Adaptive Clustering through Semidefinite Programming

Martin Royer

Adaptive Clustering through Semidefinite Programming

Martin Royer

Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS,
Université Paris-Saclay, 91405 Orsay, France.

Abstract

We analyze the clustering problem through a flexible probabilistic model that aims to identify an optimal partition on the sample $X_1, ..., X_n$. We perform exact clustering with high probability using a convex semidefinite estimator that interprets as a corrected, relaxed version of K-means. The estimator is analyzed through a non-asymptotic framework and showed to be optimal or near-optimal in recovering the partition. Furthermore, its performances are shown to be adaptive to the problem's effective dimension, as well as to K the unknown number of groups in this partition. We illustrate the method's performances in comparison to other classical clustering algorithms with numerical experiments on simulated data.

1 Introduction

Clustering, a form of unsupervised learning, is the classical problem of assembling n observations $X_1, ..., X_n$ from a p-dimensional space into K groups. Applied fields are craving for robust clustering techniques, such as computational biology with genome classification, data mining or image segmentation from computer vision. But the clustering problem has proven notoriously hard when the embedding dimension is large compared to the number of observations (see for instance the recent discussions from [2, 21]).

A famous early approach to clustering is to solve for the geometrical estimator K-means [12, 13, 19]. The intuition behind its objective is that groups are to be determined in a way to minimize the total intra-group variance. It can be interpreted as an attempt to "best" represent the observations by K points, a form of vector quantization. Although the method shows great performances when observations are homoscedastic, K-means is a NP-hard, ad-hoc method. Clustering with probabilistic frameworks are usually based on maximum likelihood approaches paired with a variant of the EM algorithm for model estimation, see for instance the works of Fraley & Raftery [10] and Dasgupta & Schulman [9]. These methods are widespread and popular, but they tend to be very sensitive to initialization and model misspecifications.

Several recent developments establish a link between clustering and semidefinite programming. Peng & Wei [16] show that the K-means objective can be relaxed into a convex, semidefinite program, leading Mixon *et al.* [15] to use this relaxation under a subgaussian mixture model to estimate the cluster centers. Chrétien *et al.* [8] use a slightly different form of a semidefinite program, inspired by work on community detection by Guédon & Vershynin [11], to recover the adjacency matrix of the cluster graph with high probability. Lastly in the different context of variable clustering, Bunea *et al.* [5] present a semidefinite program with a correction step to produce non-asymptotic exact recovery results.

In this work, we introduce a semidefinite, penalized estimator for point clustering inspired by [16] and adapted from the work and context of [5]. We analyze it through a flexible probabilistic model inducing an optimal partition that we aim to recover. We investigate the optimal conditions of exact clustering recovery with high probability and show optimal performances – including in high dimensions, improving on [15], as well as adaptability to the effective dimension of the problem. We also show that our results continue to hold without knowledge of the number of groups K. Lastly we provide evidence of our method's efficiency from simulated data and suggest a coherent alternative in case our estimator is too costly to compute.

Notation. Throughout this work we use the convention $0/0 := 0$ and $[n] = \{1, ..., n\}$. We take $a_n \lesssim b_n$ to mean that a_n is smaller than b_n up to an absolute constant factor. Let $\mathcal{S}_{d-1}$ denote the unit sphere in $\mathbb{R}^d$. For $q \in \mathbb{N}^* \cup \{+\infty\}$, $\nu \in \mathbb{R}^d$, $|\nu|_q$ is the l_q-norm and for $M \in \mathbb{R}^{d \times d'}$, $|M|_q$, $|M|_F$, $|M|_*$ and $|M|_{op}$ are respectively the entry-wise l_q-norm, the Frobenius norm associated with scalar product $\langle ., . \rangle$, the nuclear norm and the operator norm. $|D|_V$ is the variation semi-norm for a diagonal matrix D, the difference between its maximum and minimum element. Let $A \succcurlyeq B$ mean that $A - B$ is symmetric, positive semidefinite.

2 Probabilistic modeling of point clustering

Consider $X_1, ..., X_n$ and let $\nu_a = \mathbb{E}[X_a]$. The variable X_a can be decomposed into

$$X_a = \nu_a + E_a, \quad a = 1, ..., n, \tag{2.1}$$

with E_a stochastic centered variables in $\mathbb{R}^p$.

Definition 1. *For $K > 1$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_K) \in (\mathbb{R}^p)^K$, $\delta \geqslant 0$ and $\mathcal{G} = \{G_1, ..., G_K\}$ a partition of $[n]$, we say $X_1, ..., X_n$ are $(\mathcal{G}, \boldsymbol{\mu}, \delta)$-clustered if $\forall k \in [K], \forall a \in G_k, |\nu_a - \mu_k|_2 \leqslant \delta$. We then call*

$$\Delta(\boldsymbol{\mu}) := \min_{k<l} |\mu_k - \mu_l|_2 \tag{2.2}$$

the separation between the cluster means, and

$$\rho(\mathcal{G}, \boldsymbol{\mu}, \delta) := \Delta(\boldsymbol{\mu})/\delta \tag{2.3}$$

the discriminating capacity of $(\mathcal{G}, \boldsymbol{\mu}, \delta)$.

In this work we assume that $X_1, ..., X_n$ are $(\mathcal{G}, \boldsymbol{\mu}, \delta)$-clustered. Notice that this definition does not impose any constraint on the data: for any given $\mathcal{G}$, there exists a choice of $\boldsymbol{\mu}$, means and radius δ important enough so that $X_1, ..., X_n$ are $(\mathcal{G}, \boldsymbol{\mu}, \delta)$-clustered. But we are interested in partitions with greater discriminating capacity, i.e. that make more sense in terms of group separation. Indeed remark that if $\rho(\mathcal{G}, \boldsymbol{\mu}, \delta) < 2$, the population clusters $\{\nu_a\}_{a \in G_1}, ..., \{\nu_a\}_{a \in G_K}$ are not linearly separable, but a high $\rho(\mathcal{G}, \boldsymbol{\mu}, \delta)$ implies that they are well-separated from each other. Furthermore, we have the following result.

Proposition 1. *Let $(\mathcal{G}_K^*, \boldsymbol{\mu}^*, \delta^*) \in \arg\max \rho(\mathcal{G}, \boldsymbol{\mu}, \delta)$ for $(\mathcal{G}, \boldsymbol{\mu}, \delta)$ such that $X_1, ..., X_n$ are $(\mathcal{G}, \boldsymbol{\mu}, \delta)$-clustered, and $|\mathcal{G}| = K$. If $\rho(\mathcal{G}_K^*, \boldsymbol{\mu}^*, \delta^*) > 4$ then $\mathcal{G}_K^*$ is the unique maximizer of $\rho(\mathcal{G}, \boldsymbol{\mu}, \delta)$.*

So $\mathcal{G}_K^*$ is the partition maximizing the discriminating capacity over partitions of size K. Therefore in this work, we will assume that there is a $K > 1$ such that $X_1, ..., X_n$ is $(\mathcal{G}, \boldsymbol{\mu}, \delta)$-clustered with $|\mathcal{G}| = K$ and $\rho(\mathcal{G}, \boldsymbol{\mu}, \delta) > 4$. By Proposition 1, $\mathcal{G}$ is then identifiable. It is the partition we aim to recover.

We also assume that $X_1, ..., X_n$ are independent observations with subgaussian behavior. Instead of the classical isotropic definition of a subgaussian random vector (see for example [20]), we use a more flexible definition that can account for anisotropy.

Definition 2. *Let Y be a random vector in $\mathbb{R}^d$, Y has a subgaussian distribution if there exist $\Sigma \in \mathbb{R}^{d\times d}$ such that $\forall x \in \mathbb{R}^d$,*

$$\mathbb{E}\left[e^{x^T(Y-\mathbb{E}\,Y)}\right] \leqslant e^{x^T\Sigma x/2}. \tag{2.4}$$

We then call Σ a variance-bounding matrix of random vector Y, and write shorthand $Y \sim \text{subg}(\Sigma)$. Note that $Y \sim \text{subg}(\Sigma)$ implies $\text{Cov}(Y) \preccurlyeq \Sigma$ in the semidefinite sense of the inequality. To sum-up our modeling assumptions in this work:

Hypothesis 1. *Let $X_1, ..., X_n$ be independent, subgaussian, $(\mathcal{G}, \boldsymbol{\mu}, \delta)$-clustered with $\rho(\mathcal{G}, \boldsymbol{\mu}, \delta) > 4$.*

Remark that the modelization of Hypothesis 1 can be connected to another popular probabilistic model: if we further ask that $X_1, ..., X_n$ are identically-distributed within a group (and hence $\delta = 0$), the model becomes a realization of a *mixture model*.

3 Exact partition recovery with high probability

Let $\mathcal{G} = \{G_1, ..., G_K\}$ and $m := \min_{k\in[K]} |G_k|$ denote the minimum cluster size. $\mathcal{G}$ can be represented by its caracteristic matrix $B^* \in \mathbb{R}^{n\times n}$ defined as $\forall k, l \in [K]^2, \forall (a, b) \in G_k \times G_l$,

$$B^*_{ab} := \begin{cases} 1/|G_k| & \text{if } k = l \\ 0 & \text{otherwise.} \end{cases}$$

In what follows, we will demonstrate the recovery of $\mathcal{G}$ through recovering its caracteristic matrix B^*. We introduce the sets of square matrices

$$\mathcal{C}_K^{\{0,1\}} := \{B \in \mathbb{R}_+^{n\times n} : B^T = B, \text{tr}(B) = K, B1_n = 1_n, B^2 = B\} \tag{3.1}$$

$$\mathcal{C}_K := \{B \in \mathbb{R}_+^{n\times n} : B^T = B, \text{tr}(B) = K, B1_n = 1_n, B \succcurlyeq 0\} \tag{3.2}$$

$$\mathcal{C} := \bigcup_{K\in\mathbb{N}} \mathcal{C}_K. \tag{3.3}$$

We have: $\mathcal{C}_K^{\{0,1\}} \subset \mathcal{C}_K \subset \mathcal{C}$ and $\mathcal{C}_K$ is convex. Notice that $B^* \in \mathcal{C}_K^{\{0,1\}}$. A result by Peng, Wei (2007) [16] shows that the K-means estimator $\bar{B}$ can be expressed as

$$\bar{B} = \underset{B\in\mathcal{C}_K^{\{0,1\}}}{\arg\max} \langle \widehat{\Lambda}, B\rangle \tag{3.4}$$

for $\widehat{\Lambda} := (\langle X_a, X_b\rangle)_{(a,b)\in[n]^2} \in \mathbb{R}^{n\times n}$, the observed Gram matrix. Therefore a natural relaxation is to consider the following estimator:

$$\widehat{B} := \underset{B\in\mathcal{C}_K}{\arg\max} \langle \widehat{\Lambda}, B\rangle. \tag{3.5}$$

Notice that $\mathbb{E}\,\widehat{\Lambda} = \Lambda + \Gamma$ for $\Lambda := (\langle \nu_a, \nu_b\rangle)_{(a,b)\in[n]^2} \in \mathbb{R}^{n\times n}$, and $\Gamma := \mathbb{E}\left[\langle E_a, E_b\rangle\right]_{(a,b)\in[n]^2} = \text{diag}\left(|\,\text{Var}(E_a)|_*\right)_{1\leqslant a\leqslant n} \in \mathbb{R}^{n\times n}$. The following two results demonstrate that Λ is the signal structure that lead the optimizations of (3.4) and (3.5) to recover B^*, whereas Γ is a bias term that can hurt the process of recovery.

Proposition 2. *There exist $c_0 > 1$ absolute constant such that if $\rho^2(\mathcal{G}, \boldsymbol{\mu}, \delta) > c_0(6 + \sqrt{n}/m)$ and $m\Delta^2(\boldsymbol{\mu}) > 8|\Gamma|_V$, then we have*

$$\underset{B\in\mathcal{C}_K^{\{0,1\}}}{\arg\max} \langle \Lambda + \Gamma, B\rangle = B^* = \underset{B\in\mathcal{C}_K}{\arg\max} \langle \Lambda + \Gamma, B\rangle. \tag{3.6}$$

This proposition shows that the $\widehat{B}$ estimator, as well as the K-means estimator, would recover partition $\mathcal{G}$ on the population Gram matrix if the variation semi-norm of Γ were sufficiently small compared to the cluster separation. Notice that to recover the partition on the population version, we require the discriminating capacity to grow as fast as $1 + (\sqrt{n}/m)^{1/2}$ instead of simply 1 from Hypothesis 1. The following proposition demonstrates that if the condition on the variation semi-norm of Γ is not met, $\mathcal{G}$ may not even be recovered on the population version.

Proposition 3. *There exist $\mathcal{G}, \boldsymbol{\mu}, \delta$ and Γ such that $\rho^2(\mathcal{G}, \boldsymbol{\mu}, \delta) = +\infty$ but we have $m\Delta^2(\boldsymbol{\mu}) < 2|\Gamma|_V$ and*

$$B^* \notin \underset{B \in \mathcal{C}_K^{\{0,1\}}}{\arg\max} \langle \Lambda + \Gamma, B \rangle \quad \text{and} \quad B^* \notin \underset{B \in \mathcal{C}_K}{\arg\max} \langle \Lambda + \Gamma, B \rangle. \tag{3.7}$$

So Proposition 3 shows that even if the population clusters are perfectly discriminated, there is a configuration for the variances of the noise that makes it impossible to recover the right clustering by K-means. This shows that K-means may fail when the random variable homoscedasticity assumption is violated, and that it is important to correct for Γ.
The estimator from [5] can be adapted to our context. We introduce the following estimator, for $(a, b) \in [n]^2$ let $V(a, b) := \max_{(c,d) \in ([n] \setminus \{a,b\})^2} \left| \langle X_a - X_b, \frac{X_c - X_d}{|X_c - X_d|_2} \rangle \right|$, $b_1 := \arg\min_{b \in [n] \setminus \{a\}} V(a, b)$ and $b_2 := \arg\min_{b \in [p] \setminus \{a, b_1\}} V(a, b)$. Then for $a \in [n]$, let

$$\widehat{\Gamma}^{corr} := \operatorname{diag}\left(\langle X_a - X_{b_1}, X_a - X_{b_2} \rangle_{a \in [n]} \right). \tag{3.8}$$

Computing $\widehat{\Gamma}^{corr}$ can be interpreted as a correcting term to de-bias $\widehat{\Lambda}$ as an estimator of Λ. The result from Proposition 2 demonstrates the interest of studying the following semi-definite estimator of the projection matrix B^*, let

$$\widehat{B}^{corr} := \underset{B \in \mathcal{C}_K}{\arg\max} \langle \widehat{\Lambda} - \widehat{\Gamma}^{corr}, B \rangle. \tag{3.9}$$

In order to demonstrate the recovery of B^* by this estimator, we introduce different quantitative measures of the "spread" of our stochastic variables, that affect the quality of the recovery. By Hypothesis 1 there exist $\Sigma_1, ..., \Sigma_n$ such that $\forall a \in [n]$, $X_a \sim \operatorname{subg}(\Sigma_a)$. Let

$$\sigma^2 := \max_{a \in [n]} |\Sigma_a|_{op}, \quad \mathcal{V}^2 := \max_{a \in [n]} |\Sigma_a|_F, \quad \gamma^2 := \max_{a \in [n]} |\Sigma_a|_*. \tag{3.10}$$

We are now ready to introduce this paper's main result: a condition on the separation between the cluster means sufficient for ensuring recovery of B^* with high probability.

Theorem 1. *Assume that $m > 2$. For $c_1, c_2 > 0$ absolute constants, if*

$$m\Delta^2(\boldsymbol{\mu}) \geqslant c_2 \big(\sigma^2(n + m \log n) + \mathcal{V}^2(\sqrt{n + m \log n}) + \gamma(\sigma\sqrt{\log n} + \delta) + \delta^2(\sqrt{n} + m) \big), \tag{3.11}$$

then with probability larger than $1 - c_1/n$ we have $\widehat{B}^{corr} = B^*$, and therefore $\widehat{\mathcal{G}}^{corr} = \mathcal{G}$.

We call the right hand side of (3.11) the separating rate. Notice that we can read two kinds of requirements coming from the separating rate: requirements on the radius δ, and requirements on $\sigma^2, \mathcal{V}^2, \gamma$ dependent on the distributions of observations. It appears as if $\delta + \sigma\sqrt{\log n}$ can be interpreted as a geometrical width of our problem. If we ask that δ is of the same order as $\sigma\sqrt{\log n}$, a maximum gaussian deviation for n variables, then all conditions on δ from (3.11) can be removed. Thus for convenience of the following discussion we will now assume $\delta \lesssim \sigma\sqrt{\log n}$.

How optimal is the result from Theorem 1? Notice that our result is adapted to anisotropy in the noise, but to discuss optimality it is easier to look at the isotropic scenario: $\mathcal{V}^2 = \sqrt{p}\sigma^2$ and $\gamma^2 = p\sigma^2$. Therefore $\Delta^2(\boldsymbol{\mu})/\sigma^2$ represents a signal-to-noise ratio. For simplicity let us also assume that all groups have equal size, that is $|G_1| = ... = |G_K| = m$ so that $n = mK$ and the sufficient condition (3.11) becomes

$$\frac{\Delta^2(\boldsymbol{\mu})}{\sigma^2} \gtrsim \big(K + \log n\big) + \sqrt{(K + \log n)\frac{pK}{n}}. \tag{3.12}$$

Optimality. To discuss optimality, we distinguish between low and high dimensional setups.
In the low-dimensional setup $n \vee m \log n \gtrsim p$, we obtain the following condition:

$$\frac{\Delta^2(\boldsymbol{\mu})}{\sigma^2} \gtrsim \big(K + \log n\big). \tag{3.13}$$

Discriminating with high probability between n observations from two gaussians in dimension 1 would require a separating rate of at least $\sigma^2 \log n$. This implies that when $K \lesssim \log n$, our result is minimax. Otherwise, to our knowledge the best clustering result on approximating mixture center is from [15], and on the condition that $\Delta^2(\boldsymbol{\mu})/\sigma^2 \gtrsim K^2$. Furthermore, the $K \gtrsim \log n$ regime is known in the stochastic-block-model community as a hard regime where a gap is surmised to exist between the minimal information-theoretic rate and the minimal achievable computational rate (see for example [7]).
In the high-dimensional setup $n \vee m \log n \lesssim p$, condition (3.12) becomes:

$$\frac{\Delta^2(\boldsymbol{\mu})}{\sigma^2} \gtrsim \sqrt{(K + \log n)\frac{pK}{n}}. \tag{3.14}$$

There are few information-theoretic bounds for high-dimension clustering. Recently, Banks, Moore, Vershynin, Verzelen and Xu (2017) [3] proved a lower bound for Gaussian mixture clustering detection, namely they require a separation of order $\sqrt{K(\log K)p/n}$. When $K \lesssim \log n$, our condition is only different in that it replaces $\log(K)$ by $\log(n)$, a price to pay for going from detecting the clusters to exactly recovering the clusters. Otherwise when K grows faster than $\log n$ there might exist a gap between the minimal possible rate and the achievable, as discussed previously.
Adaptation to effective dimension. We can analyse further the condition (3.11) by introducing an effective dimension r_*, measuring the largest volume repartition for our variance-bounding matrices $\Sigma_1, ..., \Sigma_n$. Let

$$r_* := \frac{\gamma^2}{\sigma^2} = \frac{\max_{a\in[n]} |\Sigma_a|_*}{\max_{a\in[n]} |\Sigma_a|_{op}}, \tag{3.15}$$

r_* can also be interpreted as a form of global effective rank of matrices Σ_a. Indeed, define $Re(\Sigma) := |\Sigma|_*/|\Sigma|_{op}$, then we have $r_* \leqslant \max_{a\in[n]} Re(\Sigma_a) \leqslant \max_{a\in[n]} \operatorname{rank}(\Sigma_a) \leqslant p$.
Now using $\mathcal{V}^2 \leqslant \sqrt{r_*}\sigma^2$ and $\gamma = \sqrt{r_*}\sigma$, condition (3.11) can be written as

$$\frac{\Delta^2(\boldsymbol{\mu})}{\sigma^2} \gtrsim \big(K + \log n\big) + \sqrt{(K + \log n)\frac{r_* K}{n}}. \tag{3.16}$$

By comparing this equation to (3.12), notice that r_* is in place of p, indeed playing the role of an effective dimension for the problem. This also shows that our estimator adapts to this effective dimension, without any dimension reduction step. In consequence, equation (3.16) distinguishes between an actual high-dimensional setup: $n \vee m \log n \lesssim r_*$ and a "low" dimensional setup $r_* \lesssim$

$n \vee m \log n$ under which, regardless of the actual value of p, our estimators recovers under the near-minimax condition of (3.13).
This informs on the effect of correcting term $\widehat{\Gamma}^{corr}$ in the theorem above when $n+m\log n \lesssim r_*$. The un-corrected version of the semi-definite program (3.5) has a leading separating rate of $\gamma^2/m = \sigma^2 r_*/m$, but with the $\widehat{\Gamma}^{corr}$ correction on the other hand, (3.16) has leading separating factor smaller than $\sigma^2\sqrt{(K+\log n)r_*/m} = \sigma^2\sqrt{n+m\log n}\times\sqrt{r_*}/m$. This proves that in a high-dimensional setup, our correction enhances the separating rate of at least a factor $\sqrt{(n+m\log n)/r_*}$.

4 Adaptation to the unknown number of group K

It is rarely the case that K is known, but we can proceed without it. We produce an estimator adaptive to the number of groups K: let $\widehat{\kappa} \in \mathbb{R}_+$, we now study the following adaptive estimator:

$$\widetilde{B}^{corr} := \underset{B\in\mathcal{C}}{\arg\max}\langle \widehat{\Lambda} - \widehat{\Gamma}^{corr}, B\rangle - \widehat{\kappa}\,\mathrm{tr}(B). \tag{4.1}$$

Theorem 2. *Suppose that $m > 2$ and (3.11) is satisfied. For $c_3, c_4, c_5 > 0$ absolute constants suppose that the following condition on $\widehat{\kappa}$ is satisfied*

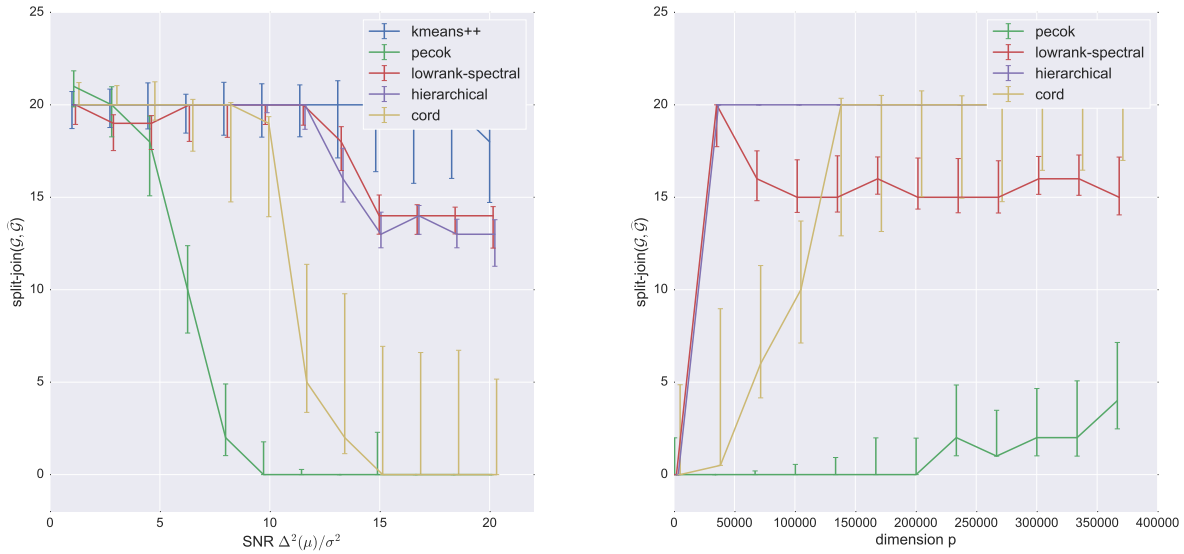
$$c_4\Big(\mathcal{V}^2\sqrt{n} + \sigma^2 n + \gamma(\sigma\sqrt{\log n}+\delta) + \delta^2\sqrt{n}\Big) < c_5\widehat{\kappa} < m\Delta^2(\boldsymbol{\mu}), \tag{4.2}$$

then we have $\widetilde{B}^{corr} = B^*$ with probability larger than $1 - c_3/n$

Notice that condition (4.2) essentially requires $\widehat{\kappa}$ to be seated between $m\Delta^2(\boldsymbol{\mu})$ and some components of the right-hand side of (3.11). So under (4.2), the results from the previous section apply to the adaptive estimator $\widetilde{B}^{corr}$ as well and this shows that it is not necessary to know K in order to perform well for recovering $\mathcal{G}$. Finding an optimized, data-driven parameter $\widehat{\kappa}$ using some form of cross-validation is outside of the scope of this paper.

5 Numerical experiments

We illustrate our method on simulated Gaussian data in two challenging, high-dimensional setup experiments for comparing clustering estimators. Our sample are drawn from $K = 3$ identically-sized, identically distributed and perfectly discriminated clusters of non-isovolumic Gaussians. The distributions are chosen to be isotropic, and the ratio between the lowest and the highest standard deviation is of 1 to 10. We draw points of a $\mathbb{R}^p$ space in two different scenarii. In $(\mathcal{S}_1)$, for a given dimension space $p = 2000$ and a fixed isotropic noise level, we report the algorithms' compared performances as the signal-to-noise ratio $\Delta^2(\boldsymbol{\mu})/\sigma^2$ is increased from 1 to 20. In $(\mathcal{S}_2)$ we impose a fixed signal to noise ratio, and observe the algorithm's decay in performance as the space dimension p is increased from 100 to 400 000. All points of the simulated space are reported as a median value with asymmetric standard deviations in the form of errorbars over a hundred simulations.
Solving for estimator $\widehat{B}^{corr}$ is a hard problem as n grows. For this task we implemented an ADMM solver from the work of Boyd *et al.* [4] with multiple stopping criterions including a fixed number of iterations of $T = 3000$. The results we report use $n = 30$ samples. For reference, we compare the recovering capacities of $\widehat{\mathcal{G}}^{corr}$, labeled 'pecok' in Figure 1 with other classical clustering algorithm. We chose three different but standard clustering procedures: Lloyd's K-means algorithm [12] with K-means++ initialization [1] (although in scenario $(\mathcal{S}_2)$, it is too slow to converge as p grows so we do not report it), Ward's method for Hierarchical Clustering [22] and the low-rank clustering



Scenario $(\mathcal{S}_1)$ Scenario $(\mathcal{S}_2)$

Figure 1: Performance comparison for classical clustering estimators and ours $\widehat{\mathcal{G}}^{corr}$, labeled 'pecok' in reference to [5]. The lower split-join, the better the clustering performance and split-join$(\mathcal{G}, \widehat{\mathcal{G}}) = 0$ implies $\widehat{\mathcal{G}} = \mathcal{G}$.

algorithm applied to the Gram matrix, a spectral method appearing in McSherry [14]. Lastly we include the CORD algorithm from Bunea *et al.* [6].

We measure the performances of estimators by computing the split-join metric on the cluster graphs, counting the number of edges to remove or add to go from one graph to the other. In the two experiments, the results of $\widehat{\mathcal{G}}^{corr}$ are markedly better than that of other methods. Scenario $(\mathcal{S}_1)$ shows it can achieve exact recovery with a lesser signal to noise ratio than its competitors, whereas scenario $(\mathcal{S}_2)$ shows its performances are decaying at a much lower rate than the others when the space dimension is increased.

Because of the slow convergence of ADMM, $\widehat{\mathcal{G}}^{corr}$ comes with important computation times. Of course all of the compared methods have a very hard time reaching high sample sizes n in the high dimensional context and to that regard, the low-rank clustering method is by far the most promising.

6 Conclusion

In this paper we analyzed a new semidefinite positive algorithm for clustering within the context of a flexible probabilistic model and exhibit the key quantities that guarantee non-asymptotic exact recovery. It implies an essential bias-removing correction that significanty improves the recovering rate in the high-dimensional setup. Hence we showed the estimator to be near-minimax, adapted to an effective dimension of the problem. We demonstrated that our estimator can in theory be optimally adapted to a data-driven choice of K. Lastly we illustrated on high-dimensional experiments that our approach is empirically stronger than other classical clustering methods.

Our method is computationally intensive even though it is of polynomial order. As the $\widehat{\Gamma}^{corr}$

correction step of the algorithm can be interpreted as an independent, denoising step for the Gram matrix, we suggest using it as such for other notably faster algorithm such as the spectral algorithms.

Acknowledgements

This work is supported by a public grant overseen by the French National research Agency (ANR) as part of the "Investissement d'Avenir" program, through the "IDI 2015" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02. It is also supported by the CNRS PICS funding HighClust. We thank Christophe Giraud for a shrewd, unwavering thesis direction.

References

[1] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[2] Martin Azizyan, Aarti Singh, and Larry Wasserman. Efficient Sparse Clustering of High-Dimensional Non-spherical Gaussian Mixtures. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 37–45, San Diego, California, USA, 09–12 May 2015. PMLR.

[3] J. Banks, C. Moore, N. Verzelen, R. Vershynin, and J. Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *ArXiv e-prints*, July 2016.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, January 2011.

[5] F. Bunea, C. Giraud, M. Royer, and N. Verzelen. PECOK: a convex optimization approach to variable clustering. *ArXiv e-prints*, June 2016.

[6] Florentina Bunea, Christophe Giraud, and Xi Luo. Minimax optimal variable clustering in g-models via cord. *arXiv preprint arXiv:1508.01939*, 2015.

[7] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.*, 17(1):882–938, January 2016.

[8] Stéphane Chrétien, Clément Dombry, and Adrien Faivre. A semi-definite programming approach to low dimensional embedding for unsupervised clustering. *CoRR*, abs/1606.09190, 2016.

[9] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *J. Mach. Learn. Res.*, 8:203–226, May 2007.

[10] Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[11] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *CoRR*, abs/1411.4686, 2014.

[12] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, September 1982.

[13] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[14] F. McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42Nd IEEE Symposium on Foundations of Computer Science*, FOCS '01, pages 529–, Washington, DC, USA, 2001. IEEE Computer Society.

[15] D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures with k-means. In *2016 IEEE Information Theory Workshop (ITW)*, pages 211–215, Sept 2016.

[16] Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM J. on Optimization*, 18(1):186–205, February 2007.

[17] Philippe Rigollet. *High-Dimensional Statistics*. Massachusetts Institute of Technology: MIT OpenCourseWare, Spring 2015.

[18] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18(82):1–9, 2013.

[19] H. Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1:801–804, 1956.

[20] R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. Chapter 5 of: Compressed Sensing, Theory and Applications. Cambridge University Press, 2012.

[21] N. Verzelen and E. Arias-Castro. Detection and Feature Selection in Sparse Mixture Models. *ArXiv e-prints*, May 2014.

[22] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

Appendix

A Intermediate results

A.1 Generic controls for exact recovery

Let $\widehat{\Gamma}$ be any estimator of Γ and let $\widehat{B} := \arg\max_{B\in\mathcal{C}_K}\langle\widehat{\Lambda}-\widehat{\Gamma}, B\rangle$.

Theorem 3. *For $c_1, c_2 > 0$ absolute constants suppose that $|\widehat{\Gamma}-\Gamma|_V \leqslant \bar{\gamma}_n^2$ with probability $1-c_1/n$, and that*

$$m\Delta^2(\boldsymbol{\mu}) \geqslant c_2\Big(\sigma^2(n+m\log n) + \mathcal{V}^2(\sqrt{n+m\log n}) + \bar{\gamma}_n^2 + \delta^2(\sqrt{n}+m)\Big), \tag{A.1}$$

then we have $\widehat{B} = B^*$ with probability larger than $1-c_1/n$

In the case where the number of groups is unknown we study $\widetilde{B} := \arg\max_{B\in\mathcal{C}}\langle\widehat{\Lambda}-\widehat{\Gamma}, B\rangle - \widehat{\kappa}\,\mathrm{tr}(B)$ for $\widehat{\kappa}\in\mathbb{R}$.

Theorem 4. *For $c_3, c_4, c_5 > 0$ absolute constants suppose that $|\widehat{\Gamma}-\Gamma|_\infty \leqslant \bar{\gamma}_n^2$ with probability $1-c_3/n$. Suppose that (A.1) is satisfied and that the following condition on $\widehat{\kappa}$ is satisfied*

$$c_4\Big(\mathcal{V}^2\sqrt{n} + \sigma^2 n + \bar{\gamma}_n^2 + \delta^2\sqrt{n}\Big) < c_5\widehat{\kappa} < m\Delta^2(\boldsymbol{\mu}), \tag{A.2}$$

then we have $\widetilde{B} = B^*$ with probability larger than $1-c_3/n$

A.2 On estimating Γ

In the general case we have $\widehat{\Gamma} = 0$ hence a deterministic perturbation term $\bar{\gamma}_n^2 = |\Gamma|_\infty$ weighing on the separation requirements. For $\widehat{\Gamma}^{corr}$, we have the following result.

Proposition 4. *Assume that $m > 2$. For $c_6, c_7 > 0$ absolute constants, with probability larger than $1-c_6/n$ we have*

$$|\widehat{\Gamma}^{corr}-\Gamma|_\infty \leqslant c_7\Big(\sigma^2\log n + (\delta+\sigma\sqrt{\log n})\gamma + \delta^2\Big). \tag{A.3}$$

A.3 Concentration of random subgaussian Gram matrices

A key result in our proof is the following concentration bound on the Gram matrix of centered, subgaussian, independent random variables.

Lemma 1. *For some absolute constant $c_* > 0$, for $a\in[n]$ let E_a be centered, independent random vectors in $\mathbb{R}^d$, $E_a \sim subg(\Sigma_a)$. Let $\mathbf{E} := \begin{bmatrix}\cdots\\ E_a^T\\ \cdots\end{bmatrix} \in \mathbb{R}^{n\times d}$ then $\forall t \geqslant 0$*

$$\mathbb{P}\left[|\mathbf{E}\mathbf{E}^T - \mathbb{E}\left[\mathbf{E}\mathbf{E}^T\right]|_{op} \geqslant 2\max_{a\in[n]}|\Sigma_a|_F\sqrt{t} + 2\max_{a\in[n]}|\Sigma_a|_{op}t\right] \leqslant 9^n 2e^{-c_* t}. \tag{A.4}$$

B Main proofs

B.1 Proof of Proposition 1: identifiability

Suppose that $X_1, ..., X_n$ are $(\mathcal{G}, \boldsymbol{\mu}, \delta)$-clustered with $|\mathcal{G}| = K$, and $\rho(\mathcal{G}, \boldsymbol{\mu}, \delta) > 4$. Then we remark that for $(a, b) \in [n]^2$, $a \overset{\mathcal{G}}{\sim} b$ is equivalent to $|\nu_a - \nu_b|_2 \leqslant 2\delta$ because:

- if $a \overset{\mathcal{G}}{\sim} b$ then there exist $k \in [K]$ such that $|\nu_a - \nu_b|_2 \leqslant |\nu_a - \mu_k|_2 + |\mu_k - \nu_b|_2 \leqslant 2\delta$
- if $a \overset{\mathcal{G}}{\not\sim} b$ then there exist $(k, l) \in [K]^2$ such that $|\nu_a - \nu_b|_2 \geqslant |\mu_k - \mu_l|_2 - |\nu_a - \mu_k|_2 - |\nu_b - \mu_l|_2 > 4\delta - 2\delta > 2\delta$.

Now suppose there exist $\mathcal{G}'$ such that $X_1, ..., X_n$ are $(\mathcal{G}', \boldsymbol{\mu}', \delta')$-clustered with $|\mathcal{G}'| = K$ and $\rho(\mathcal{G}', \boldsymbol{\mu}', \delta') > 4$. By symmetry we can assume $\delta' \leqslant \delta$, and the previous remark shows that $\mathcal{G}'$ is a sub-partition of $\mathcal{G}$, ie $\mathcal{G}$ preserves the structure of $\mathcal{G}'$. But since $|\mathcal{G}| = |\mathcal{G}'|$ this implies $\mathcal{G} = \mathcal{G}'$. □

B.2 Exact recovery with high probability

The proof for Theorem 1 (respectively Theorem 2) is a composition of Theorem 3 (respectively Theorem 4) and Proposition 4.
In this section, under Hypothesis 1, we have $\forall k \in [K], \forall a \in G_k : X_a \sim \text{subg}(\Sigma_a)$. For $k \in [K]$, we define $\sigma_k^2 := \max_{a \in G_k} |\Sigma_a|_{op} \leqslant \sigma^2, \mathcal{V}_k^2 := \max_{a \in G_k} |\Sigma_a|_F \leqslant \mathcal{V}^2, \gamma_k^2 := \max_{a \in G_k} |\Sigma_a|_* \leqslant \gamma^2$.
A number of proofs in this section are adapted from the proof ensemble of [5]. In it the authors use a latent model for variable clustering. A comparable model in this work would require to impose the following conditions on $X_1, ..., X_n$: identically distributed variables within a group (implying $\delta = 0$) and isovolumic, Gaussian distributions.

B.2.1 Proof of Theorem 3

In this theorem we only need to consider $B \in \mathcal{C}_K$, but the proof of Theorem 4 is similar to this one, hence we will start by considering the more general $B \in \mathcal{C}$ and use $B \in \mathcal{C}_K$ at a later stage of the proof. Thus we want to prove that under some conditions, with high probability:

$$\langle \widehat{\Lambda} - \widehat{\Gamma}, B^* - B \rangle > 0 \text{ for all } B \in \mathcal{C} \setminus \{B^*\} \tag{B.1}$$

For $(a, b) \in G_k \times G_l$ for $(k, l) \in [K]^2$, let:

$$\begin{aligned}
(S_1)_{ab} &:= -|\mu_k - \mu_l|_2^2/2 \\
(W_1)_{ab} &:= \langle \nu_a - \mu_k, \nu_b - \mu_l \rangle \\
(W_2)_{ab} &:= \langle \mu_k - \nu_a + \nu_b - \mu_l + E_b - E_a, \mu_k - \mu_l \rangle \\
(W_3)_{ab} &:= \langle E_b - E_a, \nu_a - \mu_k + \mu_l - \nu_b \rangle \\
(W_4)_{ab} &:= (\langle E_a, E_b \rangle - \Gamma_{ab}) \\
(W_5)_{ab} &:= (\Gamma - \widehat{\Gamma})_{ab}
\end{aligned} \tag{B.2}$$

Lemma 2. *Proving* (B.1) *reduces to proving*

$$\langle S_1 + W_1 + W_2 + W_3 + W_4 + W_5, B^* - B \rangle > 0 \text{ for all } B \in \mathcal{C} \setminus \{B^*\}. \tag{B.3}$$

The proof for Lemma 2 is found in section B.2.3. So we need only concern ourselves with the quantities $S_1, W_1, W_2, W_3, W_4, W_5$. The term S_1 contains our uncorrupted signal and since $\langle S_1, B^* \rangle = 0$ it writes:

$$\langle S_1, B^* - B \rangle = \sum_{1 \leqslant k \neq l \leqslant K} \frac{1}{2} |\mu_k - \mu_l|_2^2 |B_{G_k G_l}|_1 \tag{B.4}$$

The other parts are noisy and must be controlled. The term W_2 is a simple subgaussian form controlled through the following lemma, proved in section B.2.4:

Lemma 3. *For $c_2' > 0$ absolute constant, with probability greater than $1 - 1/n$:*

$$\forall B \in \mathcal{C}, \quad |\langle W_2, B^* - B \rangle| \leqslant \sum_{1 \leqslant k \neq l \leqslant K} \Big(2\delta + \sqrt{c_2'(\log n)(\sigma_k^2 + \sigma_l^2)} \Big) |\mu_k - \mu_l|_2 |B_{G_k G_l}|_1. \tag{B.5}$$

To control the other noisy terms we now introduce a deterministic result:

Lemma 4. *For any symmetric matrix $W \in \mathbb{R}^{n \times n}$ we have:*

$$\forall B \in \mathcal{C}, \quad |\langle W, B^* - B \rangle| \leqslant 6 |B^* W|_\infty \sum_{1 \leqslant k \neq l \leqslant K} |B_{G_k G_l}|_1 + |W|_{op} \Big[\sum_{1 \leqslant k \neq l \leqslant K} |B_{G_k G_l}|_1 / m + (\mathrm{tr}(B) - K) \Big]. \tag{B.6}$$

The proof for Lemma 4 will be found in [5], p.21-22 until eq. (58).
As $B^* 1 = 1$ and $B^* \geqslant 0$, $|B^* W|_\infty \leqslant |W|_\infty$ so we use the lemma on terms W_1 and W_3 by bounding $|W|_\infty$ and $|W|_{op}$: for the term W_1 we use $|W_1|_\infty \leqslant \delta^2$ so $|W_1|_{op} \leqslant \delta^2 \sqrt{n}$. To control the term W_3, we use the subgaussian tail bound of (B.25) with $|\nu_a - \mu_k + \mu_l - \nu_b|_2 \leqslant 2\delta$ and a union bound over $(a, b) \in [n]^2$. We get that for $c_3' > 0$ absolute constant, with probability greater than $1 - 1/n$, $|W_3|_\infty \leqslant \sqrt{c_3'(\log n)\sigma^2 \delta^2}$ and $|W_3|_{op} \leqslant \sqrt{c_3'(\log n)\sigma^2 \delta^2} \times \sqrt{n}$ therefore with probability greater than $1 - 1/n$, $\forall B \in \mathcal{C}$:

$$|\langle W_1, B^* - B \rangle| \leqslant \delta^2 \Big[\sum_{1 \leqslant k \neq l \leqslant K} |B_{G_k G_l}|_1 (6 + \frac{\sqrt{n}}{m}) + \sqrt{n} (\mathrm{tr}(B) - K)_+ \Big] \tag{B.7}$$

$$|\langle W_3, B^* - B \rangle| \leqslant \sqrt{c_3'(\log n)\sigma^2 \delta^2} \Big[\sum_{1 \leqslant k \neq l \leqslant K} |B_{G_k G_l}|_1 (6 + \frac{\sqrt{n}}{m}) + \sqrt{n} (\mathrm{tr}(B) - K)_+ \Big] \tag{B.8}$$

For the term W_4 we introduce the following lemma, proved in section B.2.5:

Lemma 5. *For $c_4', c_4'' > 0$ absolute constants, with probability larger than $1 - 2/n$:*

$$\forall B \in \mathcal{C}, \quad |\langle W_4, B^* - B \rangle| \leqslant \Big[6c_4' (\mathcal{V}^2 \sqrt{\log n} + \sigma^2 \log n)/\sqrt{m} + c_4'' (\mathcal{V}^2 \sqrt{n} + \sigma^2 n)/m \Big] \sum_{1 \leqslant k \neq l \leqslant K} |B_{G_k G_l}|_1 + (\mathrm{tr}(B) - K)_+ c_4'' (\mathcal{V}^2 \sqrt{n} + \sigma^2 n). \tag{B.9}$$

Lastly as the term W_5 is diagonal we have $|W_5|_{op} = |W_5|_\infty$ and $|B^* W_5|_\infty \leqslant |W_5|_\infty / m$ therefore:

$$\forall B \in \mathcal{C}, \quad |\langle W_5, B^* - B \rangle| \leqslant |W_5|_\infty \Big[\frac{7}{m} \sum_{1 \leqslant k \neq l \leqslant K} |B_{G_k G_l}|_1 + (\mathrm{tr}(B) - K)_+ \Big] \tag{B.10}$$

Using those controls of W_1, W_2, W_3, W_4, W_5, in combination in a union bound in (B.3) we get for $c_1' > 0$ absolute constant, with probability greater than $1 - c_1'/n$: $\forall B \in \mathcal{C}$,

$$\langle S_1 + W_1 + W_2 + W_3 + W_4 + W_5, B^* - B\rangle \geqslant \sum_{1\leqslant k\neq l\leqslant K} \Big[\frac{1}{2}|\mu_k - \mu_l|_2^2 - \Big(2\delta + \sqrt{2c_2'(\log n)\sigma^2}\Big)|\mu_k - \mu_l|_2$$
$$- (6c_4'\frac{\mathcal{V}^2\sqrt{\log n} + \sigma^2\log n}{\sqrt{m}} + c_4''\frac{\mathcal{V}^2\sqrt{n} + \sigma^2 n}{m}) - \frac{7}{m}|W_5|_\infty - (6 + \frac{\sqrt{n}}{m})(\delta^2 + \sqrt{c_3'(\log n)\sigma^2\delta^2})\Big]|B_{G_kG_l}|_1$$
$$- (\mathrm{tr}(B) - K)_+[c_4''(\mathcal{V}^2\sqrt{n} + \sigma^2 n) + (\delta^2 + \sqrt{c_3'(\log n)\sigma^2\delta^2})\sqrt{n} + |W_5|_\infty] \tag{B.11}$$

We now use the fact that for this theorem we are only considering $B \in \mathcal{C}_K$, ie matrices such that $\mathrm{tr}(B) = K$ so we can discard the last line of (B.11). In this particular context we can improve the control provided by Lemma 4 for W_5: as $\mathrm{tr}(B^*) = K$, we have for $\alpha \in \mathbb{R}$: $|\langle W_5, B^* - B\rangle| \leqslant |\langle W_5 - \alpha I_n, B^* - B\rangle| + |\alpha(\mathrm{tr}(B) - K)|$. So by choosing $\alpha = (\max_a(W_5)_{aa} + \min_a(W_5)_{aa})/2$, we have $|W_5 - \alpha I_n|_{op} = |W_5 - \alpha I_n|_\infty = |W_5|_V/2$ and therefore:

$$\forall B \in \mathcal{C}_K \quad |\langle W_5, B^* - B\rangle| \leqslant |W_5|_V\frac{7}{2m}\sum_{1\leqslant k\neq l\leqslant K}|B_{G_kG_l}|_1 \tag{B.12}$$

In consequence we can replace $|W_5|_\infty$ by $|W_5|_V/2$ in the second line of (B.11), and with another union bound, by assumption we replace $|W_5|_V/2$ by $\bar{\gamma}_n^2/2$.
Lastly Lemma 3 p. 17 from [5] shows the only matrix in $\mathcal{C}_K$ whose support is included in $supp(B^*)$ is B^*, therefore $B \in \mathcal{C}_K \setminus \{B^*\}$ implies $\sum_{1\leqslant k\neq l\leqslant K}|B_{G_kG_l}|_1 > 0$. Hence for $c_2 > 0$ absolute constant, the following condition on $\Delta(\boldsymbol{\mu})$ is sufficient to ensure exact recovery with probability larger than $1 - c_1/n$:

$$\Delta^2(\boldsymbol{\mu}) \geqslant c_2\big[\sigma^2 m\log n + \mathcal{V}^2\sqrt{m\log n} + \mathcal{V}^2\sqrt{n} + \sigma^2 n + \bar{\gamma}_n^2 + \delta^2(\sqrt{n} + m)\big] \times \frac{1}{m} \tag{B.13}$$

This concludes the proof for Theorem 3. □

B.2.2 Proof of Theorem 4: adaptive exact recovery

In this Theorem we need to take into account the additional penalization term $\widehat{\kappa}\,\mathrm{tr}(B)$. Notice it is equivalent to a correction by $\widehat{\kappa}I_n$ of our estimator $\widehat{\Lambda} - \widehat{\Gamma}$, therefore for $B \in \mathcal{C}$, $\langle \widehat{\Lambda} - \widehat{\Gamma} - \widehat{\kappa}I_n, B^* - B\rangle = \langle \widehat{\Lambda} - \widehat{\Gamma}, B^* - B\rangle + \widehat{\kappa} \times (\mathrm{tr}(B) - K)$. Therefore for Theorem 4 we can follow the same proof as in Theorem 3 until establishing (B.11), at which point we can use a union bound to use the assumption $|W_5|_\infty \leqslant \bar{\gamma}_n^2$. Consequently we have with probability greater than $1 - c_1'/n$: $\forall B \in \mathcal{C}$,

$$\langle S_1 + W_1 + W_2 + W_3 + W_4 + W_5, B^* - B\rangle \geqslant \sum_{1\leqslant k\neq l\leqslant K} \Big[\frac{1}{2}|\mu_k - \mu_l|_2^2 - \Big(2\delta + \sqrt{2c_2'(\log n)\sigma^2}\Big)|\mu_k - \mu_l|_2$$
$$- (6c_4'\frac{\mathcal{V}^2\sqrt{\log n} + \sigma^2\log n}{\sqrt{m}} + c_4''\frac{\mathcal{V}^2\sqrt{n} + \sigma^2 n}{m}) - \frac{7}{m}\bar{\gamma}_n^2 - (6 + \frac{\sqrt{n}}{m})(\delta^2 + \sqrt{c_3'(\log n)\sigma^2\delta^2})\Big]|B_{G_kG_l}|_1$$
$$- (\mathrm{tr}(B) - K)_+[c_4''(\mathcal{V}^2\sqrt{n} + \sigma^2 n) + (\delta^2 + \sqrt{c_3'(\log n)\sigma^2\delta^2})\sqrt{n} + \bar{\gamma}_n^2] + \widehat{\kappa}(\mathrm{tr}(B) - K) \tag{B.14}$$

Using the assumption (A.1) of Theorem 4 there exist $c_2' > 0$ such that with probability greater than $1 - c_1'/n$: $\forall B \in \mathcal{C}$,

$$\langle S_1 + W_1 + W_2 + W_3 + W_4, B^* - B\rangle \geqslant c_2'\Delta^2(\boldsymbol{\mu})\sum_{1\leqslant k\neq l\leqslant K}|B_{G_kG_l}|_1$$
$$- (\mathrm{tr}(B) - K)_+[c_4''(\mathcal{V}^2\sqrt{n} + \sigma^2 n) + (\delta^2 + \sqrt{c_3'(\log n)\sigma^2\delta^2})\sqrt{n} + \bar{\gamma}_n^2] + \widehat{\kappa}(\mathrm{tr}(B) - K) \tag{B.15}$$

From here, when $\mathrm{tr}(B) > K$, the left-hand side of (A.2) is sufficient to ensure recovery. When $\mathrm{tr}(B) = K$, we already established that $\sum_{1\leqslant k\neq l\leqslant K} |B_{G_kG_l}|_1 > 0$ for all matrices $B \in \mathcal{C}_K \setminus \{B^*\}$ so (A.1) is sufficient in that case. Lastly note that $K - \mathrm{tr}(B) \leqslant \frac{1}{m}\sum_{1\leqslant k\neq l\leqslant K} |B_{G_kG_l}|_1$ (see [5] eq. (57) p.21) so the right-hand side of (A.2) is sufficient condition for recovery when $\mathrm{tr}(B) - K < 0$. This concludes the proof of Theorem 4. □

B.2.3 Proof of Lemma 2

$$
\begin{aligned}
(\widehat{\Lambda} - \widehat{\Gamma})_{ab} &= \langle X_a, X_b\rangle - \widehat{\Gamma}_{ab} = \langle \nu_a, \nu_b\rangle + \langle \nu_a, E_b\rangle + \langle \nu_b, E_a\rangle + \langle E_a, E_b\rangle - \widehat{\Gamma}_{ab} && \text{(B.16)}\\
&= \langle \nu_a, \nu_b\rangle + \langle \nu_a - \nu_b, E_b - E_a\rangle + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} && \text{(B.17)}\\
&= \langle \nu_a, \nu_b\rangle + \langle \mu_k - \mu_l, E_b - E_a\rangle + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} && \text{(B.18)}\\
&= -\langle \mu_k, \mu_l\rangle + \langle \nu_a - \mu_k, \nu_b - \mu_l\rangle + \langle \nu_a, \mu_l\rangle + \langle \mu_k, \nu_b\rangle \\
&\quad + \langle \mu_k - \mu_l, E_b - E_a\rangle + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} && \text{(B.19)}\\
&= -(S_1)_{ab} - \frac{1}{2}(|\mu_k|_2^2 + |\mu_l|_2^2) + (W_1)_{ab} + \langle \nu_a, \mu_l\rangle + \langle \mu_k, \nu_b\rangle \\
&\quad + \langle \mu_k - \mu_l, E_b - E_a\rangle + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} && \text{(B.20)}\\
&= -(S_1)_{ab} - \frac{1}{2}(|\mu_k|_2^2 + |\mu_l|_2^2) + (W_1)_{ab} + \langle \nu_a, \mu_k\rangle + \langle \mu_l, \nu_b\rangle \\
&\quad + \langle \mu_k - \mu_l, \nu_b - \nu_a + E_b - E_a\rangle + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} && \text{(B.21)}\\
&= -(S_1)_{ab} - \frac{1}{2}(|\mu_k|_2^2 + |\mu_l|_2^2) + (W_1)_{ab} + \langle \nu_a, \mu_k\rangle + \langle \mu_l, \nu_b\rangle \\
&\quad + 2(S_1)_{ab} + (W_2)_{ab} + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} && \text{(B.22)}
\end{aligned}
$$

Now since $(\langle \nu_a, \mu_k\rangle)_{(a,b)\in[n]^2} = (\langle \nu_a, \mu_k\rangle)_{a\in[n]} \times 1_n^T$, $(|\mu_k|_2^2)_{(a,b)\in[n]^2} = (|\mu_k|_2^2)_{a\in[n]} \times 1_n^T$, $(\langle \nu_b, \mu_l\rangle)_{(a,b)\in[n]^2} = 1_n \times (\langle \nu_b, \mu_l\rangle)_{b\in[n]}$, $(|\mu_l|_2^2)_{(a,b)\in[n]^2} = 1_n \times (|\mu_l|_2^2)_{b\in[n]}$, $(\langle \nu_a, E_a\rangle)_{(a,b)\in[n]^2} = (\langle \nu_a, E_a\rangle)_{a\in[n]} \times 1_n^T$, $(\langle \nu_b, E_b\rangle)_{(a,b)\in[n]^2} = 1_n \times (\langle \nu_b, E_b\rangle)_{b\in[n]}$ and since $B1_n = B^*1_n = (1_n^T B)^T = (1_n^T B^*)^T = 1_n$, we have:

$$
\langle \widehat{\Lambda} - \widehat{\Gamma}, B^* - B\rangle = \langle S_1 + W_1 + W_2 + W_3 + W_4 + W_5, B^* - B\rangle \qquad \text{(B.23)}
$$

□

B.2.4 Proof of Lemma 3: control of $|\langle W_2, B^* - B\rangle|$

By definition, $(W_2)_{ab} = 0$ when $k = l$ and $(B^*)_{ab} = 0$ when $k \neq l$ so we have $\langle W_2, B^*\rangle = 0$. Let $\langle A, B\rangle_{G_kG_l} = \sum_{(a,b)\in G_k\times G_l} A_{ab}B_{ab}$, we have:

$$
\langle W_2, B^* - B\rangle = -\langle W_2, B\rangle = -\sum_{1\leqslant k\neq l\leqslant K} \langle W_2, B\rangle_{G_kG_l} \leqslant \sum_{1\leqslant k\neq l\leqslant K} |W_{2|G_kG_l}|_\infty |B_{G_kG_l}|_1 \qquad \text{(B.24)}
$$

Let $(a, b) \in G_k \times G_l$, we look at $(W_2)_{ab} = \langle E_b - E_a - (\nu_a - \mu_k) + (\nu_b - \mu_l), \mu_k - \mu_l\rangle = \langle E_a - E_b, \mu_k - \mu_l\rangle + \langle -(\nu_a - \mu_k) + (\nu_b - \mu_l), \mu_k - \mu_l\rangle$. The term on the right is a constant offset bounded by $2\delta|\mu_k - \mu_l|_2$. Let $z := \mu_k - \mu_l$, by Lemma 7 $\langle E_a - E_b, z\rangle$ is a subgaussian variable with variance bounded by $(\sigma_k^2 + \sigma_l^2)|z|_2^2$ therefore its tails are characteristically bounded (see for example [20]), there exist $c_* > 0$ absolute constant such that $\forall t \geqslant 0$:

$$
\mathbb{P}\left[|\langle E_b - E_a, z\rangle| \geqslant |z|_2\sqrt{\sigma_k^2 + \sigma_l^2} \times t\right] \leqslant e^{1-c_*t^2} \qquad \text{(B.25)}
$$

This implies that $\forall t \geqslant 0, \mathbb{P}\left[|(W_2)_{ab}| \geqslant |\mu_k - \mu_l|_2(2\delta + \sqrt{\sigma_k^2 + \sigma_l^2} \times t)\right] \leqslant e^{1-c_* t^2}$. We conclude with a union bound over all $(a,b) \in G_k \times G_l$, a union bound over all $(k,l) \in [K]^2$, $k \neq l$ and by taking $t = \sqrt{(1+3\log n)/c_*}$. □

B.2.5 Proof of Lemma 5: control of $|\langle W_4, B^* - B\rangle|$

Recall $(W_4)_{ab} = \langle E_a, E_b\rangle - \Gamma_{ab}$. We will prove Lemma 5 by using the derivation of (B.6) combined with Lemma 1 for control of the operator norm and the following lemma for the remaining part.

Lemma 6. *For $c_4' > 0$ absolute constant, with probability greater than $1 - 1/n$:*

$$|B^* W_4|_\infty \leqslant c_4' \times (\mathcal{V}^2\sqrt{\log n} + \sigma^2 \log n)/\sqrt{m}. \tag{B.26}$$

Proof. Let $(a,b) \in G_k \times G_l$, we rewrite $(B^*W_4)_{ab}$ as the sum of the following two terms:

$$(B^*W_4)_{ab} = \frac{u_b}{|G_k|} \times \mathbf{1}_{k=l} + \langle \widetilde{E}_k, E_b\rangle \text{ with } \begin{cases} u_b & := |E_b|_2^2 - \Gamma_{bb} \\ \widetilde{E}_k & := \frac{1}{|G_k|}\sum_{c \in G_k, c \neq b} E_c \end{cases} \tag{B.27}$$

The bound for u_b uses Lemma 9: $\forall t \geqslant 0\ \mathbb{P}\left[||E_b|_2^2 - \mathbb{E}\,|E_b|_2^2| \geqslant \mathcal{V}_l^2\sqrt{t} + \sigma_l^2 t\right] \leqslant 2e^{-c_* t}$ so only the scalar product remains to be controlled. Notice that by Lemma 7, $\sqrt{|G_k|}\widetilde{E}_k$ is a centered subgaussian with variance-bounding matrix $\widetilde{\Sigma} = \frac{1}{|G_k|}\sum_{c \in G_k, c \neq b} \Sigma_c$, therefore $|\widetilde{\Sigma}|_F \leqslant \mathcal{V}_k^2$ and $|\widetilde{\Sigma}|_{op} \leqslant \sigma_k^2$. So using Lemma 9 again we find $\forall t \geqslant 0$:

$$\mathbb{P}\left[2|\sqrt{|G_k|}\langle \widetilde{E}_k, E_b\rangle| \geqslant \sqrt{2}\langle \widetilde{\Sigma}, \Sigma_b\rangle^{1/2}\sqrt{t} + |\widetilde{\Sigma}^{1/2}\Sigma_b^{1/2}|_{op} t\right] \leqslant 2e^{-c_* t} \tag{B.28}$$

Therefore using a union bound, then $\langle \widetilde{\Sigma}, \Sigma_b\rangle^{1/2} \leqslant \mathcal{V}_k\mathcal{V}_l \leqslant \mathcal{V}^2$ (Cauchy-Schwarz) and applying another union bound over all $(a,b) \in [n]^2$ with $t = (\log 4 + 3\log n)/c_*$ yields the result. □

We are ready to wrap-up the proof. From Lemma 1 applied to W_4, taking $t = (\log 2 + n\log 9 + \log n)/c_*$ there exists $c_4'' > 0$ absolute constant such that we have with probability greater than $1 - 1/n$: $|W_4|_{op} \leqslant c_4''(\mathcal{V}^2\sqrt{n} + \sigma^2 n)$. Now applying Lemma 4 to W_4:

$$|\langle W_4, B^* - B\rangle| \leqslant 6|B^*W_4|_\infty \sum_{1 \leqslant k \neq l \leqslant K} |B_{G_k G_l}|_1 + |W_4|_{op}\Big[\sum_{1 \leqslant k \neq l \leqslant K} |B_{G_k G_l}|_1/m + (\mathrm{tr}(B) - K)\Big] \tag{B.29}$$

Therefore combining the lemma with the derivations above and a union bound, we get with probability greater than $1 - 2/n$:

$$\begin{aligned} |\langle W_4, B^* - B\rangle| \leqslant &\Big[6c_4'(\mathcal{V}^2\sqrt{\log n} + \sigma^2\log n)/\sqrt{m} + c_4''(\mathcal{V}^2\sqrt{n} + \sigma^2 n)/m\Big] \sum_{1 \leqslant k \neq l \leqslant K} |B_{G_k G_l}|_1 \\ &+ (\mathrm{tr}(B) - K)_+ c_4''(\mathcal{V}^2\sqrt{n} + \sigma^2 n) \end{aligned} \tag{B.30}$$

This concludes the proof for Lemma 5. □

B.3 Proof of Proposition 4, Gamma estimator $\widehat{\Gamma}^{corr}$

Let $a \in G_k$, $b_1 \in G_{l_1}, b_2 \in G_{l_2}$, using (2.1) and $2|xy| \leqslant x^2 + y^2$ we have for $a \in [n]$:

$$|\widehat{\Gamma}_{aa} - \Gamma_{aa}| = |\langle X_a - X_{b_1}, X_a - X_{b_2}\rangle - \Gamma_{aa}| \leqslant U_1 + \frac{3}{2}U_2 + 2U_3 + 3U_4 \tag{B.31}$$

$$\begin{aligned} \text{where:} \quad U_1 &:= ||E_a|_2^2 - \Gamma_{aa}| \\ U_2 &:= |\nu_a - \nu_{b_1}|_2^2 + |\nu_a - \nu_{b_2}|_2^2 \\ U_3 &:= \sup_{(b,c)\in[n]^2} \langle \frac{\nu_a - \nu_c}{|\nu_a - \nu_c|_2}, E_b\rangle^2 \\ U_4 &:= \sup_{(b,c)\in[n]^2, b\neq c} |\langle E_b, E_c\rangle| \end{aligned}$$

Control of $U_1 = ||E_a|_2^2 - \Gamma_{aa}|$: by using the first inequality from Lemma 9 with $t = (2\log n + \log 2)/c_*$ there exists $c_1' > 0$ such that with probability greater than $1 - 1/n^2$:

$$U_1 \leqslant c_1' \times (\mathcal{V}_k^2\sqrt{\log n} + \sigma_k^2 \log n) \tag{B.32}$$

Control of $U_3 = \sup_{(b,c)\in[n]^2} \langle \frac{\nu_a - \nu_c}{|\nu_a - \nu_c|_2}, E_b\rangle^2$: write $z = (\nu_a - \nu_c)/|\nu_a - \nu_c|_2$ and $Y = \Sigma_b^{-1/2} E_b \sim \text{subg}(I_p)$ and $A = \Sigma_b^{1/2^T}(zz^T)\Sigma_b^{1/2}$, so that: $\langle z, E_b\rangle^2 = E_b^T zz^T E_b = Y^T AY$. Because $|z|_2 = 1$ and zz^T is symmetric of rank 1 we have $|A|_F = |A|_{op} = \text{tr}(A) \leqslant \sigma^2$ therefore we use Lemma 8 with $t = (4\log n + \log 2)/c_*$ and then a union bound over all $(b,c) \in [n]^2$ so that with probability greater than $1 - 1/n^2$:

$$U_3 \leqslant c_3' \times \sigma^2 \log n \tag{B.33}$$

Control of $U_4 = \sup_{(b,c)\in[n]^2, b\neq c} |\langle E_b, E_c\rangle|$: using the fact that E_b and E_c are independent and the second inequality of Lemma 9 with $t = (4\log n + \log 2)/c_*$, a union bound over all $(b,c) \in [n]^2$, there exists $c_4' > 0$ such that we have with probability greater than $1 - 1/n^2$:

$$U_4 \leqslant c_4' \times (\sigma^2 \log n + \mathcal{V}^2\sqrt{\log n}) \tag{B.34}$$

Control of $U_2 = |\nu_a - \nu_{b_1}|_2^2 + |\nu_a - \nu_{b_2}|_2^2$: here we use the requirement that all groups are of length at least $m \geqslant 3$, there exist $(a_1, a_2) \in G_k \setminus \{a\}$, $(c,d) \in ([n] \setminus \{a, a_1, a_2\})^2$, let $Z = (X_c - X_d)/|X_c - X_d|_2$. For $a_u \in \{a_1, a_2\}$ we have $\langle X_a - X_{a_u}, Z\rangle = \langle \nu_a - \nu_{a_u}, Z\rangle + \langle E_a - E_{a_u}, Z\rangle$. By independence and Lemma 7, $\langle E_a - E_{a_u}, Z\rangle$ is subgaussian with variance bounded by $2\sigma^2$. Therefore using the subgaussian tail bounds of (B.25) and a union bound, there exists $c_2' > 0$ absolute constant such that with probability over $1 - 1/n^2$: $V(a, a_1) \vee V(a, a_2) \leqslant 2\delta + c_2'\sigma\sqrt{\log n}$. Hence for $b_u \in \{b_1, b_2\}$ with probability over $1 - 1/n^2$:

$$|\langle X_a - X_{b_u}, X_c - X_d\rangle| \leqslant (2\delta + c_2'\sigma\sqrt{\log n})|X_c - X_d|_2 \tag{B.35}$$

Now suppose $l_1 \neq k$, choose $c \in G_k \setminus \{a\}, d \in G_{l_1} \setminus \{b_1\}$. We have $|X_c - X_d|_2 \leqslant |\mu_k - \mu_{l_1}|_2 + 2\delta + |E_c - E_d|_2$. We also have $\langle X_a - X_{b_1}, X_c - X_d\rangle = \langle \nu_a - \nu_{b_1} + E_a - E_{b_1}, \nu_c - \nu_d + E_c - E_d\rangle = \langle \mu_k - \mu_{l_1} + \delta_{ab} + E_a - E_{b_1}, \mu_k - \mu_{l_1} + \delta_{cd} + E_c - E_d\rangle$ for $\delta_{ab} = (\nu_a - \nu_{b_1}) - (\mu_k - \mu_{l_1})$ and $\delta_{cd} = (\nu_c - \nu_d) - (\mu_k - \mu_{l_1})$. Therefore:

$$\begin{aligned} |\langle X_a - X_{b_1}, X_c - X_d\rangle| &\geqslant |\mu_k - \mu_{l_1}|_2^2/2 - 4\delta|\mu_k - \mu_{l_1}|_2 \qquad \text{(B.36)} \\ &\quad - \frac{1}{2}\langle \frac{\mu_k - \mu_{l_1}}{|\mu_k - \mu_{l_1}|_2}, E_c + E_a - E_d - E_{b_1}\rangle^2 - 2\sup_{(b,c,d)\in[n]^3} \langle \frac{\delta_{cd}}{|\delta_{cd}|_2}, E_b\rangle^2 - 4U_4 - 12\delta^2 \\ &\geqslant |\mu_k - \mu_{l_1}|_2^2/2 - 4\delta|\mu_k - \mu_{l_1}|_2 - 8U_3' - 2U_3'' - 4U_4 - 12\delta^2 \qquad \text{(B.37)} \end{aligned}$$

where $U_3' = \sup_{(b,l)\in[n]\times[K]} \langle \frac{\mu_k-\mu_l}{|\mu_k-\mu_l|_2}, E_b\rangle^2$, $U_3'' = \sup_{(b,c,d)\in[n]^3} \langle \frac{\delta_{cd}}{|\delta_{cd}|_2}, E_b\rangle^2$.
So combining the last derivations:

$$|\mu_k-\mu_{l_1}|_2^2/2 - 4\delta|\mu_k-\mu_{l_1}|_2 \leqslant (2\delta + c_2'\sigma\sqrt{\log n})(|\mu_k-\mu_{l_1}|_2 + 2\delta + |E_c-E_d|_2) \\ +8U_3' + 2U_3'' + 4U_4 + 12\delta^2 \tag{B.38}$$

Notice that U_3', U_3'' can be controlled exactly as U_3 was, and simultaneously: for $c_3'' > 0$ absolute constant, with probability greater than $1-1/n^2$: $8U_3' + 2U_3'' \leqslant c_3''\sigma^2\log n$.
We now control $|E_c - E_d|_2$: notice that by Lemma 7, $E_c - E_d$ is $\mathrm{subg}(\Sigma_c + \Sigma_d)$. We have $\mathbb{E}\left[|E_c-E_d|_2^2\right] \leqslant |\Sigma_c+\Sigma_d|_* \leqslant 2\gamma^2$, $|\Sigma_c+\Sigma_d|_F \leqslant 2\mathcal{V}^2 \leqslant 2\sigma\gamma$ and $|\Sigma_c+\Sigma_d|_{op} \leqslant 2\sigma^2$. Therefore by the first inequality of Lemma 9 with $t = (4\log n + \log 2)/c_*$ and a union bound over all $(c,d)\in[n]^2$, there exists $c_2'' > 0$ absolute constant such that we have simultaneously with probability greater than $1-1/n^2$:

$$\sup_{(c,d)\in[n]^2} |E_c - E_d|_2 \leqslant c_2''\sqrt{\gamma^2 + \sigma\gamma\sqrt{\log n} + \sigma^2\log n} \leqslant c_2''(\gamma + \sigma\sqrt{\log n}) \tag{B.39}$$

Therefore with a union bound, with probability greater than $1-4/n^2$:

$$|\mu_k-\mu_{l_1}|_2^2/2 - (c_2'\sigma\sqrt{\log n} + 6\delta)|\mu_k-\mu_{l_1}|_2 \leqslant (2\delta + c_2'\sigma\sqrt{\log n})\big(2\delta + (\gamma+\sigma\sqrt{\log n})(c_2'' + \frac{c_3''}{c_2'} + \frac{4c_4'}{c_2'})\big) \\ + 12\delta^2 \tag{B.40}$$

Hence for $c_5' > 0$ absolute constant we have with probability greater than $1-4/n^2$: $|\mu_k-\mu_{l_1}|_2^2 \leqslant c_5'(\delta+\sigma\sqrt{\log n})(\delta+\sigma\sqrt{\log n}+\gamma)$. The same control can be derived simultaneously for $|\mu_k-\mu_{l_2}|_2^2$ by replacing $d\in G_{l_1}\setminus\{b_1\}$ by $d'\in G_{l_2}\setminus\{b_1,b_2\}$. We conclude that for $c_5'' > 0$ absolute constant, we have with probability greater than $1-4/n^2$:

$$U_2 \leqslant 2|\mu_k-\mu_{l_1}|_2^2 + 2|\mu_k-\mu_{l_2}|_2^2 + 16\delta^2 \leqslant c_5''(\delta+\sigma\sqrt{\log n})(\delta+\sigma\sqrt{\log n}+\gamma) \tag{B.41}$$

Therefore with a union bound over all four terms U_1, U_2, U_3, U_4 and $a\in[n]$, for $c_6, c_7 > 0$ absolute constants we have with probability greater than $1-c_6/n$: $|\widehat{\Gamma}-\Gamma|_\infty \leqslant c_7(\delta+\sigma\sqrt{\log n})(\delta+\sigma\sqrt{\log n}+\gamma)$.
This concludes the proof of Proposition 4 □

B.4 Proof of Proposition 2

For this proof we rely heavily on the proof of Theorem 3: let $\widehat{\Gamma} = 0$ so that $W_5 = \Gamma$, notice that W_3 and W_4 are centered. We take expectation of (B.3), therefore proving $\langle \Lambda+\Gamma, B^*-B\rangle > 0$ for all $B\in\mathcal{C}_K\setminus\{B^*\}$ is equivalent to proving:

$$\langle S_1 + W_1 + \mathbb{E}[W_2] + \Gamma, B^*-B\rangle > 0 \text{ for all } B\in\mathcal{C}_K\setminus\{B^*\} \tag{B.42}$$

Notice that for $(a,b)\in G_k\times G_l$, $\mathbb{E}[(W_2)_{ab}] \leqslant 2\delta|\mu_k-\mu_l|_2$. Using this in combination with other

arguments from the proof of Theorem 3, that is using (B.4), (B.7) and (B.12), we have $\forall B \in \mathcal{C}_K$:

$$\langle S_1, B^* - B\rangle = \sum_{1\leqslant k\neq l\leqslant K} \frac{1}{2}|\mu_k - \mu_l|_2^2 |B_{G_kG_l}|_1 \tag{B.43}$$

$$|\langle W_1, B^* - B\rangle| \leqslant \sum_{1\leqslant k\neq l\leqslant K} \delta^2(6 + \frac{\sqrt{n}}{m})|B_{G_kG_l}|_1 \tag{B.44}$$

$$|\langle \mathbb{E}\left[W_2\right], B^* - B\rangle| \leqslant \sum_{1\leqslant k\neq l\leqslant K} 2\delta|\mu_k - \mu_l|_2 |B_{G_kG_l}|_1 \tag{B.45}$$

$$|\langle W_5, B^* - B\rangle| \leqslant \sum_{1\leqslant k\neq l\leqslant K} \frac{7|\Gamma|_V}{2m}|B_{G_kG_l}|_1 \tag{B.46}$$

Thus we have:

$$\langle S_1 + W_1 + \mathbb{E}\left[W_2\right] + W_5, B^* - B\rangle \geqslant \sum_{1\leqslant k\neq l\leqslant K} \Big[\frac{1}{2}|\mu_k - \mu_l|_2^2 - 2\delta|\mu_k - \mu_l|_2 - \delta^2(6 + \frac{\sqrt{n}}{m}) - \frac{7|\Gamma|_V}{2m}\Big]|B_{G_kG_l}|_1 \tag{B.47}$$

Hence we deduce that there exist c_0 absolute constant such that if $\rho^2(\mathcal{G}, \boldsymbol{\mu}, \delta) > c_0(6 + \sqrt{n}/m)$ and $m\Delta^2(\boldsymbol{\mu}) > 8|\Gamma|_V$, then we have $\arg\max_{B\in\mathcal{C}_K}\langle \Lambda + \Gamma, B\rangle = B^*$. Lastly as B^* is in $\mathcal{C}_K^{\{0,1\}} \subset \mathcal{C}_K$, this concludes the proof. □

B.5 Proof of Proposition 3

Assume $X_1, ..., X_n$ is $(\mathcal{G}, \boldsymbol{\mu}, \delta)$-clustered with caracterizing matrix B^* and define the following:

- $\delta = 0$ implying maximum discriminating capacity for $\mathcal{G}$ ie $\rho(\mathcal{G}, \boldsymbol{\mu}, \delta) = +\infty$.
- Let

$$B^* := \begin{bmatrix} \boxed{\frac{1}{m}} & & \\ & \boxed{\frac{1}{m}} & \\ & & \boxed{\frac{1}{m}} \end{bmatrix} \in \mathcal{C}_K^{\{0,1\}} \qquad \text{and} \qquad B_1 := \begin{bmatrix} \boxed{2/m} & & \\ & \boxed{2/m} & \\ & & \boxed{\frac{1}{2m}} \end{bmatrix} \in \mathcal{C}_K^{\{0,1\}}$$

  where $\boxed{\frac{1}{m}}$ represents constant square blocks of size m and value $1/m$, and the other values in the matrices are zeros.
- $K = 3$ and for some $\Delta > 0$, $\mu_1 = (\Delta/\sqrt{2}, 0, 0)^T$ and $\mu_2 = (0, \Delta/\sqrt{2}, 0)^T$, $\mu_3 = (0, 0, \Delta/\sqrt{2})^T$ so that for $(a, b) \in G_k \times G_l$: $\Lambda_{ab} = \langle \mu_k, \mu_l\rangle = \Delta^2/2 \times \mathbf{1}\{a \overset{\mathcal{G}}{\sim} b\}$. Then $\Delta^2(\boldsymbol{\mu}) = \Delta^2$ and $\Lambda = (\Delta^2/2)mB^*$.
- For $\gamma_+ > \gamma_- > 0$ let $\Gamma = \text{diag}\,(\underbrace{\gamma_+, ..., \gamma_+}_{m}, \underbrace{\gamma_-, ..., \gamma_-}_{m}, \underbrace{\gamma_-, ..., \gamma_-}_{m})$

Then we have the following: $\langle B^*, \Gamma\rangle = \gamma_+ + 2\gamma_-$, $\langle B_1, \Gamma\rangle = 2\gamma_+ + \gamma_-$, $\langle B^*, \Lambda\rangle = \Delta^2/2 \times 3m$, $\langle B_1, \Lambda\rangle = \Delta^2/2 \times 2m$. Thus we have $\langle B^*, \Lambda + \Gamma\rangle < \langle B_1, \Lambda + \Gamma\rangle$ as soon as $m\Delta^2(\boldsymbol{\mu}) < 2(\gamma_+ - \gamma_-)$. This concludes the proof. □

C Subgaussian properties and controls

Lemma 7. *$\forall a \in [n]$ let $Y_a \sim subg(\Sigma_a)$, independent, $\Sigma_a \in \mathbb{R}^{d\times d}$ then*

$$Y = (Y_1^T, ..., Y_n^T)^T \sim subg(\operatorname{diag}(\Sigma_a)_{a\in[n]}), \tag{C.1}$$

$$Z = \sum_{a\in[n]} c_a Y_a \sim subg(\sum_{a\in[n]} c_a^2 \Sigma_a). \tag{C.2}$$

Proof. By independence for $z = \{z_1^T, ..., z_n^T\}^T \in \mathbb{R}^{nd}, z_a \in \mathbb{R}^d$ we have

$$\mathbb{E}\left[e^{z^T(Y-\mathbb{E}Y)}\right] = \prod_{a=1}^n \mathbb{E}\left[e^{z_a^T(Y_a-\mathbb{E}Y_a)}\right] \leqslant \prod_{a=1}^n e^{z_a^T\Sigma_a z_a/2} = e^{z^T \operatorname{diag}(\Sigma_a)_{a\in[n]} z/2}$$

$$\mathbb{E}\left[e^{z_1^T(Z-\mathbb{E}Z)}\right] = \prod_{a=1}^n \mathbb{E}\left[e^{z_1^T c_a(Y_a-\mathbb{E}Y_a)}\right] \leqslant \prod_{a=1}^n e^{z_1^T c_a^2\Sigma_a z_1/2} = e^{z_1^T(\sum_{a\in[n]} c_a^2\Sigma_a)z_1/2}$$

□

Lemma 8. *Hanson-Wright inequality for subgaussian variables*
Let Y be a centered random vector, $Y \sim subg(I_d)$, let A be a matrix of size $d \times d$. There exists $c_* > 0$ such that for any $t \geqslant 0$

$$\mathbb{P}\left[|Y^TAY - \mathbb{E}\left[Y^TAY\right]| \geqslant |A|_F\sqrt{t} + |A|_{op}t\right] \leqslant 2e^{-c_*t}. \tag{C.3}$$

Proof. A variation of the original Hanson-Wright inequality (Theorem 1.1 from [18]), it holds as $\sigma = 1$ bounds the subgaussian norm $|Y|_{\Psi_2} := \sup_{x\in\mathcal{S}_{d-1}} \sup_{p\geqslant 1} p^{-1/2}(\mathbb{E}|x^TY|^p)^{1/p}$, a consequence of Lemma 5.5 from [20]. □

Lemma 9. *Subgaussian quadratic forms*
Let E, E' be centered, independent random vectors, $E \sim subg(\Sigma)$, $E' \sim subg(\Sigma')$, then for $t \geqslant 0$

$$\mathbb{P}\left[||E|_2^2 - \mathbb{E}|E|_2^2| \geqslant |\Sigma|_F\sqrt{t} + |\Sigma|_{op}t\right] \leqslant 2e^{-c_*t} \tag{C.4}$$

$$\mathbb{P}\left[2|\langle E, E'\rangle| \geqslant \sqrt{2}\langle\Sigma, \Sigma'\rangle^{1/2}\sqrt{t} + |\Sigma^{1/2}\Sigma'^{1/2}|_{op}t\right] \leqslant 2e^{-c_*t}. \tag{C.5}$$

Proof. For the first inequality, we use Lemma 8 with $Y = \Sigma^{-1/2}E$ and $A = \Sigma$. As for the second inequality, by Lemma 7 we have $Y = (E^T\Sigma^{-1/2}, E'^T\Sigma'^{-1/2T})^T \sim \text{subg}(I_{2d})$. Then let us use Lemma 8 with

$$A = \begin{pmatrix} 0 & \Sigma^{1/2}\Sigma'^{1/2} \\ \Sigma'^{1/2^T}\Sigma^{1/2^T} & 0 \end{pmatrix}$$

Notice that $|A|_F^2 = 2\langle\Sigma, \Sigma'\rangle$ and $|A|_{op} \leqslant |\Sigma^{1/2}\Sigma'^{1/2}|_{op}$ so the results follow. □

Proof of Lemma 1: concentration of random subgaussian Gram matrices.
Let $W := \mathbf{E}\mathbf{E}^T - \mathbb{E}[\mathbf{E}\mathbf{E}^T]$. Using the epsilon-net method as in Lemma 4.2 from [17], let $\mathcal{N}$ be a 1/4-net for $\mathcal{S}_{n-1}$ such that $|\mathcal{N}| \leqslant 9^n$ (see Lemma 5.2 [20]), we have for $u, v \in \mathcal{S}_{n-1}^2$: $u^TWv \leqslant \max_{x\in\mathcal{N}} x^TWv + \frac{1}{4}\max_{u\in\mathcal{S}_{n-1}} u^TWv \leqslant \max_{x,y\in\mathcal{N}^2} x^TWy + \frac{1}{2}\max_{u,v\in\mathcal{S}_{n-1}^2} u^TWv$ hence

$$|W|_{op} \leqslant 2\max_{x,y\in\mathcal{N}^2} x^TWy \quad \text{and} \quad \mathbb{P}\left[|W|_{op} \geqslant t\right] \leqslant \sum_{x,y\in\mathcal{N}^2} \mathbb{P}\left[x^TWy \geqslant t/2\right] \tag{C.6}$$

Notice that this rewrites $x^T W y = \sum_{a=1}^n \sum_{b=1}^n x_a (E_a^T E_b - \Gamma_{ab}) y_b = (\sum_{a=1}^n E_a^T x_a)(\sum_{b=1}^n E_b^T y_b)^T - \mathbb{E}(\sum_{a=1}^n E_a^T x_a)(\sum_{b=1}^n E_b^T y_b)^T$. For $x, y \in \mathcal{N}^2$, let $x \otimes \Sigma^{1/2} := (x_1 \Sigma_1^{1/2}, ..., x_n \Sigma_n^{1/2})^T \in \mathbb{R}^{np \times p}$ and $Y = (E_1^T \Sigma_1^{-1/2}, ..., E_n^T \Sigma_n^{-1/2})^T \in \mathbb{R}^{np \times 1}$ (by Lemma 7 we have $Y \sim \text{subg}(I_{np})$). We have

$$x^T W y = Y^T (x \otimes \Sigma^{1/2})(y \otimes \Sigma^{1/2})^T Y - \mathbb{E}[Y^T (x \otimes \Sigma^{1/2})(y \otimes \Sigma^{1/2})^T Y] \tag{C.7}$$

Now define $A := (x \otimes \Sigma^{1/2})(y \otimes \Sigma^{1/2})^T$: we have $|A|_{op} \leqslant \max_{a \in [n]} |\Sigma_a|_{op}$ because for $z \in \mathbb{R}^p$, $|(x \otimes \Sigma^{1/2}) z|_2^2 = \sum_{b=1}^n x_b^2 |\Sigma_b^{1/2} z|_2^2 \leqslant \max_{a \in [n]} |\Sigma_a|_{op} |z|_2^2$. As for the Frobenius norm, by Cauchy-Schwarz: $|(x \otimes \Sigma^{1/2})(y \otimes \Sigma^{1/2})^T|_F^2 = \sum_{a=1}^n \sum_{b=1}^n x_a^2 y_b^2 |\Sigma_a^{1/2} \Sigma_b^{1/2}|_F^2 \leqslant \max_{a \in [n]} |\Sigma_a|_F^2$. Therefore using Lemma 8 on Y we have $\forall t \geqslant 0 : \mathbb{P}\left[|Y^T A Y - \mathbb{E}\left[Y^T A Y\right]| \geqslant \max_{a \in [n]} |\Sigma_a|_F \sqrt{t} + \max_{a \in [n]} |\Sigma_a|_{op} t\right] \leqslant 2e^{-ct}$. Hence in conjunction with (C.6) we conclude the proof. □