A APPENDIX

A.1 FURTHER RESPONSE TO REVIEWERS

Sample complexity bound. To verify the sample complexity bound for LA-SDP in Theorem 2 ($O(\log(n))$) is tight, we will change n and adjust the squared distance between clusters by multiplying $\log(n)$. More precisely, we let $d = \lambda\sqrt{\log(n)},\ \lambda > 0$. The diagonal of the covariance matrices are placed at a simplex of $\mathbb{R}^p$ that are not identical to the corresponding centers. i.e. $\mu_k = \lambda \cdot e_k$, $\Sigma_k = L \cdot \text{diag}(e_{k+1}),\ \forall l \in [K]$, where $e_{K+1} = e_1$. This guarantees the symmetry of the construction. We set $L = 10,\ p = 4,\ K = 4$. Each time we draw the $n = 120/240/480$ data from the GMM. The results of the simulation for the second plot in Figure 5 are obtained through 20 total replicates, where we can observe the same pattern across different settings for n. This shows that the order $\log(n)$ for separation bound in Theorem 2 should be tight.

Figure 5: Mis-clustering error (with shaded error bars for the left plot) vs λ for iLA-SDP for different n.

Computational complexity for banknote authentication dataset. Now if we look at the results of time cost for clustering banknote authentication dataset in Table 1, we can observe that the time cost for iLA-SDP is relatively high and to reduce the time cost, we could consider sub-sampling methods, e.g., the subsampling idea (Zhuang et al., 2022b). This will be set as our future goal.

Table 1: Time cost (SD) for clustering banknote authentication dataset for 20 replicates.

EM	KM	iLA-SDP	SC
0.1719 (0.0853)	0.0013 (0.0013)	2100 (1882)	0.0395 (0.0959)

A.2 ENHANCED ILA-SDPS FOR HIGH-DIMENSIONAL AND LARGE-SIZE DATA

In this section, we propose two variations of iLA-SDP that can handle high-dimensional and large-size data with better computational and statistical efficiency.

High dimensional data. If the number attributes of the data are large, it would be hard to approximate the true covariance matrices since there are $O(p^2)$ many unknown parameters. Thus, we propose two dimension reduction procedures that based on hierarchical clustering, Fisher's LDA and F-test. The detailed algorithm have been shown in Algorithm 2 and Algorithm 3. To reduce the dimension, we proposed two procedure.

1. If the number of clusters K is small and the difference between centers are sparse, we shall use HC as a benchmark method for feature selection and assume the group means according to HC

as ground true. Specifically, for i-th attribute, we calculate the F-statistics and its p-value based on the H_0 that all group means w.r.t. i-th attribute are the same. At last, each attribute would likely to be selected if the p-value $\mathcal{P}_i$ for i-th attribute is significantly small among p-values for all attributes.

2. First we use the hierarchical clustering to get the clustering results for all possible input cluster number $\tilde{K} \in [p]$. If we assume all the clusters have identical covariance matrices, then we may use the assignments from HC to estimate the within-cluster covariance $\tilde{W}$ (with group means $\tilde{\mu}_l$) and get the signal-to-noise ratio $\Delta(\tilde{K}) := \min_{k \neq l} \|\tilde{W}^{-1/2}(\tilde{\mu}_k - \tilde{\mu}_l)\|$. Here, HC serves as a benchmark method for data initial processing. We will then choose the largest $\tilde{K}$ within target range such that the signal-to-noise ratio $\Delta(\tilde{K})$ is maximized. Then it will lead to the new dataset with dimension $q = \tilde{K} - 1$ after running Fisher's LDA on the assignments from HC with clusters number equals $\tilde{K}$. Finally we perform Algorithm 1 on the new dataset and extract the cluster labels.

Large-size data. As we know that the time complexity for solving SDP is as high as $O(n^{3.5})$. We might use subsampling methods to bring down the time cost while maintain the superior behavior for LA-SDP (Zhuang et al., 2022b). The proposed algorithm is shown in Algorithm 4.

Algorithm 2: Likelihood adjusted SDP based iterative algorithm with unknown covariance matrices $\Sigma_1, \ldots, \Sigma_K$ for large p.

Input: Data matrix $X \in \mathbb{R}^{p \times n}$ containing n points. Cluster numbers K. The stopping criterion parameters p_0, ϵ and S. $\alpha \in [0, 1],\ C > 0$.

1 Run hierarchical clustering with data X, clusters number K and extract the cluster labels $G_1^{(0)}, \ldots, G_K^{(0)}$ as prior assignments for $[n]$. Suppose the assignments have true centers $\mu_k^{(0)},\ k \in [K]$.

2 **for** $i = 1, \ldots, p$ **do**

3 Calculate the p-value $\mathcal{P}_i$ of the F-test $\mathcal{F}_i$ under H_0: $\mu_{1,i}^{(0)} = \cdots = \mu_{K,i}^{(0)}$, where $\mu_{k,i}^{(0)}$ corresponds to the i-th component of $\mu_i^{(0)}$.

4 Keep p_0 attributes with p_0 smallest p-values $\mathcal{P}_i$.

5 **if** *there is no clear cutoff between $\mathcal{P}_i$'s, i.e.* $\max_{i \in [p]} \mathcal{P}_i / \min_{i \in [p]} \mathcal{P}_i < C$, **then**

6 we further keep other $p - p_0$ attributes with probability $\alpha > 0$.

7 Get dimension reduced data $\tilde{X}$.

8 Run Algorithm 1 on $\tilde{X}$ with initialization obtained from K clusters of HC and stopping criterion parameters ϵ and S. Then extract the cluster labels $\hat{G}_1, \ldots, \hat{G}_K$ as a partition estimate for $[n]$.

Output: A partition estimate $\hat{G}_1, \ldots, \hat{G}_K$ for $[n]$.

A.3 Experiment results

In this section, we provide more details of the settings and post the results for simulation experiments. For all the dimension reduction procedures used in the simulation experiments, we perform step 1-7 in Algorithm 2 followed by Algorithm 3 with input parameters $\alpha = 0.7,\ C = 10^{10},\ p_0 = 2K,\ p_1 = 15\ \epsilon = 10^{-2},\ S = 50$. The initialization we use is hierarchical clustering from *mclust* package in R. Here we test our algorithm on Gaussian mixture models and real datasets. We compared our algorithm iLA-SDP (HC as initialization) with HC, EM algorithm (HC as initialization), K-means (HC as initialization) and original SDP.

Improvements of iLA-SDP over SDP. Recall in Theorem 2, we define the signal-to-noise ratio as $\Delta^2 := \min_{k \neq l} \|\Sigma_k^{-1/2}(\mu_k - \mu_l)\|^2$. To verify the validity of the definition and compare iLA-SDP and SDP, we change the conditional number for covariance matrices $\Sigma_1, \ldots, \Sigma_K$. Here we choose $n = 200,\ p = 4,\ K = 4$. Recall $M := \max_{k \neq l} \|\Sigma_l^{1/2} \Sigma_k^{-1} \Sigma_l^{1/2}\|_{\text{op}}$, we choose all the covariance matrices to be the same such that M is fixed. The covariance matrices are set to be identity matrix except that the first entry at the diagonal are set to be $L + 1$, which refers to the condition number of matrices. We consider two cases where $L = 10,\ 100$. Now denote $e_k \in \mathbb{R}^p$ as the vector with k-th

Algorithm 3: Likelihood adjusted SDP based iterative algorithm with unknown covariance matrices $\Sigma_1, \ldots, \Sigma_K$ for large p.

Input: Data matrix $X \in \mathbb{R}^{p\times n}$ containing n points. Cluster numbers K. The stopping criterion parameters p_1, ϵ and S.

Select a bench mark clustering method (HC) as a way to provide a prior assignments.

for $\tilde{K} = K, K+1, \ldots, p_1 - 1, p_1$ **do**

  Run hierarchical clustering with data X, clusters number $\tilde{K}$ and extract the cluster labels $G_1^{(\tilde{K})}, \ldots, G_{\tilde{K}}^{(\tilde{K})}$ as prior assignments for $[n]$ and get the group means $\mu_k^{(\tilde{K})}, k \in [\tilde{K}]$.

  Calculate the within-cluster covariance matrix W, then get the signal-to-noise ratio $\Delta(\tilde{K}) := \min_{l\neq k} \|W^{-1/2}(\mu_l^{(\tilde{K})} - \mu_k^{(\tilde{K})})\|$.

Choose K^* such that $\Delta(K^*)$ is maximized for $K^* = K, K+1 \ldots, P-1, P$.

Perform the Fisher's LDA with data X, assignments $G_1^{(K^*)}, \ldots, G_{K^*}^{(K^*)}$ and get the transformed data $\tilde{X} \in \mathbb{R}^{q\times n}$ with $q = K^* - 1$.

Run Algorithm 1 on $\tilde{X}$ with initialization obtained from K clusters of HC and stopping criterion parameters ϵ and S. Then extract the cluster labels $\hat{G}_1, \ldots, \hat{G}_K$ as a partition estimate for $[n]$.

Output: A partition estimate $\hat{G}_1, \ldots, \hat{G}_K$ for $[n]$.

Algorithm 4: Sketch and lift: Likelihood adjusted SDP based iterative algorithm with unknown covariance matrices $\Sigma_1, \ldots, \Sigma_K$ for large n.

Input: Data matrix $X \in \mathbb{R}^{p\times n}$ containing n points. Cluster numbers K. The stopping criterion parameters P, ϵ and S. Sampling weights $(w_1, \ldots, w_n)$ with $w_1 = \cdots = w_n = \gamma \in (0,1)$ being the subsampling factor.

(Sketch) Independent sample an index subset $T \subset [n]$ via Ber(w_i) and store the subsampled data matrix $V = (X_i)_{i\in T}$.

Run subroutine Algorithm 1 with input V to get a partition estimate $\hat{R}_1, \ldots, \hat{R}_K$ for T.

Compute the centroids $\bar{X}_k = |\hat{R}_k|^{-1}\sum_{j\in\hat{R}_k} X_j$ and within-group sample covariance matrices $\hat{\Sigma}_k = |\hat{R}_k|^{-1}\sum_{j\in\hat{R}_k}(X_j - \bar{X}_k)(X_j - \bar{X}_k)^T$ for $k \in [K]$.

(Lift) For each $i \in [n] \setminus T$, assign $i \in \hat{G}_k$ if $\log|\hat{\Sigma}_k| + \|\hat{\Sigma}_k^{-1/2}(X_i - \bar{X}_k)\|^2 < \log|\hat{\Sigma}_l| + \|\hat{\Sigma}_l^{-1/2}(X_i - \bar{X}_l)\|^2, \quad \forall l \neq k,\ l \in [K]$. And randomly assign i to any K clusters if such k doesn't exist.

Output: A partition estimate $\hat{G}_1, \ldots, \hat{G}_K$ for $[n]$.

entry as 1, and 0 otherwise. The centers of clusters $\mu_1, \ldots, \mu_K$ are placed on vertices of a regular simplex, i.e., $\mu_k = \lambda\sqrt{1 + (1+L)^{-1}}e_k,\ k \in [K]$. This ensures that for any L, $\Delta = \lambda,\ \forall\lambda$. From Figures 2 we can observe that the signal-to-noise ratio we defined is reasonable. On the other hand, the performance of SDP becomes worse as condition number of the group covariance matrices grows since the assumption of isotropy group covariance matrices for SDP is violated and same reason for K-means.

Impact of dimension reduction. Here we want to see the performance of iLA-SDP after dimension reduction. The covariance matrices of GMM are drawn independently following $\Sigma_k := U_k\Lambda_k U_k^T,\ \forall k \in [K]$. Here U_k is a random orthogonal matrix, Λ_k is a diagonal matrix with entries drawn from $\mathcal{Z} = 1 + \beta Z \cdot \mathbf{1}(Z > 0)$, where Z is standard Gaussian distribution, $\beta > 0$ controls the condition number of Σ_k. Here we choose $n = 200,\ p = 20,\ K = 4,\ \beta = 5$. The covariance matrices are fixed once chosen and we perform Algorithm 1 on the dataset directly to get the results of iLA-SDP for each replicates. For dimension reduction, we follow the procedure of dimension reduction introduced in Algorithm 2 and Algorithm 3 in Appendix A.2 and get the transformed dataset $\tilde{X}$ with lower dimension. Then the results of pLA-SDP is obtained from running Algorithm 1 with HC as initialization on $\tilde{X}$. The results in Figure 6 shows that after reduction of dimension in our procedure, the performance of iLA-SDP becomes significantly better when the separation is large. This is because in our setting, the difference between centers $d_{(k,l)} := \mu_k - \mu_l$, is

sparse for all distinct pairs. And after performing the F-test on the covariates, the noisy terms get eliminated which results in better performance.

Failure of EM vs SDP. Recall the failure of EM for random initialization (Jin et al., 2016) in the special case that covariance matrices equal to identity matrix and it assumes equal weights. Both covariance matrices and weights are known. In this case, EM algorithm would be reduced to the version that the weights and the mean update interactively. Meanwhile, iLA-SDP would be reduced to SDP. The random initialization indicates that we pick any data point as initialization of the centers uniformly. Following the same setting from the construction of the pitfall, we choose one dimension GMM with three clusters such that the distance between two of the centers is much smaller than others. More concisely, we let $n = 300,\ K = 3,\ p = 1,\ \mu_1 = \gamma,\ \mu_2 = -\gamma,\ \mu_3 = 10 \cdot \gamma$. The results can be observed from the first plot in Figure 3 with 300 replicates, where we denote the reduced version of EM as mEM. From the figure we can observe that the reduced version of iLA-SDP, which is SDP, performs stable and achieves exact recovery when the separation is large. However, EM would fail for random initialization.

Perturbation of initialization assignments. To see how the performance of EM and iLA-SDP will change when perturbing the initialization, we set HC as initialization and proportion α ($\alpha \in [0, 1]$) of the initialization labels will be perturbed. The diagonal of the covariance matrices are placed at a simplex of $\mathbb{R}^p$ that are not identical to the corresponding centers. i.e. $\mu_k = \lambda \cdot e_k$, $\Sigma_k = L \cdot \text{diag}(e_{k+1}),\ \forall l \in [K]$, where $e_{K+1} = e_1$. This guarantees the symmetry of the construction. We set $L = 10,\ p = 4,\ K = 4$ and the distance between centers $d = 8$. Each time we draw the $n = 200$ data from the GMM and run HC as initialization. Then we randomly assign α proportion of the labels from HC to any cluster uniformly. The results of the simulation for the second plot in Figure 3 are obtained through 300 total replicates, where we can observe that iLA-SDP is fairly stable with perturbation of initialization if the separation is large while EM can go worse as α approaches 1, i.e., all the labels are selected randomly. In other words, EM is more sensitive to initialization and iLA-SDP is more stable if the signal is strong.

Empirical evidence for monotone increasing of objective function for iLA-SDP. Here we provide examples based on previous experiment settings where we set the distance between centers $d = 1/3/5/10$. and try to see how the log-likelihood function of given data changes as the iteration proceeds. From Figure 7 in Appendix we can see that our algorithm guarantees that the log-likelihood function of given data increases over iteration empirically. What is more, by our construction we can show that the log-likelihood function will increase after each step for iLA-SDP theoretically.

Figure 6: Mis-clustering error (with shaded error bars for the left plot) vs center distance D for iLA-SDP before and after dimension reduction. pLA-SDP denotes the iLA-SDP after dimension reduction.

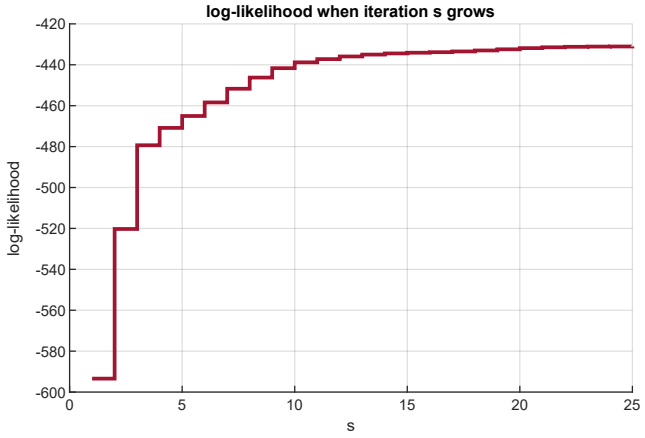

Figure 7: Log-likelihood (up to some constant) as iteration s grows for iLA-SDP.

A.4 PROOF OF THE THEOREMS AND PROPOSITIONS

In this section, we provide the proofs for the Proposition 1, Proposition 4 and a sketch proof of Theorem 2. The proof of the main theorem follows the track from the paper solving the exact recovery for original SDP (Chen & Yang, 2021b) and we will show the main differences in our proof.

First, we provide explicit expressions of some constants appearing in Theorem 2 below:

$$E_1 = \frac{4(1+2\delta)M^{5/2}}{(1-\beta)^2\eta^2}\left(M + \sqrt{M^2 + \frac{(1-\beta)^2}{(1+\delta)}\frac{p}{m\log n} + C_4 R_n}\right)$$

with

$$R_n = \frac{(1-\beta)^2}{(1+\delta)\log n}\left(\frac{\sqrt{p\log n}}{\underline{n}} + \frac{\log n}{\underline{n}}\right),$$

and

$$\begin{aligned} E_2 = &\frac{C_5(M-1)^3M^2}{(1-\beta)(1-\eta)}\left(\frac{p}{\log n}+1\right) + \frac{C_6K^2(1-\beta)}{\beta} \\ &\cdot \min\left\{\frac{1}{\beta(M-1)^2}\frac{n}{m}\left(1+\frac{\log p}{\log n}\right)\frac{p}{\log n},\ \frac{(M-1)M^2}{\beta}\left(\sqrt{\frac{p^3}{\log n}} + \sqrt{p\log n}\right)\frac{n}{\sqrt{m}}\right\}. \end{aligned} \tag{15}$$

A.4.1 Proof of Proposition 1

Proposition 1 (**SDP relaxation for K-means is a special case of LA-SDP**). Suppose $\Sigma_k = \sigma^2\mathrm{Id}_p$ for all $k \in [K]$. Let $\hat{Z}$ be the solution to (5) that achieves maximum M_1 and $\hat{Z}_k, k = 1, \ldots, K$, be the solution to (5) with maximum M_2. Then $M_1 = M_2$. And $\hat{Z} = \sum_{k=1}^K \hat{Z}_k$, if $\hat{Z}$ is unique in (5).

Proof of Proposition 1 If $\Sigma_k = \sigma^2\mathrm{Id}_p,\ \forall k \in [K]$. Then from (7) we have

$$A_k \equiv \frac{1}{2}\left[\mathrm{diag}(X^TX)\mathbf{1}_n^T + \mathbf{1}_n\mathrm{diag}(X^TX)^T\right] + X^TX,\ \forall k \in [K].$$

This implies that (8) can be written as

$$\begin{aligned} \hat{Z}_1, \ldots, \hat{Z}_K = \arg\max_{Z_1,\ldots Z_K \in \mathbb{R}^{n\times n}} &\left\langle X^TX, \left(\sum_{k=1}^K Z_k\right)\right\rangle \\ \text{subject to } Z_k \succeq 0,\ \mathrm{tr}\left(\sum_{k=1}^K Z_k\right) = K,\ &\left(\sum_{k=1}^K Z_k\right)\mathbf{1}_n = \mathbf{1}_n,\ Z_k \geqslant 0,\ \forall\ k \in [K], \end{aligned} \tag{16}$$

Since $\left\langle \mathrm{diag}(X^TX)\mathbf{1}_n^T, \left(\sum_{k=1}^K Z_k\right)\right\rangle = \mathrm{tr}(X^TX)$, which is a constant in the optimization problem (16). Now suppose $\hat{Z}$ is a solution to (5) that achieves maximum M_1 and $\hat{Z}_k, k = 1, \ldots, K$, is the solution to (16) that achieves maximum M_2, then we have

$$\left\langle X^TX, \left(\sum_{k=1}^K Z_k\right)\right\rangle \leq M1,$$

$$\left\langle X^TX, \left(\sum_{k=1}^K \tilde{Z}_k\right)\right\rangle \leq M2,$$

where $\tilde{Z}_1 := \hat{Z},\ \tilde{Z}_2 = \cdots = \tilde{Z}_K = 0$. In other words, $M_1 = M_2$, which finishes the proof. If $\hat{Z}$ is unique in (5), then we have $\hat{Z} = \sum_{k=1}^K \hat{Z}_k$ since both of them achieve the maximum in (5). ■

A.4.2 Proof of Proposition 4

Proposition 4 (**iLA-SDP is a soft clustering method**). If $\mathrm{rank}(Z_k) = 1$, then there exists weights $(w_{k,1}, \ldots, w_{k,n})$ such that $\hat{\Sigma}_k$ in Lemma 3 can be written as

$$\hat{\Sigma}_k := \frac{1}{n_k}\sum_{i=1}^n w_{k,i}(X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^\top \quad \text{with} \quad \hat{\mu}_k := \frac{1}{n_k}\sum_{i=1}^n w_{k,i}X_i, \tag{17}$$

where $n_k = \sum_{i=1}^n w_{k,i}$.

Proof of Proposition 4 If Z_k is rank 1, then there exists $a \in \mathbb{R}^n$ such that $Z_k = aa^T$. Let $w_k := a^T\mathbf{1} \cdot a$, then we have

$$Z_k = \frac{w_k w_k^T}{w_k^T \mathbf{1}},$$

i.e., $Z_{k,ij} = \frac{w_{k,i}w_{k,j}}{\sum_{i=1}^n w_{k,i}}$. Finally, by plugging in the expression of $Z_{k,ij}$ with $w_{k,i}$ we can get the target expression for $\hat{\Sigma}_k$. ■

A.4.3 SKETCH PROOF OF THEOREM 2

Theorem 2 (**Exact recovery for LA-SDP**). Suppose there exist constants $\delta > 0$ and $\beta \in (0, 1)$ such that

$$\log n \geq \max\left\{\frac{(1-\beta)^2}{\beta^2}, \frac{(1-\beta)(1-\eta)K^2}{\beta^2 \max\{(M-1)^2, 1\}}\right\}\frac{C_1 n}{m},\ \delta \leq \frac{\beta^2}{(1-\beta)^2}\frac{C_2 M^{1/2}}{K},\ m \geq \frac{4(1+\delta)^2}{\delta^2}.$$

If

$$\Delta^2 \geq (E_1 + E_2)\log n, \text{ and } \min_{k\neq l} D_{(k,l)} \geq C_3(1 + \log n/p + p/n), \tag{18}$$

where

$$E_1 = \frac{4(1+2\delta)M^{5/2}}{(1-\beta)^2\eta^2}\left(M + \sqrt{M^2 + \frac{(1-\beta)^2}{(1+\delta)}\frac{p}{m\log n} + C_4 R_n}\right)$$

with

$$R_n = \frac{(1-\beta)^2}{(1+\delta)\log n}\left(\frac{\sqrt{p\log n}}{n} + \frac{\log n}{n}\right),$$

and

$$\begin{aligned} E_2 &= \frac{C_5(M-1)^3M^2}{(1-\beta)(1-\eta)}\left(\frac{p}{\log n} + 1\right) + \frac{C_6K^2(1-\beta)}{\beta} \\ &\cdot \min\left\{\frac{1}{\beta(M-1)^2}\frac{n}{m}\left(1 + \frac{\log p}{\log n}\right)\frac{p}{\log n},\ \frac{(M-1)M^2}{\beta}\left(\sqrt{\frac{p^3}{\log n}} + \sqrt{p\log n}\right)\frac{n}{\sqrt{m}}\right\}; \end{aligned} \tag{19}$$

then the LA-SDP achieves exact recovery, or $\hat{Z} = Z^*$, with probability at least $1 - C_7K^3n^{-\delta}$ for some universal constants $C_1, \ldots, C_7$.

Sketch of the proof. Recall that we let $G_1^*, \ldots, G_K^*$ be the true partition of the index set $[n] := \{1, \ldots, n\}$ such that if $i \in G_k^*$, then

$$X_i = \mu_k + \epsilon_i, \tag{20}$$

where $\mu_k \in \mathbb{R}^p$ is the true center of the k-th cluster G_k^* (G_k for simplicity) and ϵ_i is an i.i.d. random Gaussian noise $N(0, \Sigma_k)$. First we can write down the dual problem:

$$\min_{\substack{\lambda\in\mathbb{R},\alpha\in\mathbb{R}^n,\\ B_k\in\mathbb{R}^{n\times n}}} \lambda K + \alpha^T\mathbf{1}_n,\ \text{subject to } B_k \geq 0,\ \lambda \mathrm{Id}_n + \frac{1}{2}(\alpha\mathbf{1}_n^T + \mathbf{1}_n\alpha^T) - A_k - B_k \succeq 0,\ \forall k \in [K].$$

Denote $Z_k^* := \frac{1}{|G_k|}\mathbf{1}_{G_k}\mathbf{1}_{G_k}^T,\ \forall k \in [K]$ then it can be shown that the sufficient conditions for the solution of SDP to be $Z_k = Z_k^*,\ \forall k \in [K]$ are

$$B_k \geq 0; \tag{C1}$$

$$W_k := \lambda \mathrm{Id}_n + \frac{1}{2}(\alpha\mathbf{1}_n^T + \mathbf{1}_n\alpha^T) - A_k - B_k \succeq 0; \tag{C2}$$

$$\mathrm{tr}(W_k Z_k^*) = 0; \tag{C3}$$

$$\mathrm{tr}(B_k Z_k^*) = 0. \tag{C4}$$

It can be verified that if we can find symmetric B_k such that

$$B_{k,G_kG_k} = 0;$$

$$
\begin{aligned}
[B_{k,G_lG_k}\mathbf{1}_{G_k}]_i &= -\frac{n_k+n_l}{2n_l}\cdot\lambda \\
&\quad + \frac{n_k}{2}[(\|\Sigma_k^{-1/2}(\bar{X}_k - X_i)\|^2 + \log|\Sigma_k|) - (\|\Sigma_l^{-1/2}(\bar{X}_l - X_i)\|^2 + \log|\Sigma_l|)]; \\
[B_{k,G_lG_l}\mathbf{1}_{G_l}]_j &= [A_{l,G_lG_l}\mathbf{1}_{G_l}]_j - [A_{k,G_lG_l}\mathbf{1}_{G_l}]_j; \\
[B_{k,G_{l'}G_l}\mathbf{1}_{G_l}]_j &= [B_{l,G_{l'}G_l}\mathbf{1}_{G_l}]_j + [A_{l,G_{l'}G_l}\mathbf{1}_{G_l}]_j - [A_{k,G_{l'}G_l}\mathbf{1}_{G_l}]_j,
\end{aligned}
$$

for any triple pairs (k,l,l') that are mutually distinct and $i \in G_k,\ j \in G_l$. Then (C3) and (C4) hold. In fact, the target matrices can be defined through

$$
B^{\#}_{k,G_{l'}G_l} := \frac{B_{k,G_{l'}G_l}\mathbf{1}_{G_l}\mathbf{1}_{G_{l'}}^T B_{k,G_{l'}G_l}}{\mathbf{1}_{G_{l'}}^T B_{k,G_{l'}G_l}\mathbf{1}_{G_l}}, \tag{21}
$$

for any $k \in [K],\ (l',l) \neq (k,k)$. Furthermore, the construction of B_k shows that $B_k\mathbf{1}_{G_l} = 0,\ \forall(k,l)$ pairs.

The following two lemma gives the sufficient conditions for (C1).

Lemma 6 (**Separation bound on the covariance matrices**). Let $\lambda_1,\ldots,\lambda_p$ correspond to the eigenvalues of $(\Sigma_l^{1/2}\Sigma_k^{-1}\Sigma_l^{1/2} - \mathrm{Id}_p)$ and define $D_{(k,l)} := \frac{\sum_{i=1}^p(\lambda_i - \log(1+\lambda_i))}{p\max_i|\lambda_i|}$. If there exists constant C such that

$$
\min_{k\neq l} D_{(k,l)} \geq C(1 + \log n/p + p/n),
$$

then

$$
\mathbb{P}\Big([A_{l,G_lG_l}\mathbf{1}_{G_l}]_j - [A_{k,G_lG_l}\mathbf{1}_{G_l}]_j \geq 0,\ \text{for all } (k,l)\in[K]^2 \text{ and } j\in G_l\Big) \geq 1 - CK^2/n.
$$

Lemma 7 (**Separation bound on the centers**). Let $\delta > 0,\ \beta \in (0,1),\ \eta \in (0,1)$. If we have

$$
\Delta^2 \geq \frac{4(1+\delta)M^2}{(1-\beta)^2\eta^2}\left[M^{3/2} + \sqrt{M^3 + \frac{(1-\beta)^2M}{(1+\delta)}\frac{p + 2\sqrt{p\log(nK)} + 4\log(nK)}{m\log n}}\right]\log n,
$$

and

$$
\begin{aligned}
\Delta^2 &\geq \frac{M^2(M-1)^2}{(1-\beta)^2(1-\eta)^2}\cdot \\
&\left(1 + \frac{2(1-\beta)(1-\eta)}{M}[3\log M + 4M(M-1)(p + 2\sqrt{p\log(nK)} + 4\log(nK))]\right),
\end{aligned}
$$

then

$$
\begin{aligned}
&\mathbb{P}\Big(\|\Sigma_l^{-1/2}(\bar{X}_l - X_j)\|^2 + \log|\Sigma_l|) - (\|\Sigma_{l'}^{-1/2}(\bar{X}_{l'} - X_j)\|^2 + \log|\Sigma_{l'}|) \\
&- \frac{2}{n_l}\big|[A_{l,G_{l'}G_l}\mathbf{1}_{G_l}]_j - [A_{k,G_{l'}G_l}\mathbf{1}_{G_l}]_j\big| \geq \frac{\beta}{M}\|\Sigma_l^{-1/2}(\mu_l - \mu_{l'})\|^2 + (n_l^{-1} + n_{l'}^{-1})p - r_{k,l,l'}, \\
&\text{for all triple } (k,l,l') \in [K]^3 \text{ with } (k,l,l') \neq (k,k,k) \text{ and } j \in G_{l'}\Big) \\
&\leq \frac{CK^3}{n^\delta},
\end{aligned}
$$

where

$$
r_{k,l,l'} = 4\sqrt{\frac{\log(nK)}{n_l}}\|\Sigma_l^{-1/2}(\mu_l - \mu_{l'})\| + 2(n_l^{-1} + n_{l'}^{-1})\sqrt{2p\log(nK)} + 4n_{l'}^{-1}\log(nK).
$$

for some large constant C.

The proof of Lemma 7 follows the similar steps from the original paper (Chen & Yang, 2021b). The two lemmas imply that (C1) can hold with high probability if the separation condition in the assumption holds. The remaining part is to verify the (C2).

Denote $\Gamma = \text{span}\{\mathbf{1}_{G_k} : k \in [K]\}^{\perp}$ be the othogonal complement of the linear space spanned by $\mathbf{1}_{G_k},\ k \in [K]$. Note that $W_k \mathbf{1}_{G_l} = 0,\ \forall (k,l) \in [K]^2$, we only need to check for $v \in \Gamma$,

$$v^T W_k v \geq 0,\ \forall k \in [K].$$

Note that $v^T \mathbf{1}_{G_k} = 0$, we have

$$v^T W_k v = \lambda \|v\|^2 - S_k(v) - T_k(v),$$

where $S_k(v) := v^T A_k v = v^T X^T \Sigma_k^{-1} X v$, and $T_k(v) = v^T B v$. By concentration bound we can get

$$\mathbb{P}(S_k(v) \leq MK(\sqrt{n} + \sqrt{p} + \sqrt{2 \log n}),\ \text{for all } k \in [K]) \geq 1 - \frac{K}{n}.$$

For $T_k(v)$, first we define

$$V_{k,ll'}^{(1)} := \langle \Sigma_{l'}^{1/2} \Sigma_l^{-1} (\mu_{l'} - \mu_l),\ \sum_{j \in G_{l'}} v_j \epsilon_j \rangle;$$

$$V_{k,ll'}^{(2)} := \langle \bar{\epsilon}_{l'} - \Sigma_{l'}^{1/2} \Sigma_l^{-1/2} \bar{\epsilon}_l,\ \sum_{j \in G_{l'}} v_j \epsilon_j \rangle;$$

$$V_{k,ll'}^{(3)} := \frac{1}{2} \sum_{j \in G_{l'}} \epsilon_j^T \Sigma_{l'}^{1/2} (\Sigma_l^{-1} - \Sigma_{l'}^{-1}) \Sigma_{l'}^{1/2} \epsilon_j v_j;$$

$$V_{k,ll'}^{(4)} := \frac{1}{n_l} \sum_{j \in G_{l'}} ([A_{l,G_{l'}G_l} \mathbf{1}_{G_l}]_j - [A_{k,G_{l'}G_l} \mathbf{1}_{G_l}]_j) v_j \cdot \mathbf{1}(l \neq l').$$

Then we can write $T_k(v)$ as

$$T_k(v) := \sum_{l \neq l'} \frac{n_l n_{l'}}{\mathbf{1}_n^T B_k \mathbf{1}_n} (T_{k,ll'}^{(1)} + T_{k,ll'}^{(2)} + T_{k,ll'}^{(3)} + T_{k,ll'}^{(4)} + T_{k,ll'}^{(5)}),$$

where

$$T_{k,ll'}^{(1)} := V_{k,ll'}^{(1)} \cdot V_{k,l'l}^{(1)};$$
$$T_{k,ll'}^{(2)} := V_{k,ll'}^{(2)} \cdot V_{k,l'l}^{(2)};$$
$$T_{k,ll'}^{(3)} := V_{k,ll'}^{(1)} \cdot V_{k,l'l}^{(2)} + V_{k,ll'}^{(2)} \cdot V_{k,l'l}^{(1)};$$
$$T_{k,ll'}^{(4)} := (V_{k,ll'}^{(3)} + V_{k,ll'}^{(4)}) \cdot (V_{k,l'l}^{(1)} + V_{k,l'l}^{(2)}) + (V_{k,ll'}^{(1)} + V_{k,ll'}^{(2)}) \cdot (V_{k,l'l}^{(3)} + V_{k,l'l}^{(4)});$$
$$T_{k,ll'}^{(5)} := (V_{k,ll'}^{(3)} + V_{k,ll'}^{(4)}) \cdot (V_{k,l'l}^{(3)} + V_{k,l'l}^{(4)}).$$

Now we choose $\lambda = p + \frac{\beta}{4M} m \Delta^2$, which implies that

$$\mathbf{1}_n^T B_k \mathbf{1}_n \geq \frac{n_l n_{l'}}{8} \frac{\beta}{M} \max\{\|\Sigma_{l'}^{-1/2}(\mu_l - \mu_{l'})\|^2, \|\Sigma_l^{-1/2}(\mu_l - \mu_{l'})\|^2\}.$$

From concentration bounds for Gaussians we have for all triple $(k,l,l') \in [K]^3$ such that $(k,l,l') \neq (k,k,k)$,

$$\left| \sum_{l \neq l'} \frac{n_l n_{l'}}{\mathbf{1}_n^T B_k \mathbf{1}_n} T_{k,ll'}^{(1)} \right| \leq \frac{CM^2}{\beta} \cdot (n + \sqrt{2nK \log n} + 2K \log n) \|v\|^2;$$

$$\left| \sum_{l \neq l'} \frac{n_l n_{l'}}{\mathbf{1}_n^T B_k \mathbf{1}_n} T_{k,ll'}^{(2)} \right| \leq \frac{CM^3}{1-\beta} \cdot (\delta \sqrt{mp \log n} + \sqrt{mp \log^7 n / \underline{n}}) \|v\|^2;$$

$$\left| \sum_{l \neq l'} \frac{n_l n_{l'}}{\mathbf{1}_n^T B_k \mathbf{1}_n} T_{k,ll'}^{(5)} \right| \leq \frac{CK^2}{\beta} \cdot (\frac{1-\beta}{(M-1)^2 M} (p + pM \log p / \log n) + M(M-1)) n \|v\|^2,$$

Or

$$\left| \sum_{l \neq l'} \frac{n_l n_{l'}}{\mathbf{1}_n^T B_k \mathbf{1}_n} T_{k,ll'}^{(5)} \right| \leq \frac{CK^2 M(1-\beta)(M-1)}{\beta} \cdot (\sqrt{\frac{p^3 m}{\log n}} + \sqrt{pm \log n}) n \|v\|^2,$$

with probability $\geq 1 - CK^3/n^\delta$ for some constant C.

Note that by assumption we have $\Delta^2 \geq \frac{C(M-1)^3M^2}{(1-\beta)(1-\eta)}(p + \log n) + \frac{CM^3}{(1-\beta)}\sqrt{(1+\delta)p\log n/m}$ and the fact that the remaining terms of $T_{k,ll'}$ can be bounded by the above inequalities up to multiplied by some constant, we can directly verify that (C2) is true under our assumptions. ■

Lemma 6 (**Separation bound on the covariance matrices**). Let $\lambda_1, \ldots, \lambda_p$ correspond to the eigenvalues of $(\Sigma_l^{1/2}\Sigma_k^{-1}\Sigma_l^{1/2} - \mathrm{Id}_p)$ and define $D_{(k,l)} := \frac{\sum_{i=1}^p(\lambda_i - \log(1+\lambda_i))}{p\max_i|\lambda_i|}$. If there exists constant C such that

$$\min_{k\neq l} D_{(k,l)} \geq C(1 + \log n/p + p/n),$$

then

$$\mathbb{P}\Big([A_{l,G_lG_l}\mathbf{1}_{G_l}]_j - [A_{k,G_lG_l}\mathbf{1}_{G_l}]_j \geq 0, \text{ for all } (k,l) \in [K]^2 \text{ and } j \in G_l\Big) \geq 1 - CK^2/n.$$

Sketch of the proof. Let $T := [A_{l,G_lG_l}\mathbf{1}_{G_l}]_j - [A_{k,G_lG_l}\mathbf{1}_{G_l}]_j$, $B := \Sigma_l^{1/2}\Sigma_k^{-1}\Sigma_l^{1/2} - \mathrm{Id}_p$ then by definition we have

$$\begin{aligned}
T = &-\sum_{i=1}^p \log(\lambda_i + 1) + \sum_{i=1}^p \lambda_i \\
&+ \frac{1}{2}\langle B, \epsilon_j\epsilon_j^T - \mathrm{Id}_p\rangle \\
&- \frac{1}{2}\langle B, \frac{1}{n_l}\sum_{t\in G_l}\epsilon_t\epsilon_j^T + \epsilon_j\Big(\frac{1}{n_l}\sum_{t\in G_l}\epsilon_t\Big)^T\rangle \\
&+ \frac{1}{2}\langle B, \frac{1}{n_l}\sum_{t\in G_l}\epsilon_t\epsilon_t^T - \mathrm{Id}_p\rangle,
\end{aligned}$$

where the last three terms can be bounded by concentration bounds for Gaussians. ■