^{22}, \hat{X}_r^{22} \rangle &\geq \langle \hat{S}^{22}, X_0^{22} \rangle - \epsilon_{\text{est}} \geq \max_{r_T \geq r_t > r} \langle \hat{S}^{22}, \hat{X}_{r_t}^{22} \rangle - \epsilon_{\text{est}} - \epsilon_{\text{over}} \\ &\geq \max_{r_T \geq r_t > r} \langle \hat{S}^{22}, \hat{X}_{r_t}^{22} \rangle - \Delta\end{aligned}$$

For $r_t < r$:

$$\begin{aligned}\langle \hat{S}^{22}, \hat{X}_r^{22} \rangle &\geq \langle \hat{S}^{22}, X_0^{22} \rangle - \epsilon_{\text{est}} \geq \max_{r_t < r} \langle \hat{S}^{22}, \hat{X}_{r_t}^{22} \rangle - \epsilon_{\text{est}} + \epsilon_{\text{under}} \\ &> \max_{r_t < r} \langle \hat{S}^{22}, \hat{X}_{r_t}^{22} \rangle + \Delta\end{aligned}$$

Therefore, with probability at least $1 - \delta_{\text{est}} - \delta_{\text{over}} - \delta_{\text{under}}$, $\langle \hat{S}^{22}, \hat{X}_r^{22} \rangle \geq \max_t \langle \hat{S}^{22}, \hat{X}_{r_t}^{22} \rangle - \Delta$. Let $R = \{r_t : \langle \hat{S}^{22}, \hat{X}_{r_t}^{22} \rangle \geq \max_t \langle \hat{S}^{22}, \hat{X}_{r_t}^{22} \rangle - \Delta\}$. It follows then $r \in R$.

Furthermore, with probabiltiy at least $1 - \delta_{\text{est}} - \delta_{\text{under}}$, for $r_t < r$:

$$\begin{aligned}\langle \hat{S}^{22}, \hat{X}_{r_t}^{22} \rangle &\leq \langle \hat{S}^{22}, X_0^{22} \rangle - \epsilon_{\text{under}} \leq \langle \hat{S}^{22}, \hat{X}_r^{22} \rangle + \epsilon_{\text{est}} - \epsilon_{\text{under}} \\ &< \max_t \langle \hat{S}^{22}, \hat{X}_{r_t}^{22} \rangle - \Delta.\end{aligned}$$

Therefore, for any $r_t < r$, $r_t \notin R$, and $\min\{r_t : r_t \in R\} = r$. □

B.2. Proof of Theorem 8

We first prove a concentration lemma that holds for any normalized clustering matrix X independent of A.

Lemma S3. *Consider a an adjacency matrix A and its population version P. Let X be a normalized clustering matrix independent of A. Then with probability at least $1 - O(n^{-1})$,*

$$|\langle A - P, X \rangle| \leq (1 + B_{\max}) \sqrt{\textit{trace}(X) \log n}$$

with $B_{\max} = \max_{i,j} B_{ij}$.

Proof. The result follows from Hoeffding's inequality and the fact that X is a projection matrix.

By independence between A and X,

$$\begin{aligned}P\left(\sum_{i<j}(A_{ij} - P_{ij})X_{ij} > t\right) &\leq \exp(-\frac{2t^2}{(1 + B_{\max})^2 \sum_{i<j} X_{ij}^2}) \\ &\leq \exp(-\frac{4t^2}{(1 + B_{\max})^2 \left\|X\right\|_F^2}) \\ &= \exp(-\frac{4t^2}{(1 + B_{\max})^2 \text{trace}(X)})\end{aligned}$$

Let $t = \frac{1}{2}(1 + B_{\max})\sqrt{\text{trace}(X) \log n}$, then by symmetry in A and X, $P(\langle A - P, X \rangle > (1 + B_{\max})\sqrt{\text{trace}(X) \log n}) = O(1/n)$. The other direction is the same.

□

In order to prove Theorem 8, we need to derive the three error bounds in Theorem 6 in this setting. For notational convenience, we first derive the bounds for A and a general normalized clustering matrix $\hat{X}$, with the understanding that the same asymptotic bounds apply to estimates obtained from the training graph provided the split is random and the number of training nodes is $\Theta(n)$.

Lemma S4. *For a sequence of underfitting normalized clustering matrix $\{\hat{X}_{r_t}\}_{r_t<r}$, all independent of A, provided $n\rho/\sqrt{\log n} \to \infty$, we have*

$$\max_{r_t<r} \langle A, \hat{X}_{r_t}\rangle \le \langle A, X_0\rangle - \Omega_P(n\rho\pi_{min}^2/r^2),$$

for fixed r and $\pi_{\min}$.

Proof. Let $\{\hat{C}_k\}$ be the clusters associated with $\hat{X}$. Denote $\gamma_{k,i} = |\hat{C}_k \cap C_i|$, and $\hat{m}_k = |\hat{C}_k| = \sum_i \gamma_{k,i}$, $r_t = \text{rank}(\hat{X}) < r$. First note that for each $i \in [r]$, $\exists k \in [r_t]$, s.t. $\gamma_{k,i} \ge |C_i|/r_t$. Since $r > r_t$, by the Pigeonhole principle, we see that $\exists i_0, j_0, k_0$, $i_0 \ne j_0$, such that,

$$\begin{aligned}
\gamma_{k_0,i_0} &= |\hat{C}_{k_0} \cap C_{i_0}| \ge |C_{i_0}|/r_t \ge \pi_{\min} n/r_t \\
\gamma_{k_0,j_0} &= |\hat{C}_{k_0} \cap C_{j_0}| \ge |C_{j_0}|/r_t \ge \pi_{\min} n/r_t
\end{aligned} \tag{S2.6}$$

For each $k \ne k_0$,

$$\frac{\sum_{i,j} B_{i,j}\gamma_{k,i}\gamma_{k,j}}{\hat{m}_k} \le \frac{\sum_i B_{i,i}\sum_j \gamma_{k,i}\gamma_{k,j}}{\hat{m}_k} = \sum_i B_{i,i}\gamma_{k,i}.$$

For $k = k_0$,

$$\begin{aligned}
\frac{\sum_{i,j} B_{i,j}\gamma_{k,i}\gamma_{k,j}}{\hat{m}_k} &= \frac{\sum_{i,j} B_{i,i}\gamma_{k,i}\gamma_{k,j}}{\hat{m}_k} + \frac{\sum_{i\ne j}(B_{i,j}-B_{i,i})\gamma_{k,i}\gamma_{k,j}}{\hat{m}_k} \\
&= \sum_i B_{i,i}\gamma_{k,i} + \frac{\sum_{i\ne j}(B_{i,j}-B_{i,i})\gamma_{k,i}\gamma_{k,j}}{\hat{m}_k} \\
&\le \sum_i B_{i,i}\gamma_{k,i} + \frac{(2B_{i_0,j_0} - B_{i_0,i_0} - B_{j_0,j_0})\gamma_{k,i_0}\gamma_{k,j_0}}{\hat{m}_k} \\
&= \sum_i B_{i,i}\gamma_{k,i} - \frac{((B_{i_0,i_0} - B_{i_0,j_0}) + (B_{j_0,j_0} - B_{i_0,j_0}))\,\gamma_{k,i_0}\gamma_{k,j_0}}{\hat{m}_k} \\
&\overset{(a)}{\le} \sum_i B_{i,i}\gamma_{k,i} - \frac{2\tau\gamma_{k,i_0}\gamma_{k,j_0}}{n\pi_{\min}\hat{m}_k} \\
&\le \sum_i B_{i,i}\gamma_{k,i} - \frac{2\tau\pi_{\min} n}{r_t^2\hat{m}_k},
\end{aligned} \tag{S2.7}$$

where $\tau = n\pi_{\min}p_{\text{gap}}$, $p_{\text{gap}} := \min_i(B_{i,i} - \max_{j\ne i} B_{i,j})$. (a) is true by definition of τ and Eq (S2.6).

Therefore, since $\hat{m}_{k_0} \le n$,

$$\begin{aligned}
\langle P, \hat{X}\rangle &= \sum_{k=1}^{r_t} \frac{\sum_{i,j} B_{i,j}\gamma_{k,i}\gamma_{k,j}}{\hat{m}_k} - O(\rho r_t) \le \sum_{k=1}^{r_t}\sum_{i=1}^{r} B_{i,i}\gamma_{k,i} - \Omega(\frac{\tau\pi_{\min} n}{r_t^2\hat{m}_{k_0}}) \\
&= \langle P, X_0\rangle - \Omega\left(\frac{\tau\pi_{\min}}{r_t^2}\right).
\end{aligned} \tag{S2.8}$$

Next by Lemma S3, for each X with $\text{trace}(X) \le r$,

$$|\langle A - P, X\rangle| \le (1 + B_{\max})\sqrt{r\log n}$$

with probabiltiy at least $1 - O(1/n)$. By a union bound and using the same argument, w.h.p.

$$\max_{r_t<r} |\langle A - P, \hat{X}_{r_t}\rangle| \leq (1 + B_{\max})\sqrt{r \log n} \tag{S2.9}$$

Eqs (S2.8) and (S2.9) imply w.h.p.

$$\begin{aligned}&\langle A, X_0\rangle - \max_{r_t<r}\langle A, \hat{X}_{r_t}\rangle \\ =&\Omega\left(\frac{np_{gap}\pi^2_{\min}}{r^2}\right) - O(\sqrt{r\log n}) = \Omega\left(\frac{np_{gap}\pi^2_{\min}}{r^2}\right)\end{aligned} \tag{S2.10}$$

using the condition in the Lemma.

□

Lemma S5. *For a sequence of overfitting normalized clustering matrix $\{\hat{X}_{r_t}\}_{r<r_t\leq r_T}$, all independent of A, $r_T = \Theta(r)$, we have w.h.p.*

$$\max_{r<r_t\leq r_T} \langle A, \hat{X}_{r_t}\rangle \leq \langle A, X_0\rangle + (1 + B_{\max})\sqrt{r_T \log n} + B_{\max} r.$$

Proof. First note, for any $\hat{X}$, using weak assortativity on B,

$$\begin{aligned}\langle P, \hat{X}\rangle &\leq \sum_{i,j} \hat{X}_{i,j} B_{C(i),C(j)} \\ &\leq \sum_i B_{C(i),C(i)} \sum_j \hat{X}_{i,j} \\ &\leq \langle P, X_0\rangle + B_{\max} r,\end{aligned} \tag{S2.11}$$

where $C(i)$ denotes the cluster node i belongs to. By the same argument as in Eq (S2.9), w.h.p.

$$\max_{r<r_t\leq r_T} |\langle A - P, \hat{X}_{r_t}\rangle| \leq (1 + B_{\max})\sqrt{r_T \log n} \tag{S2.12}$$

From the above

$$\max_{r<r_t\leq r_T} \langle A, \hat{X}_{r_t}\rangle \leq \langle A, X_0\rangle + (1 + B_{\max})\sqrt{r_T \log n} + B_{\max} r.$$

□

Lemma S6. *With probabilitiy at least $1 - O(1/n)$, MATR-CV achieves exact recovery on the testing nodes given the true cluster number r, i.e. $\hat{X}^{22}_r = X^{22}_0$, provided $n\pi_{\min}\rho/\log n \to \infty$, $\gamma_{train} = \Theta(1)$.*

Proof. Denote m^{11}_k, m^{22}_k as the number of nodes belonging to the cluster C_k in the training graph and testing graph respectively.

First, with Theorem 2 in (Yan et al., 2017) and Lemma S7, we know SDP-2 can achieve exact recovery on traininng graph with high probability. Now, consider a node s in testing graph, and assume it belongs to cluster C_k. The probability that it is assigned to cluster k is: $P(\frac{\sum_{j\in C_k} A^{21}_{s,j}}{m^{11}_k} \geq \max_{l\neq k} \frac{\sum_{j\in C_l} A^{21}_{s,j}}{m^{11}_l})$.

Using the Chernoff bound, for some constant c,

$$\begin{aligned}P(\frac{\sum_{j\in C_k} A^{21}_{s,j}}{m^{11}_k} \geq B_{k,k} - c\sqrt{B_{k,k}\log n/m^{11}_k}) \geq 1 - n^{-3}; \\ P(\frac{\sum_{j\in C_l} A^{21}_{s,j}}{m^{11}_l} \leq B_{l,k} + c\sqrt{B_{l,k}\log n/m^{11}_l}) \geq 1 - n^{-3}.\end{aligned} \tag{S2.13}$$

Since the graph split is random, for each k, with probability at least $1-n^{-3}$, $|m_k^{11}-\gamma_{\text{train}}m_k| \le c_1\sqrt{m_k\log n}$ for some constant c_1. By a union bound, this holds for all k with probability at least $1-rn^{-3}$. Then under this event,

$$\sqrt{\frac{B_{l,k}\log n}{m_l^{11}}} \le c_2\sqrt{\frac{B_{l,k}\log n}{n\pi_{\min}}}$$

for some c_2 since $n\pi_{\min}/\log n \to \infty$. Since $n\pi_{\min}\rho/\log n \to \infty$, by Eq (S2.13), with probabitliy at least $1-O(rn^{-3})$,

$$B_{k,k} - c\sqrt{B_{k,k}\log n/m_k^{11}} > \max_{l\neq k} B_{l,k} + c\sqrt{B_{l,k}\log n/m_k^{11}},$$

and node s is assigned correctly to cluster k. Taking a union over all s in the training set, with probability $1-O(rn^{-2})$, MATR-CV would give exact recovery for the testing graph given r.

□

Lemma S7. *If $m_k \ge \pi n$, then $m_k^{11} \ge \pi n\gamma_{train}$, and $m_k^{22} \ge \pi n(1-\gamma_{train})$, with high probability. If $\max_{k,l}\frac{m_k}{m_l} \le \delta$, then $\max_{k,l}\frac{m_k^{11}}{m_l^{11}} \le \delta + o(1)$ with high probability.*

Proof. The result follows from (Skala, 2013). □

Proof for Theorem 8

Proof. First we note that by a similar argument as in Lemma S6, $|m_k^{22}-(1-\gamma_{\text{train}})m_k| \le c\sqrt{m_k\log n}$ for all k with probability at least $1-rn^{-3}$. Then the size of the smallest cluster of the test graph A^{22} will be of the same order as $n\pi_{\min}$. Also A^{22} has size $\Theta(n)$. A^{22} is independent of any $\hat{X}^{22}$. Thus in Theorem 6, applying Lemma S4 and Lemma S5 to A^{22} shows

$$\epsilon_{under} = \Omega(n\rho\pi_{\min}^2/r^2),$$
$$\epsilon_{over} = (1+B_{\max})\sqrt{r_T\log n} + B_{\max}r,$$

and Lemma S6 shows $\epsilon_{est} = 0$, w.h.p. For fixed $r, \pi_{\min}$, we have $\epsilon_{under} \gg \epsilon_{over}$. By Theorem 6, choosing $\Delta = (1+B_{\max})\sqrt{r_T\log n} + B_{\max}r$ leads to MATR-CV returning the correct r. We can further refine Δ by noting that $r_{\max} := \arg\max_{r_t}\langle A, \hat{X}_{r_t}\rangle \ge r$ w.h.p., then it suffices to consider the candidate range $\{r_1,\dots,r_{\max}\}$. The same arguments still hold for this range, thus r_T and r in Δ can be replaced with $r_{\max}$. □

Proposition S8. *Let $\tilde{X}$ be the optimal solution of SDP-2 for $A \sim SBM(B,Z)$. Suppose $\lambda \le O(\pi_{min}^2 n\min_{k\neq l}(B_{kk}-B_{kl})) - \Omega(\sqrt{\rho n\log n/\pi_{min}})$, and for every $k < r-1$,*

$$\begin{aligned}&\Omega(\sqrt{n\rho}) + \max\{m_{r-1}(B_{r-1,r-1}-B_{r-1,r}), m_r(B_{r,r}-B_{r,r-1})\} \le \lambda \\ &\le O(\pi_{\min}n\min_{k<r-1}(B_{k,k}+B^*_{r-1,r-1}-2B^*_{k,r-1})) - \Omega(\sqrt{\rho n\log n/\pi_{min}}),\end{aligned} \tag{S2.14}$$

then $\tilde{X} = X^*$ with high probability, where X^* is the normalized clustering matrix when the last two clusters are merged, and B^* is the $(r-1)\times(r-1)$ corresponding clustering probability matrix. $k,l < r-1$, $B^*_{kl} = B_{kl}$, $B^*_{k,r-1} = \frac{m_{r-1}B_{k,r-1}+m_KB_{k,r}}{m_{r-1}+m_r}$, $B^*_{r-1,r-1} = \frac{(m_{r-1}^2B_{r-1,K-1}+2*m_rm_{r-1}B_{r-1,r}+(m_r^2B_{r,r})}{(m_{r-1}+m_r)^2}$.

Proof of Proposition S8. We develop sufficient conditions with a construction of the dual certificate which guarantees X^* to be the optimal solution. The KKT conditions can be written as below:

First order stationary:

$$-A-\Lambda+(1\alpha^T+\alpha 1^T)+\beta I-\Gamma$$

Primal feasibility:

$$X\succeq 0, X\ge 0, X\mathbf{1}_n = \mathbf{1}_n, \text{trace}(X) = r$$

Dual feasibility:

$$\Gamma\ge 0, \Lambda\succeq 0$$

Complementary slackness

$$\langle \Lambda, X \rangle = 0, \Gamma \circ X = 0.$$

Consider the following construction: denote $T_k = C_k, n_k = m_k$, for $k < r-1$, $T_{r-1} = C_{r-1} \bigcup C_r, n_{r-1} = m_{r-1} + m_r$.

$$X_{T_k} = E_{n_k}/n_k, \quad X_{T_k T_l} = 0, \text{for } k \neq l \leq r-1$$

$$\Lambda_{T_k} = -A_{T_k} + (1_{n_k}\alpha_{T_k}^T + \alpha_{T_k}1_{n_k}^T) + \lambda I_{n_k}$$

$$\Lambda_{T_k T_l} = -(I - \frac{E_{n_k}}{n_k})A_{T_k T_l}(I - \frac{E_{n_l}}{n_l})$$

$$\Gamma_{T_k} = 0, \Gamma_{T_k,T_l} = -A_{T_k T_l} - \Lambda_{T_k T_l} + (1_{n_k}\alpha_{T_l}^T + \alpha_{T_k}1_{n_l}^T)$$

$$\alpha_{T_k} = \frac{1}{n_k}(A_{T_k}1_{n_k} + \phi_k 1_{n_k})$$

$$\phi_k = -\frac{1}{2}(\beta + \frac{1_{n_k}^T A_{T_k} 1_{n_k}}{n_k})$$

All the KKT conditions are satisfied by construction except for positive semidefiniteness of Λ and positiveness of Γ. Now, we show it one by one.

Positive semidefiniteness of Λ Since span$(1_{T_k}) \subset$ ker(Λ), it suffices to show that for any $u \in span(1_{T_k})^\perp, u^T\Lambda u \geq 0$. Consider $u = \sum_k u_{T_k}$, where $u_{T_k} := u \circ 1_{T_k}$, then $u_{T_k} \perp 1_{n_k}$.

$$\begin{aligned} u^T\Lambda u &= -\sum_k u_{T_k}^T A_{T_k} u_{T_k} + \lambda \sum_k u_{T_k}^T u_{T_k} - \sum_{k \neq l} u_{T_k}^T A_{T_k T_l} u_{T_l} \\ &= -\sum_k u_{T_k}^T (A-P)_{T_k} u_{T_k} - \sum_{k \neq l} u_{T_k}^T (A-P)_{T_k T_l} u_{T_l} + \lambda \|u\|_2^2 - u^T P u \\ &= -u^T(A-P)u + \lambda \|u\|_2^2 - u_{T_{r-1}}^T P_{T_{r-1}T_{r-1}} u_{T_{r-1}} \end{aligned} \tag{S2.15}$$

Now consider $u_{T_{r-1}}^T P_{T_{r-1}T_{r-1}} u_{T_{r-1}}$, and note that $T_{r-1} = C_{r-1} \bigcup C_r$, and

$$P_{T_{r-1}T_{r-1}} = \begin{bmatrix} B_{r-1,r-1}E_{m_{r-1}m_{r-1}}, B_{r-1,r}E_{m_{r-1}m_r} \\ B_{r,r-1}E_{m_r m_{r-1}}, B_{r,r}E_{m_r m_r} \end{bmatrix}$$

Since $u_{T_{r-1}} \perp 1_{n_{r-1}}$,

$$u_{T_{r-1}}^T \begin{bmatrix} B_{r-1,r}E_{m_{r-1}m_{r-1}}, B_{r-1,r}E_{m_{r-1}m_r} \\ B_{r,r-1}E_{m_r m_{r-1}}, B_{r,r-1}E_{m_r m_r} \end{bmatrix} u_{T_{r-1}} = 0,$$

therefore

$$\begin{aligned} u_{T_{r-1}}^T P_{T_{r-1}T_{r-1}} u_{T_{r-1}} &= u_{T_{r-1}}^T \begin{bmatrix} (B_{r-1,r} - B_{r-1,r-1})E_{m_{r-1}m_{r-1}}, 0 \\ 0, (B_{r-1,r} - B_{r,r})E_{m_r m_r} \end{bmatrix} u_{T_{r-1}} \\ &\leq \max\{m_{r-1}(B_{r-1,r-1} - B_{r-1,r}), m_r(B_{r,r} - B_{r,r-1})\} \|u\|_2^2 \end{aligned} \tag{S2.16}$$

Since $\|A - P\| \leq c_0\sqrt{np}$ provided $p \geq c_0 \log n/n$, Therefore, a sufficient condition is:

$$\lambda \geq \Omega(\sqrt{np_{max}}) + \max\{m_{r-1}(B_{r-1,r-1} - B_{r-1,r}), m_r(B_{r,r} - B_{r,r-1})\} \tag{S2.17}$$

Positiveness of Γ Define $d_i^*(T_k) = \sum_{j \in T_k} A_{i,j}$, $\bar{d}_i^*(T_k) = \frac{d_i^*(T_k)}{n_k}$, and $\bar{d}^*(T_k T_l) = \frac{\sum_{i \in T_l} \bar{d}_i^*(T_k)}{n_l}$. Then consider $x \in T_k$, $y \in T_l$, we need

$$\bar{d}_x^*(T_k) - \bar{d}_x^*(T_l) + \frac{1}{2}(\bar{d}^*(T_k T_l) - \bar{d}^*(T_k T_k)) + \bar{d}_y^*(T_l) - \bar{d}_y^*(T_k) + \frac{1}{2}(\bar{d}^*(T_k T_l) - \bar{d}^*(T_l T_l)) - \frac{\lambda}{2n_l} - \frac{\lambda}{2n_k} \geq 0,$$

Using Chernoff bound, for positiveness of Γ with high probability we only need

$$\frac{1}{2}(B^*_{kk}+B^*_{ll}-2B^*_{kl})-\sqrt{6\log n}(\sqrt{\frac{B^*_{kk}}{n_k}}+\sqrt{\frac{B^*_{ll}}{n_l}})-\sqrt{18B^*_{kl}\log n(\frac{1}{n_k}+\frac{1}{n_l})}\geq\frac{\lambda}{2n_l}+\frac{\lambda}{2n_k}$$

where for $k,l<r-1$,

$$B^*_{kl}=B_{kl},$$
$$B^*_{k,r-1}=\frac{m_{r-1}B_{k,r-1}+m_rB_{k,r}}{m_{r-1}+m_r},$$
$$B^*_{r-1,r-1}=\frac{(m^2_{r-1}B_{r-1,r-1}+2*m_rm_{r-1}B_{r-1,r}+(m^2_rB_{r,r})}{(m_{r-1}+m_r)^2}.$$

If $k,l<r-1$, then $B^*_{kl}=B_{kl},n_l=m_l$, the condition becomes

$$\frac{1}{2}(B_{kk}+B_{ll}-2B_{kl})-\sqrt{6\log n}(\sqrt{\frac{B_{kk}}{m_k}}+\sqrt{\frac{B_{ll}}{m_l}})-\sqrt{18B_{kl}\log n(\frac{1}{m_k}+\frac{1}{m_l})}\geq\frac{\lambda}{2m_l}+\frac{\lambda}{2m_k},$$

which is equivalent to

$$\lambda\leq O(\pi_{\min}n\min_{k\neq l}(B_{kk}-B_{kl}))-\Omega(\sqrt{\rho n\log n/\pi_{\min}}).$$

Now, suppose $k<r-1,l=r-1$, the condition becomes:

$$\begin{aligned}\frac{1}{2}(B_{kk}+B^*_{ll}-2B^*_{kl})-\sqrt{6\log n}(\sqrt{\frac{B_{kk}}{m_k}}+\sqrt{\frac{B^*_{ll}}{m_{r-1}+m_r}})\\-\sqrt{18B^*_{kl}\log n(\frac{1}{m_k}+\frac{1}{m_{r-1}+m_r})}\geq\frac{\lambda}{2m_{r-1}+2m_r}+\frac{\lambda}{2m_k}.\end{aligned}\tag{S2.18}$$

Since $\sqrt{6\log n}(\sqrt{\frac{B_{kk}}{m_k}}+\sqrt{\frac{B^*_{ll}}{m_{r-1}+m_r}})\frac{m_k(m_{r-1}+m_r)}{m_k+m_{r-1}+m_r}=O(\sqrt{\rho n\log n/\pi_{\min}})$, and similarly for other terms, then we have the sufficient condition for positiveness of Γ on λ:

$$\lambda\leq\frac{(m_km_{r-1}+m_km_r)(B_{k,k}+B^*_{r-1,r-1}-2B^*_{k,r-1})}{m_k+m_{r-1}+m_r}-\Omega(\sqrt{\rho n\log n/\pi_{\min}}).$$

$$\lambda\leq O(\pi_{\min}n\min_{k<r-1}(B_{k,k}+B^*_{r-1,r-1}-2B^*_{k,r-1}))-\Omega(\sqrt{\rho n\log n/\pi_{\min}}).$$

□

C. Proof of Theorem 10

The proof is similar in spirit to that of Theorem 8. First we have the following concentration result from the proof of Theorem 1 in Yan & Sarkar (2016):

$$\|\hat{S}-S\|_\infty=O(\sqrt{\frac{\log n}{d}})\tag{S3.19}$$

with probability at least $1-1/n$. This implies

$$|\langle\hat{S}-S,X\rangle|=O(n\sqrt{\frac{\log n}{d}})\tag{S3.20}$$

with probability $1-1/n$.

For notational convenience, we derive the underfitting and overfitting bounds for $\hat{S}$ and a general normalized clustering matrix $\hat{X}$, with the understanding that the same asymptotic bounds apply to estimates obtained from the training graph provided the split is random and the number of training nodes is $\Theta(n)$.

Lemma S9. *Recall that $p_{gap} = \min_k(a_{kk} - \max_{\ell\neq k} a_{k\ell}) > 0$, $\tau = n\pi_{\min}p_{gap}$. For a sequence of underfitting normalized clustering matrix $\{\hat{X}_{r_t}\}_{r_t<r}$, provided r, $\pi_{\min}$, p_{gap} are all fixed, and $d/\log n \to \infty$, we have w.h.p.*

$$\max_{r_t<r}\langle \hat{S}, \hat{X}_{r_t}\rangle \leq \langle \hat{S}, X_0\rangle - \Omega(\tau\pi_{min}/r^2).$$

Proof. By the same argument as in Eq (S2.8), for any underfitting $\hat{X}_{r_t}$,

$$\langle S, \hat{X}_{r_t}\rangle \leq \langle S, X_0\rangle - \Omega\left(\frac{\tau\pi_{\min}}{r_t^2}\right)$$

The above combined with Eq (S3.20) gives the desired result provided $d/\log n \to \infty$. □

Lemma S10. *For a sequence of overfitting normalized clustering matrix $\{\hat{X}_{r_t}\}_{r<r_t\leq r_T}$, $r_T = \Theta(r)$ and is fixed, we have w.h.p.*

$$\max_{r<r_t\leq r_T}\langle \hat{S}, \hat{X}_{r_t}\rangle \leq \langle \hat{S}, X_0\rangle + O(n\sqrt{\frac{\log n}{d}}).$$

Proof. By the same argument as in Eq (S2.11) (without the diagonal effect),

$$\langle S, \hat{X}_{r_t}\rangle \leq \langle S, X_0\rangle$$

for any overfitting $\hat{X}_{r_t}$. Combined with (S3.20), we have the desired result. □

Lemma S11. *With high probability, MATR-CV achieves exact recovery on the testing set given the true cluster number r, i.e. $\hat{X}_r^{22} = X_0^{22}$, provided $d/\log n \to \infty$, $\min_{k\neq\ell}\|\mu_k - \mu_\ell\| \geq \eta > 0$, $\max_{k\neq\ell}|a_{k\ell}| < \infty$, and $\gamma_{train} = \Theta(1)$.*

Proof. Denote $\{m_k^{11}\}$ as the cluster sizes and $\{\hat{m}_k^{11}\}$ as the estimated cluster sizes in the training set; $\{C_k^{11}\}$ as the true clusters, and $\{\hat{C}_k\}$ as the estimated clusters from the training set.

For a datapoint t in the testing set, assume it belongs to cluster C_k. The probability that it is assigned to cluster k is:

$$P(\frac{\sum_{j\in\hat{C}_k}\hat{S}_{t,j}^{21}}{\hat{m}_k^{11}} \geq \max_{l\neq k}\frac{\sum_{j\in\hat{C}_l}\hat{S}_{t,j}^{21}}{\hat{m}_l^{11}}).$$

We first note that

$$\frac{\sum_{j\in\hat{C}_k}(\hat{S}_{t,j}^{21} - S_{t,j}^{21})}{\hat{m}_k^{11}} \geq -c\sqrt{\frac{\log n}{d}} \tag{S3.21}$$

for all t, k with probability at least $1 - 1/n$ by Eq (S3.19). By Theorem 4 in Yan & Sarkar (2016),

$$|\hat{m}_k^{11}/m_k^{11} - 1| = O(\log d/d) \tag{S3.22}$$

for all k w.h.p., assuming r is fixed, and $\min_{k\neq\ell}\|\mu_k - \mu_\ell\| \geq \eta > 0$. Then in (S3.21),

$$\begin{aligned}\frac{\sum_{j\in\hat{C}_k}S_{t,j}^{21}}{\hat{m}_k^{11}} &= \frac{|\hat{C}_k\cap C_k^{11}|a_{kk} + \sum_{\ell\neq k}|\hat{C}_k\cap C_\ell^{11}|a_{k\ell}}{\hat{m}_k^{11}}\\ &\geq \frac{|\hat{C}_k\cap C_k^{11}|a_{kk} + |\hat{C}_k\cap \tilde{C}_k^{11}|\min_{\ell\neq k}a_{k\ell}}{\hat{m}_k^{11}}\\ &\geq a_{kk}\,(1 - O(\log d/d)) + O(\log d/d)\min_{\ell\neq k}a_{k\ell}\end{aligned} \tag{S3.23}$$

by Eq (S3.22), where $\tilde{C}_k^{11}$ denotes the complement of C_k^{11} in the training set. Using (S3.23) in (S3.21),

$$\frac{\sum_{j\in\hat{C}_k}\hat{S}_{t,j}^{21}}{\hat{m}_k^{11}} \geq a_{kk} - O(\log d/d) - c\sqrt{\frac{\log n}{d}} \tag{S3.24}$$

w.h.p. Similarly we can show w.h.p.

$$\frac{\sum_{j\in\hat{C}_\ell}\hat{S}^{21}_{t,j}}{\hat{m}^{11}_k} \le a_{k\ell} + O(\log d/d) + c\sqrt{\frac{\log n}{d}}. \tag{S3.25}$$

Then Eqs (S3.24) and (S3.25) imply the event

$$\frac{\sum_{j\in\hat{C}_k}\hat{S}^{21}_{t,j}}{\hat{m}^{11}_k} \ge \max_{l\neq k}\frac{\sum_{j\in\hat{C}_l}\hat{S}^{21}_{t,j}}{\hat{m}^{11}_l}$$

holds w.h.p. for large n, for $d/\log n \to \infty$. Since all the bounds are uniform in t, we have strong consistency with MATR-CV. □

Proof of Theorem 10. Now the proof directly comes from applying Theorem 6 and the results in Lemma S9-S11. We have $\epsilon_{\text{under}} = \Omega(\tau\pi_{\min}/r^2) = \Omega(n)$ since all the other parameters have constant order, $\epsilon_{\text{over}} = O(n\sqrt{\log n/d})$, and $\epsilon_{\text{est}} = 0$. Thus it suffices to choose $\Delta = n\sqrt{\frac{(\log n)^{1.1}}{d}}$. □

D. Detailed parameter settings in experiments and additional results

D.1. Details on motivating examples in Section 3 (Figure 1)

Figure 1(a): We generate an adjacency matrix from a SBM model with four communities, each having 50 nodes, and

$$B = \begin{bmatrix} 0.8 & 0.6 & 0.4 & 0.4 \\ 0.6 & 0.8 & 0.4 & 0.4 \\ 0.4 & 0.4 & 0.8 & 0.6 \\ 0.4 & 0.4 & 0.6 & 0.8 \end{bmatrix}.$$

The visualization of the underlying probability matrix is shown in Figure 1(a).

Figure 1(b): We consider a four-component Gaussian mixture model, where the means $\mu_1, \ldots, \mu_4$ are generated from Gaussian distributions centered at $(0,0), (0,0), (5,5), (10,10)$ with covariance $6I$, so that the first two clusters are closer to each other than the rest. Then we generate 1000 data points centered at these means with covariance $0.5I$, each point assigned to one of the four clusters independently with probability $(0.48, 0.48, 0.02, 0.02)$. Finally, we introduce correlation between the two dimensions by multiplying each point by $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. A scatter plot example of the datapoints is shown in Figure 1(c).

D.2. Additional results on real data in Section 5.2

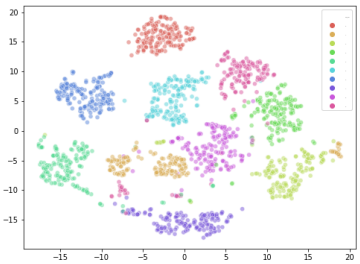
In Figure S1, we show the 2D projection of the handwritten digits dataset using tSNE and color the points by the clusters they belong to according to each method. We see that as indicated by the NMI scores reported in the main paper, the clustering by MATR, MST and KNN correspond more closely to the true clustering than DS.

D.3. Additional settings and results on simulated data in Section 5.3 (Figure 5)

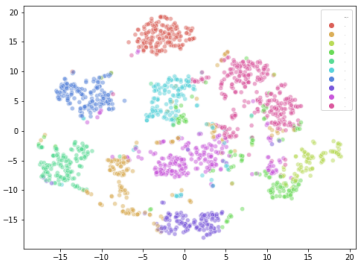
We first consider graphs generated from a SBM with equal sized clusters, where

$$B = \rho \times \begin{bmatrix} 0.8 & 0.5 & 0.3 & 0.3 \\ 0.5 & 0.8 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.8 & 0.5 \\ 0.3 & 0.3 & 0.5 & 0.8 \end{bmatrix}.$$

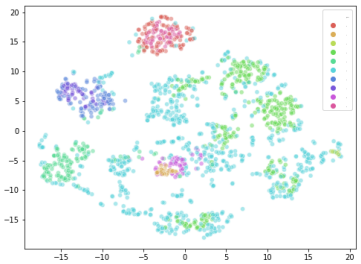
In Table S1 (a,b), we show the median number of clusters selected by each method as ρ changes for both the equal sized and unequal sized cases described above. The ground truth is 4 clusters.



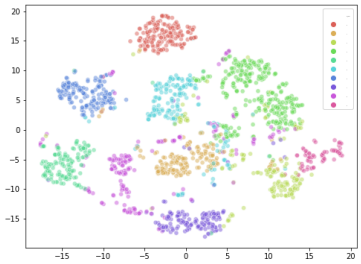
(a) True clustering



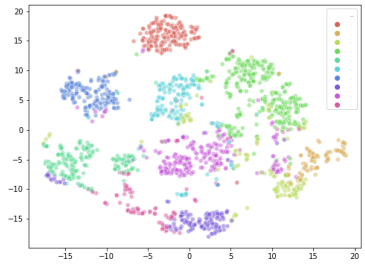
(b) Clustering by MATR



(c) Clustering by DS



(d) Clustering by KNN



(e) Clustering by MST

Figure S1: Visualization of clustering results on handwritten digits dataset.

ρ	MATR-CV	BH	ECV
0.2	2	2	2
0.3	2	2	2
0.4	4	2	2
0.5	4	2	2
0.6	4	4	2

(a) Median number of clusters selected for equal sized case

ρ	MATR-CV	BH	ECV
0.2	2	2	2
0.3	2	2	2
0.4	3	2	2
0.5	4	2	2
0.6	4	3	2

(b) Median number of clusters selected for unequal sized case

Table S1: Comparison of model selection results along with ρ for all algorithms.

D.4. Additional results on simulated data in Section 5.4

In Table S2, we compare model selection results obtained from different methods on simulated mixture of Gaussian data. The same setting as in Figure 2(c)-(d) is used (described in Section D.2) with $r = 4$ and equal probabilities of cluster assignment. MATR-CV performs similarly as GAP but better than SIL.

separation	1	1.5	2.2	3.3	5.0
MATR-CV	1	1	1	1	1
GAP	1	1	1	1	1
SIL	0	0	1	1	1

Table S2: Exact recovery fractions for balanced 4 clusters

References

Lei, J. and Rinaldo, A. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 02 2015. doi: 10.1214/14-AOS1274. URL https://doi.org/10.1214/14-AOS1274.

Skala, M. Hypergeometric tail inequalities: ending the insanity. *arXiv preprint arXiv:1311.5939*, 2013.

Yan, B. and Sarkar, P. On robustness of kernel clustering. In *Advances in Neural Information Processing Systems*, pp. 3098–3106, 2016.

Yan, B., Sarkar, P., and Cheng, X. Provable estimation of the number of blocks in block models. *arXiv preprint arXiv:1705.08580*, 2017.