Statistically Optimal Clustering through Convex Optimisation

Martin Royer

NNT : 2018SACLS442

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Laboratoire d'accueil : Laboratoire de mathématiques d'Orsay, UMR 8628 CNRS

Spécialité de doctorat : Mathématiques appliquées

Martin ROYER

Optimalité statistique du partitionnement par l'optimisation convexe

Date de soutenance : le 13 novembre 2018

Après avis des rapporteurs : PIERRE LATOUCHE (Université Paris-Descartes)
STÉPHANE CHRÉTIEN (National Physical Laboratory)

Jury de soutenance :

FLORENTINA BUNEA	(Cornell University)	Codirecteur de thèse
BENOÎT CADRE	(ENS Rennes)	Examinateur
STÉPHANE CHRÉTIEN	(NPL)	Rapporteur
ROMAIN COUILLET	(Centrale-Supélec)	Examinateur
CHRISTOPHE GIRAUD	(LMO)	Directeur de thèse
PASCAL MASSART	(LMO)	Président

.

Remerciements

Ces travaux de doctorat sont le fruit d'une construction collective dans une organisation ouverte du monde de la recherche où peut-être plus que nulle part ailleurs, les progrès de l'individu doivent aux efforts d'autrui.

Sous la direction bienveillante de Christophe et Flori, je me suis vu offrir un formidable cap à suivre. I was confronted with demanding, multifaceted challenges in the course of a transatlantic, back-and-forth exploratory journey. Les épreuves en furent passionnantes, propices à la découverte, et comme il fallait composer avec mes propres limitations je ne doute pas leur avoir quelquefois fait perdre le sang-froid ; c'était bien malgré moi. Nevertheless at all times I always could rely on them, and owing to their patient merit, many times I did.

Nicolas Verzelen m'a pris sous son aile, a exercé de loin, en éminence grise et sans discontinuer son influence attentive, salutaire, quand rien ne l'y forçait.

During collaborative projects I had the opportunity of partnering with and learning from talented researchers : Jishnu Das, Xi (Rossi) Luo and exceptional Xin (Mike) Bing.

Dans le labyrinthe de la recherche mathématique j'ai été guidé par les conseils de chercheurs : Pascal Massart, Giles Hooker, Elisabeth Gassiat et d'autres d'une longue série, commencée quand j'ai rencontré Marc Hoffmann et Nicolas Fournier.

Pierre Latouche et Stéphane Chrétien ont eu la générosité (et l'infinie patience) de rapporter ces travaux. Romain Couillet et Benoît Cadre celle de participer à mon jury.

Ces travaux se sont effectués grâce au programme investissement d'avenir de l'Agence Nationale de la Recherche, à travers le projet IDI 2015 subventionné par l'IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

Working as a TA to Jacob Bien I was impressed with the command, tact and balance deployed in and outside the classroom that made for enlightening teaching.

Alors qu'un doctorat dans les *deux mondes* des universités de Cornell et Paris-Sud a des exigences techniques et administratives redoutables, il m'a été donné de toujours être au contact d'interlocuteurs aimables et compétents, dont Diana Drake, Hal Coghill et Sue Bishop, Frédéric Paulin, Valérie Lavigne, Florence Rey et Corentin Guéneron, Olivier Chaudet, Sylvain Faure et Suzanne Varet, mais d'autres encore.

Enfin le jeune chercheur côtoie surtout ses semblables, les doctorants, véritables marsouins de la grande armada mathématique qui découvrent, questionnent et se passionnent pour tout fraternellement : Maxime, Robert, charmante Elodie, Eugène, Joseph, Thomas, mon fidèle Luc, Augustin, Armand qui a conquis Mambrin, Thomas, fulgurant Guillaume, Camille, Pierre, Julien, l'insatiable Jeanne, Gabriele, Claire, Anthony, sage Salim, Hugo, Thibault, Romain, Pierre bourre-pif !, François, Gabriel si délicat, et encore Clément, Clémence, Emmanuel et l'impétueuse Milica, puis tant d'autres. Across the ocean their counterparts : undaunted Chang, Yichen, witty Daniel and the Ben's, Skyler and Wenyu, musical Sarah, Keegan and Xiaoyun and many others.

A tous je présente mes remerciements chaleureux et sincères.

Je veux encore saluer ici l'aide que j'ai reçue quelques années auparavant de mon professeur de classes préparatoires, un homme au grand courage, Jean-Claude Jacquens.

A mes parents.

Table des matières

Chapitre 1

Introduction

Contents

Séparer un ensemble en parties distinctes, porteuses de sens pour ses entités, c'est ce que font la grammaire avec les mots de la langue, le cerveau en distinguant les signaux perçus, la biologie en classifiant les gènes, la médecine pour les maladies humaines. La tâche du partitionnement, c'est de faire apparaître de la structure dans un ensemble composite d'entités. Ses applications directes aux sciences et problématiques modernes sont omniprésentes. On le trouve dans les domaines de l'expression génomique ou protéinique lorsqu'il s'agit d'identifier un gène ou une séquence défectueuse, dans la segmentation d'images pour détecter certaines formes ou évènements automatiquement, on l'utilise pour comprendre la parcellation du cerveau grâce à l'imagerie par résonance magnétique fonctionnelle, pour modéliser des interactions sociales et encore dans l'analyse de documents, la robotique, l'astronomie, les problématiques de quantification et de réduction de dimension, sans parler d'innombrables applications industrielles.

1 Le problème du partitionnement

Soit $\mathcal{X}$ ensemble fini d'un espace euclidien.

Définition 1.1 (Problème du partitionnement)**.** *On appelle problème du partitionnement la tâche de séparer $\mathcal{X}$ en parties selon un critère de similitude, c'est-à-dire de trouver une partition $\mathcal{G}$ de $\mathcal{X}$ telle que tout élément G de $\mathcal{G}$ regroupe des éléments de $\mathcal{X}$ similaires.*

Le sens des mots *similitude* et *similaires* doivent être précisées. On peut dire pour l'instant qu'on dispose d'une mesure de similarité entre les éléments de $\mathcal{X}$, et que pour deux éléments de $\mathcal{X}$ plus cette mesure est élevée, plus la similarité entre les entités est grande. Une formulation plus précise sera introduite au chapitre 2, qui considère que deux éléments sont proches si leurs rapports à tous les autres éléments de l'ensemble sont comparables. On pourrait ainsi dire que le problème du partitionnement consiste à faire des groupes dont les éléments soient à la fois similaires à l'intérieur d'un groupe, et dissimilaires de ceux d'autres groupes.

1.1 Partitionnement statistique et computationnel

Les travaux présentés ici s'articulent autour de la problématique suivante :

> **Problématique.** *On souhaite présenter et analyser des algorithmes de complexité polynomiale et optimaux pour le partitionnement exact dans les cadres statistiques du partitionnement de points et de variables, avec applications notamment à la biologie.*

Admettons pour l'instant qu'il existe une partition $\mathcal{G}$ de $\mathcal{X}$ qui s'exprime à travers ce qu'on observe, c'est-à-dire qui structure d'une façon ou d'une autre les rapports entre éléments de $\mathcal{X}$. C'est cette partition qui nous intéresse par définition, mais son niveau d'expression pourrait être suffisamment fort pour qu'on puisse aisément la retrouver, ou au contraire trop faible de sorte que cela soit strictement impossible. Sous quelles conditions est-on en mesure de retrouver $\mathcal{G}$? On peut se figurer pour l'instant que la dissimilarité entre les groupes est résumée par un coefficient $\Delta > 0$, indiquant la force du signal, et fonctionnant ainsi : plus Δ est élevé, plus les groupes sont uniformément séparés à travers les lois des observations, et donc plus le problème du partitionnement est simple à résoudre. Plus formellement, si dans un modèle statistique $\mathcal{M}_\Delta$ les lois des éléments de chaque groupe de $\mathcal{G}$ sont suffisamment distinguables, dis-similaires les unes des autres, on devrait pouvoir retrouver $\mathcal{G}$, avec grande probabilité, en analysant les observations. Dans cette approche probabiliste on comparera souvent cette force du signal avec une quantité mesurant la force du bruit : par exemple on représentera par $\sigma > 0$ l'intensité dans la direction de variance maximale pour les observations.

Example 1.1 (Contrôle de trois gaussiennes)**.** *Dans le cas du partitionnement de points (voir exemple figure 1.1), on suppose que des points sont distribués équitablement selon trois gaussiennes dont les centres sont à distance euclidienne au moins $\Delta > 0$, et de déviations maximales inférieures à σ. Si la dimension est petite, alors on montre qu'un estimateur computationnel retrouve $\mathcal{G}$ avec grande probabilité dès lors que*[1]

$$\Delta^2 \gtrsim \sigma^2 \log n. \tag{1.1}$$

[1] dans ce manuscrit, $a \gtrsim b$ signifie qu'il existe $c > 0$ constante numérique telle que $a \geqslant c \times b$

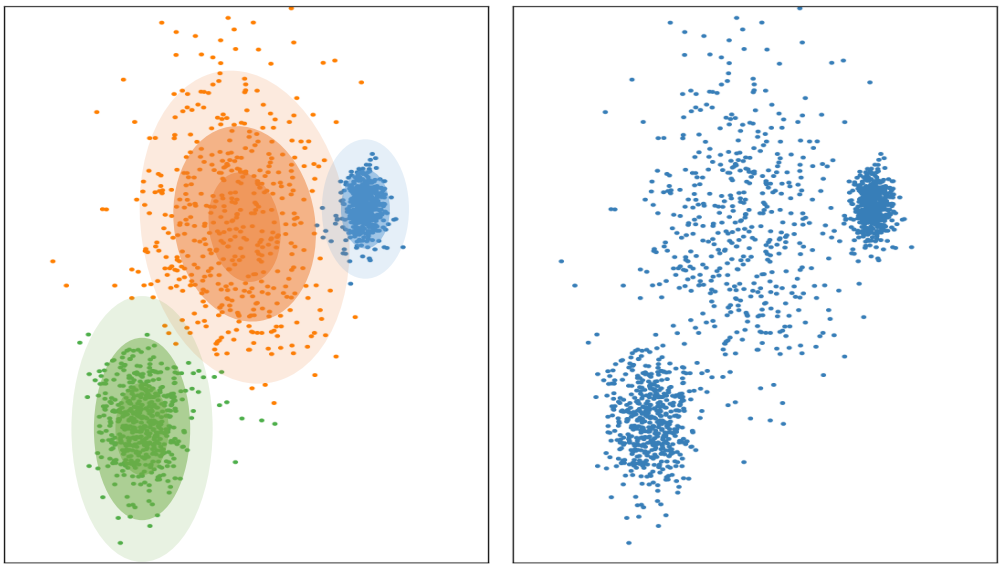


FIGURE 1.1 – Le partitionnement de points dans $\mathbb{R}^2$. A droite les observations qui sont les réalisations de trois gaussiennes anisotropiques et anisovolumiques, à gauche une partition indiquée par la couleur.

Si l'on est capable de retrouver $\mathcal{G}$, une autre question essentielle apparaît : quelle est la manière optimale de procéder ? Soit avec ce qu'on a introduit précédemment, quelle est la plus petite séparation Δ qui permette encore de retrouver $\mathcal{G}$ avec grande probabilité ? C'est l'approche de l'analyse dite minimax. On appelle estimateur une fonction des observations $\widehat{\mathcal{G}} : (X_1, \dots, X_n) \to \widehat{\mathcal{G}}(X_1, \dots, X_n)$; une façon de caractériser cette capacité à retrouver $\mathcal{G}$ pour n'importe quel estimateur dans le modèle statistique $\mathcal{M}_\Delta$ est d'introduire et de se comparer au risque minimax, défini comme

$$\mathcal{R}_\Delta := \inf_{\widehat{\mathcal{G}}} \sup_{\mathbb{P} \in \mathcal{M}_\Delta} \mathbb{P}\left[\widehat{\mathcal{G}}(X_1, \dots, X_n) \neq \mathcal{G}\right]. \tag{1.2}$$

On verra que ce risque tend vers 0 quand Δ grandit à une certaine vitesse fonction des paramètres du problème (dimension de l'espace, nombre de groupes dans $\mathcal{G}$, etc), et peut être minoré par une constante strictement positive sinon.

L'exemple de résultat qu'on a reproduit ci-dessus est un cas particulier (petite dimension, nombre de structures constant). Plus généralement peut-on produire un estimateur qui retrouve $\mathcal{G}$ en temps polynomial à la vitesse minimax ? Ou il y a-t-il un écart entre la vitesse théorique $\mathcal{R}_\Delta$ (1.2), et ce qu'un algorithme polynomial peut réellement accomplir ? En établissant ces éléments, on est amené à se poser de nombreuses questions connexes à celle du partitionnement, par exemple la question de l'influence du nombre de groupes $|\mathcal{G}|$ ou de la plus petite taille de groupe $\min_{G \in \mathcal{G}} |G|$ dans notre capacité à retrouver le partitionnement des données. La condition de séparation entre les différents groupes peut-elle être exprimée de façon purement locale, c'est-à-dire entre groupes deux-à-deux, ou ne peut-on l'exprimer que globalement, pour un ensemble de points donné ?

Enfin il est indispensable de se demander quelles sont les conditions qui permettent à la partition $\mathcal{G}$ évoquée plus haut d'être identifiable.

Remarque 1.1 (A propos du partitionnement exact)**.** *Le risque minimax considéré étudie l'espérance de la fonction de perte $\mathcal{G}_1, \mathcal{G}_2 \mapsto \mathbf{1}\{\mathcal{G}_1 \neq \mathcal{G}_2\}$, qui correspond au problème du partitionnement exact : elle indique 1 si l'on a retrouvé la partition recherchée, 0 sinon. Ce choix peut paraître exigeant ou peu réaliste, c'est selon, néanmoins on montrera par la suite en s'appuyant sur nos résultats qu'il n'est pas si pénalisant qu'il y paraît.*

1.2 Cadres d'étude

On considère n points (ou observations) $X_1, \dots, X_n$, les réalisations de vecteurs aléatoires indépendants à valeurs dans $\mathbb{R}^p$, et on note :

$$\forall a \in 1 \dots n, \quad X_a = (X_a^{(1)}, \dots, X_a^{(p)})^T \in \mathbb{R}^p \tag{1.3}$$

$$\forall i \in 1 \dots p, \quad X^{(i)} = (X_1^{(i)}, \dots, X_n^{(i)}) \in \mathbb{R}^n, \tag{1.4}$$

et soit la matrice des observations :

$$\mathbf{X} := (X_a^{(i)})_{a=1..n}^{i=1..p} \in \mathbb{R}^{n \times p}. \tag{1.5}$$

Schématiquement, on peut imaginer faire du partitionnement dans deux contextes à distinguer : en partitionnant les indices des lignes de la matrice des observations, ou bien en partitionnant les indices des colonnes, comme on peut le voir sur la table 1.1.

$$\left[\begin{array}{c} \mathbf{X}_{G_1} \\ \hline \mathbf{X}_{G_2} \\ \hline \dots \\ \hline \mathbf{X}_{G_K} \end{array}\right] \quad \text{ou} \quad \left[\begin{array}{c|c|c|c} \mathbf{X}_{G_1} & \mathbf{X}_{G_2} & \dots & \mathbf{X}_{G_K} \end{array}\right]$$

TABLE 1.1 – partitionnement de points (gauche) ou de variables (droite)

On nommera **partitionnement de points** le problème du partitionnement appliqué à $\mathcal{X} = \{X_1, \dots, X_n\} \subset \mathbb{R}^p$. Ce problème de partitionnement cherche à distinguer des groupes parmi un ensemble d'observations, comme l'illustre la figure 1.1. On peut se faire une représentation visuelle ou géométrique du problème : il consiste à regrouper des points dans l'espace $\mathbb{R}^p$ qui seraient proches ou concentrés. Le modèle simplifié suivant, qui postule que les points d'un même groupe ont même moyenne, vient correctement illustrer cette vision :

Hypothèse 1.1 (Cluster model, Chrétien, Dombry et Faivre, 2016)**.** *Pour tout point X_a appartenant à un groupe $G_k \in \mathcal{G}$ on a :*

$$\mathbb{E}[X_a] = \mu_k \in \mathbb{R}^p. \tag{1.6}$$

Ce contexte sera le cadre d'étude du chapitre 3.

Remarque 1.2 (Connexion SBM)**.** *Si l'on traite du partitionnement d'un graphe, on peut chercher à grouper ses noeuds en appliquant le problème du partitionnement à $\mathcal{X} = \{X_1, \dots, X_n\}$ où $X_a \in \mathbb{R}^n$ encode alors l'indicatrice des connexions du noeud a aux autres noeuds du graphe ; la matrice $(X_a^{(i)})_{ai}$ est la matrice d'adjacence du graphe, et un modèle statistique populaire est le modèle à blocs stochastiques (SBM). Alors la structure d'inter-dépendance des arêtes apporte facilités et difficultés bien spécifiques.*

Le deuxième contexte statistique annoncé considère le groupement des coordonnées d'un même vecteur aléatoire X dont $X_1, \dots, X_n$ seraient les réalisations. On suppose alors que ces coordonnées sont des variables aléatoires centrées. On nommera **partitionnement de variables** le problème du partitionnement appliqué à $\mathcal{X} = \{X^{(1)}, \dots, X^{(p)}\} \subset \mathbb{R}^n$. Dans le cas du partitionnement de variables, on ne cherche plus à partitionner des entités indépendantes, au contraire celles-ci sont possiblement fortement corrélées entre elles. On utilise souvent la modélisation par variables latentes :

Hypothèse 1.2 (Modèle latent)**.** *On suppose qu'il existe $Z = (Z^{(1)}, \dots, Z^{(K)})$ (où $K := |\mathcal{G}|$) et $E = (E^{(1)}, \dots, E^{(p)})$ vecteurs aléatoires indépendants tels que pour tout point $X^{(i)}$ appartenant à un groupe G_k de $\mathcal{G}$, on a :*

$$X^{(i)} = Z^{(k)} + E^{(i)} \tag{1.7}$$

Ce contexte sera le cadre d'étude du chapitre 2.

2 Estimateurs pour le partitionnement

2.1 Les estimateurs à vraisemblance et K-moyennes

Dans le cadre du problème du partitionnement, les méthodes reposant sur un modèle statistique font usage de la vraisemblance des données, et cherchent soit à estimer les distributions modélisées, soit à établir une partition sous-jacente. Les modèles de mélange Pearson, 1894 sont employés à cet effet depuis longtemps (voir par exemple Mclachlan et Basford, 1988). Conceptuellement la méthode utilisée consiste à estimer les paramètres du mélange en cherchant à maximiser la vraisemblance des données observées avec l'algorithme EM, Dempster, Laird et Rubin, 1977, puis d'en déduire la partition la plus probable par le maximum a posteriori de la vraisemblance. Ces méthodes produisent des résultats pour l'estimation (Dasgupta et Schulman, 2007 apprend le mélange si $\Delta^2 \gtrsim \sigma^2 p$), mais sont confrontées à un obstacle sérieux : le problème d'optimisation considéré n'étant pas convexe, les solutions proposées trouvent leur convergence dans des optima locaux et ont une grande sensibilité à l'initialisation.

Intéressons-nous à un cas particulier de plus près. Lorsque l'on est en présence d'observations gaussiennes isotropiques et homoscedastiques, le maximum de vraisemblance est réalisé par un estimateur bien connu : celui qui cherche à minimiser sur l'ensemble des partitions à K éléments la variance intra-groupes, soit l'estimateur des K-moyennes ou K-means.

Définition 2.1 (Estimateur des K-moyennes)**.**

$$\underset{\mathcal{G}=\{G_k\}_{1\leq k\leq K}}{\operatorname{argmin}} \sum_{k\leqslant K} \sum_{a\in G_k} \left\| x_a - \frac{1}{|G_k|} \sum_{b\in G_k} x_b \right\|^2 \tag{2.1}$$

où $\mathcal{X} = \{x_1, \dots, x_N\}$ est l'ensemble qu'on cherche à partitionner.

Le programme (2.1) est NP-dur, voir Aloise et al., 2009. Pour approcher cet objectif, Lloyd propose un algorithme en 1957 Lloyd, 1982 qui part d'une estimation initiale des centres, puis alterne à la manière d'un algorithme EM (l'algorithme de Lloyd est en fait un CEM, voir Celeux et Govaert, 1992) entre une phase d'estimation des groupes et une phase de ré-ajustement des centres. Si la convergence de l'algorithme est certaine, elle peut se révéler très lente et à défaut d'une bonne initialisation ne permettra d'atteindre qu'un minimum local de l'objectif présenté. Néanmoins cet estimateur doit nous intéresser : sa formulation est générique donc interprétable en dehors du contexte de la modélisation qui précède, elle permet de donner une réponse au problème du partitionnement, réponse qu'on sait optimale dans un cas particulier.

Remarque 2.1 (A propos de la kernelisation)**.** *Dans ces travaux on traite avec des données relationnelles sous leur forme la plus standard, à savoir la matrice de Gram du produit scalaire canonique (voir ci-dessous). Ca signifie que pour mesure de similarité entre entités on choisira toujours le produit scalaire associé à la norme euclidienne de l'espace considéré, comme l'illustre l'estimateur* (2.1)*. Ils s'étendent au cas où l'on considère un noyau Φ en traitant $x'_a := \Phi(x_a)$, mais ce manuscrit n'aborde pas plus ces aspects.*

2.2 Optimisation semi-définie positive

Pour des matrices, $A \succcurlyeq B$ signifie que $A - B$ est symétrique, semi-définie positive.

L'objectif (2.1) a un intérêt central dans ces travaux. Dans cette partie, nous expliquons comment on peut le connecter à des méthodes d'optimisation semi-définie positive (SDP) qui répondent au problème du partitionnement. Avec des opérations simples on peut d'abord reformuler cet objectif de la façon suivante :

$$\sum_{k \leqslant K} \sum_{a \in G_k} \left\| x_a - \frac{1}{|G_k|} \sum_{b \in G_k} x_b \right\|^2 = \frac{1}{2} \sum_{k \leqslant K} \frac{1}{|G_k|} \sum_{a,b \in G_k^2} \left\| x_a - x_b \right\|^2 \tag{2.2}$$

$$= - \sum_{k \leqslant K} \sum_{a,b \in G_k^2} \frac{1}{|G_k|} \langle x_a, x_b \rangle + \sum_{a \leqslant n} \left\| x_a \right\|^2. \tag{2.3}$$

Dans la formulation (2.3), plusieurs choses apparaissent. D'abord le terme de droite ne dépend pas de $\mathcal{G}$, on peut l'oublier dans le contexte de minimisation. Le terme de gauche peut être vu comme un produit scalaire matriciel entre la matrice de Gram $\widehat{\Lambda} := (\langle x_a, x_b \rangle)_{a,b}$ et une matrice B qui encode exactement le partitionnement $\mathcal{G}$. En effet soit $1_{G_k} = (\mathbf{1}\{a \in G_k\})_{a=1\dots n}$, on a :

$$(2.1) \Leftrightarrow \underset{\mathcal{G}=\{G_k\}_{1 \leq k \leq K}}{\operatorname{argmin}} \langle -\widehat{\Lambda}, B_{\mathcal{G}} \rangle \text{ s.t. } B_{\mathcal{G}} := \sum_{k \leqslant K} \frac{1}{|G_k|} 1_{G_k} 1_{G_k}^T \in \mathbb{R}^{n \times n}. \tag{2.4}$$

On a utilisé une caractérisation du partitionnement $\mathcal{G}$ par la matrice $B_{\mathcal{G}} := \sum_{G \in \mathcal{G}} \frac{1}{|G|} 1_G 1_G^T$ qu'on nomme *matrice caractéristique*, et le problème du partitionnement est alors vu comme un problème d'estimation de cette matrice caractéristique, une matrice positive, idempotente, dont les lignes et colonnes somment à 1, et de trace K. Ce jeu d'écriture mène à l'équivalence suivante, facilement

démontrable (voir Peng et Wei, 2007) :

$$(2.1) \Leftrightarrow \underset{B \text{ sym.}}{\operatorname{argmin}} \langle -\widehat{\Lambda}, B \rangle \text{ s.t. } \begin{array}{l} \bullet\ B \geqslant 0, B.1 = 1 \\ \bullet\ \operatorname{tr}(B) = K \\ \bullet\ B^2 = B \end{array} . \tag{2.5}$$

Cette nouvelle formulation illustre bien les difficultés de l'estimateur des K-moyennes : il est équivalent à un programme de minimisation sur l'ensemble des matrices idempotentes, qui n'est pas convexe. Mais en interprétant $B^2 = B$ comme contrainte opérant sur le spectre de B, on peut relâcher celle-ci par la condition suivante :

$$I \succcurlyeq B \succcurlyeq 0 \tag{2.6}$$

au sens de l'inégalité semi-définie positive. En remarquant que $I \succcurlyeq B$ est redondant avec les contraintes $B.1 = 1$ et $B \geqslant 0$, Peng et Wei, 2007 proposent la relaxation convexe suivante :

Définition 2.2 (K-MEANS SDP, Peng et Wei, 2007)**.**

$$\underset{B \textit{ sym.}}{\operatorname{argmin}} \langle -\widehat{\Lambda}, B \rangle \textit{ s.t. } \begin{array}{l} \bullet\ B \geqslant 0, B.1 = 1 \\ \bullet\ \operatorname{tr}(B) = K \\ \bullet\ B \succcurlyeq 0 \end{array} \tag{2.7}$$

Le programme proposé ci-dessus cherche à minimiser une forme linéaire sous des contraintes linéaires et une contrainte conique, il est donc convexe. On a approché un problème dur par un problème qu'on sait optimiser en temps polynomial !

L'ensemble des matrices considéré contient bien les matrices caractéristiques, mais par construction il est plus large que ça, et on est en droit de se demander comment passer d'une solution de (2.7) à une partition de l'ensemble $\mathcal{X}$. En fait la question ne se posera pas pour notre analyse théorique : comme on traite le problème du partitionnement exact, nous chercherons précisément à décrire un régime où la solution de (2.7) est admissible pour les K-moyennes, i.e. appartient à l'ensemble des matrices considéré dans (2.5), les matrices caractéristique de partitionnements.

Remarque 2.2 (De la relaxation SDP au partitionnement)**.** *Empiriquement on a besoin de cette étape supplémentaire comme la résolution numérique de* (2.7) *n'est jamais exacte. Les auteurs de cette relaxation utilisent une procédure d'arrondissement correspondant à une projection spectrale en dimension* $\mathbb{R}^{K-1}$*. On pourrait aussi appliquer un algorithme de Lloyd, ou de partitionnement hiérarchique. Cependant on constate en pratique que cette étape est simple, au sens où à partir du minimiseur, ces algorithmes produisent facilement la même partition.*

Cette reformulation s'inscrit dans un contexte bien plus large de découvertes à partir de relaxations SDP du maximum de vraisemblance. Pour ne citer que les plus récentes, Chrétien, Dombry et Faivre, 2016, s'inspirant des travaux de Guédon et Vershynin, 2016 dans la communauté SBM, mais aussi Montanari et Sen, 2016 ; Chen et Xu, 2016 ; Amini et Levina, 2018 ; Cai et Li, 2015 ; Abbe, Bandeira et Hall, 2016 ; Perry et Wein, 2017 ; Li et al., 2017. Tous ces travaux optimisent à l'aide de formulations SDP des approximations plus ou moins fidèles de la vraisemblance ou de la coupe minimum, sans nécessiter de procédure d'initialisation particulière. Elles semblent par ailleurs exhiber des propriétés de robustesse, aux points extrêmes, perturbations et mis-spécifications,

comme discutées dans Moitra, Perry et Wein, 2016 ; Cai et Li, 2015 ; Javanmard, Montanari et Ricci-Tersenghi, 2016 ou par la suite dans ce manuscrit. Enfin elles sont particulièrement intéressantes pour les possibilités d'analyses qu'elles présentent, voir section 3.2 page 12.

Une force spécifique du K-means SDP (2.7), qui optimise sur une relaxation convexe de l'ensemble des matrices caractéristiques de partitions $B_{\mathcal{G}} := \sum_{G\in\mathcal{G}} \frac{1}{|G|} 1_G 1_G^T$, est qu'il ne demande pas de connaissances a priori sur $\mathcal{G}$, à l'exception du nombre de structures K. Cette propriété est liée à la renormalisation $\frac{1}{|G|}$ dans $B_{\mathcal{G}}$, qui n'est pas présente dans les programmes SDP usuellement étudiés qui chercheraient plutôt à estimer la matrice d'adjacence de la partition $\sum_{G\in\mathcal{G}} 1_G 1_G^T$. Par exemple dans le cas particulier où la partition recherchée $\mathcal{G}$ a des groupes de même taille m, la relaxation par Amini et Levina, 2018 dans le cadre des modèles à blocs stochastiques correspond à :

$$\underset{B \text{ sym.}}{\operatorname{argmin}} \langle -\mathbf{X}, B\rangle \text{ s.t. } \begin{array}{l} \bullet\ B.1 = m1 \\ \bullet\ \operatorname{diag}(B) = 1 \\ \bullet\ B \succcurlyeq 0 \end{array} . \tag{2.8}$$

Ce programme exige de connaître les tailles de groupes a priori. Néanmoins ce SDP a tout son intérêt, Li et al., 2017 montrent qu'il permet de répondre au moins partiellement à la question de la localisation de la condition de séparation évoquée en problématique.

Enfin on présente une relaxation plus forte de l'objectif (2.1) qui permet de faire un lien avec un autre estimateur classique, l'estimateur spectral de McSherry, 2001. Celui-ci calcule les K premiers vecteurs propres de la matrice de Gram, puis applique un algorithme de partitionnement sur la matrice des vecteurs propres obtenue. On montre (voir section 5.4 page 41) que cette procédure revient à appliquer ce même algorithme de partitionnement au produit de la relaxation de (2.5) suivante :

Définition 2.3 (Spectral SDP, Peng et Wei, 2007)**.**

$$\underset{B \textit{ sym.}}{\operatorname{argmin}} \langle -\widehat{\Lambda}, B\rangle \textit{ s.t. } \begin{array}{l} \bullet\ \operatorname{tr}(B) = K \\ \bullet\ 1 \succcurlyeq B \succcurlyeq 0 \end{array} . \tag{2.9}$$

Nombreux sont les estimateurs du partitionnement qui partagent des liens de parenté avec les K-moyennes, et qu'on peut mieux comprendre grâce à ce passage dans le monde de l'optimisation. Il est intéressant de noter qu'il existe des travaux antérieurs qui montrent hors de ce contexte que les K-moyennes se relâchent par des estimateurs spectraux, voir par exemple Zha et al., 2001.

3 Contributions

Au chapitre 4 de ce manuscrit on présente de façon plus détaillée des rapports étroits entre le partitionnement et le domaine de l'optimisation. La question du partitionnement exact à vitesse optimale a été traitée dans deux articles, Bunea et al., 2018a et Royer, 2017 correspondant respectivement aux chapitres 2 et 3 de ce manuscrit, le premier traitant du partitionnement de variable, le second du partitionnement de point. Enfin le chapitre 5 présente des travaux de partitionnement appliqué à la biologie.

3.1 PECOK : penalisation, convexification des K-means

On a posé précédemment le problème du partitionnement, et plus précisément la question de savoir s'il est possible de faire du partitionnement exact à une vitesse optimale au sens minimax, dans les cadres du partitionnement de points et de variables. On a montré que l'optimisation SDP était une voie prometteuse, qu'on a souhaité attaquer à travers la brèche des K-moyennes – estimateur qu'on sait optimal dans un cas précis. Peut-on généraliser ce cas d'optimalité ?

Le biais des K-moyennes

En considérant l'estimateur des K-moyennes dans le cas d'un modèle simplifié précédent (1.6), en espérance on obtient :

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{2}\sum_{k\leqslant K}\frac{1}{|G_k|}\sum_{a,b\in G_k^2}\left\|X_a - X_b\right\|^2\right]\\ =&\frac{1}{2}\sum_{k\leqslant K}\frac{1}{|G_k|}\sum_{a,b\in G_k^2}\left\|\mathbb{E}[X_a] - \mathbb{E}[X_b]\right\|^2 + \sum_{a=1}^{n}\Gamma_a - \sum_{k\leqslant K}\frac{1}{|G_k|}\sum_{a\in G_k}\Gamma_a\end{aligned} \tag{3.1}$$

où

$$\forall a \in 1\dots n, \quad \Gamma_a := \operatorname{tr}\operatorname{Cov}(X_a). \tag{3.2}$$

Pour l'analyse on pourra ignorer le terme du milieu qui est invariant selon la partition considérée. Le terme de gauche représente le critère des K-moyennes appliqué aux espérances des points : sa minimisation en une partition rassemblant les points "proches" (au sens des moyennes) permettra donc bien de retrouver la partition minimisant la variance. Mais on voit apparaître à droite un terme de biais qui opère de la façon suivante : si un groupe G_k est composé d'éléments ayant une dispersion sensiblement plus large que comparé aux éléments des autres groupes, alors ce critère aura tendance à le pénaliser sensiblement plus. Ainsi la minimisation du critère pourra conduire à le séparer en plus petites parties, indépendamment du comportement des moyennes des points. Autrement dit cet exemple montre que le critère des K-moyennes peut systématiquement se tromper dans le partitionnement d'un problème simple et bien posé, par faute de ce biais. Ce critère, optimal dans le cas d'observations homoscédastiques, est donc en général biaisé.

Remarque 3.1 (Un biais volumique)**.** *Il faut préciser ici que ce biais provient de la dispersion entre les* $\operatorname{tr}\operatorname{Cov}(X_a)$*, quantités réelles qui mesurent uniquement l'étalement des variables aléatoires considérées. Autrement dit le biais des K-moyennes ne provient pas de l'anisotropie des distributions, mais de leur caractère anisovolumique, cf Figure 1.2*

Estimation des volumes Γ

On souhaite à présent corriger ce biais, à l'aide d'une estimation de Γ. En évaluant à nouveau (3.2) et (1.6), on peut remarquer que si b appartient au même groupe que a, on peut écrire

$$\Gamma_a = \mathbb{E}\langle X_a, X_a\rangle - \langle \mu_k, \mu_k\rangle = \mathbb{E}\langle X_a, X_a - X_b\rangle. \tag{3.3}$$

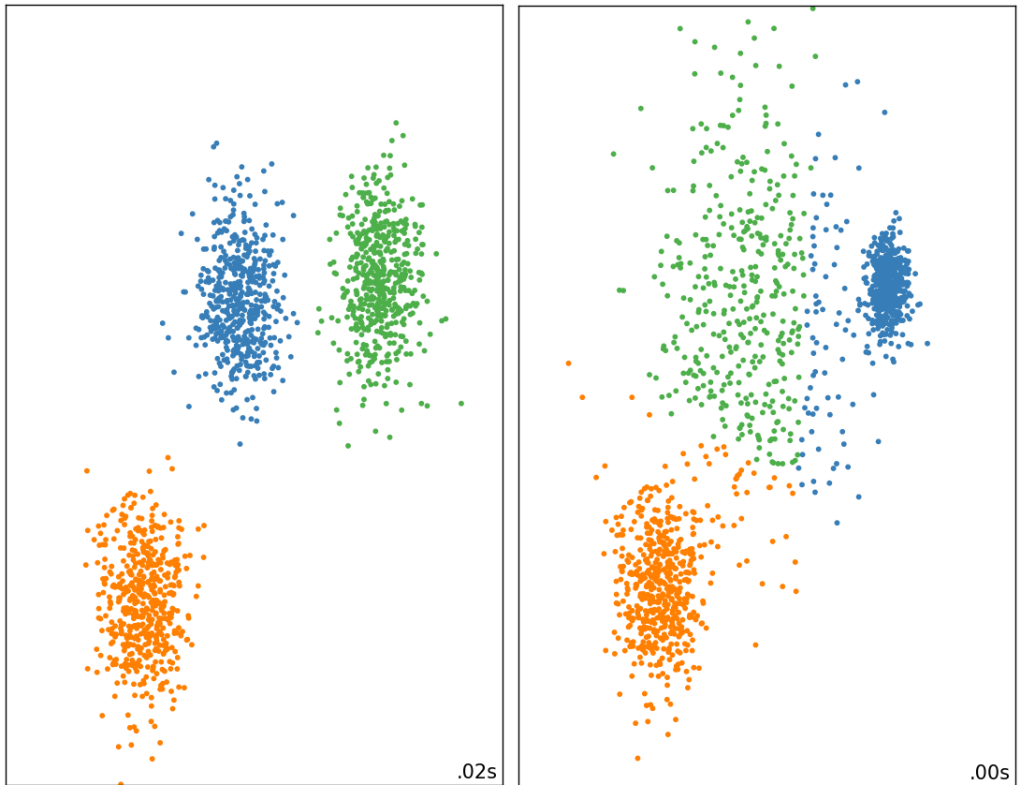


FIGURE 1.2 – Le biais des K-moyennes dans $\mathbb{R}^2$. A gauche les réalisations de trois gaussiennes isovolumiques ovoïdales, sur lesquelles on applique un algorithme de Lloyd. A droite on a introduit une dispersion entre les variances par groupe, et ce même algorithme alors échoue.

Intuitivement on voudrait donc trouver pour chaque indice a un "voisin" $\widehat{b}(a)$, et construire une correction du type :

$$\widehat{\Gamma}_a := \langle X_a, X_a - X_{\widehat{b}(a)} \rangle. \tag{3.4}$$

Voici un exemple de contrôle qu'on peut produire avec ce genre d'idée :

Lemme 3.1 (Un premier estimateur de Γ). *Soit $\widehat{\Gamma}^{(0)}$ l'estimateur défini par, pour tout $a \in 1 \dots n$:*

$$\widehat{\Gamma}^{(0)}_a := \langle X_a, X_a - X_{\widehat{b}_0(a)} \rangle \tag{3.5}$$

où

$$\widehat{b}_0(a) := \operatorname*{argmin}_{b \in [n], b \neq a} \max_{d \in [n], d \neq a, b} |\langle X_a - X_b, X_d \rangle|. \tag{3.6}$$

Si on suppose qu'on est dans le modèle simplifié (1.6)*, que les observations sont sous-gaussiennes de*

covariances dominées par $\sigma^2 I_p$, que $\min_{G\in\mathcal{G}} |G| > 1$, alors on a avec probabilité plus grande que $1 - 1/n$:

$$|\widehat{\Gamma}^{(0)} - \Gamma|_\infty \lesssim \sigma^2 \Big(\log n + (\sqrt{p} + \max_{l\in[K]} |\mu_l|_2/\sigma)\sqrt{\log n} \Big) \tag{3.7}$$

Il faut noter que $\widehat{b}_0(a)$ n'est pas nécessairement un voisin de a au sens de la partition $\mathcal{G}$!

Démonstration simplifiée. Soit $a \in G_k$ et $c \in G_k \setminus \{a\}$, on a par pivot :

$$\langle X_a - X_{\widehat{b}_0(a)}, X_a \rangle = \langle X_a - X_c, X_a \rangle + \langle X_c, X_a - X_{\widehat{b}_0(a)} \rangle - \langle X_{\widehat{b}_0(a)}, X_a - X_c \rangle. \tag{3.8}$$

Dès lors on peut majorer comme ceci :

$$|\langle X_a - X_{\widehat{b}_0(a)}, X_a \rangle - \Gamma_a| \leqslant |\langle X_a - X_c, X_a \rangle - \Gamma_a| + 2 \max_{d\in[n], d\neq a,c} |\langle X_a - X_c, X_d \rangle|, \tag{3.9}$$

par construction de $\widehat{b}_0(a)$. En notant $\forall d \in 1 \dots n, E_d := X_d - \mathbb{E}[X_d]$ on obtient :

$$|\widehat{\Gamma}^{(0)}_a - \Gamma_a| \leqslant ||E_a|^2 - \mathbb{E}\,|E_a|^2| + 5 \max_{b,d\in[n]^2, b\neq d} |\langle E_b, E_d \rangle| + 6 \max_{b,l\in[n]\times[K]} |\langle E_b, \mu_l \rangle|. \tag{3.10}$$

Les trois quantités majorantes sont des valeurs absolues de formes gaussiennes quadratiques ou linéaires. On se réfèrera aux résultats du chapitre 3 pour les outils dérivés de l'inégalité de Hanson-Wright qui permettent de majorer (à chaque fois avec probabilité plus grande que $1 - 1/(3n)$) ces deux premières quantités par $\sigma^2(\log n + \sqrt{p \log n})$ (voir par exemple les contrôles (B.32) page 106 et (B.34) page 106). La troisième quantité est majorée par $\sigma \max_{l\in[K]} |\mu_l|_2 \sqrt{\log n}$ (voir par exemple (B.25) page 104). Le résultat final est obtenu par unions de bornes. □

Le contrôle uniforme (3.6) pour le choix de $\widehat{b}_0(a)$ est motivé par l'inégalité (3.9). Si le terme $\max_{l\in[K]} |\mu_l|_2$ nous gène pour l'analyse, c'est parce que l'estimateur est imparfait : on est en droit d'espérer obtenir l'indépendance à l'origine (donc voir apparaître plutôt $\max_{k,l\in[K]^2} |\mu_k - \mu_l|_2$) et même l'absence de ce terme contenant du signal. Ces deux souhaits se réalisent grâce à l'estimateur $\widehat{\Gamma}$ (voir définition (5.8) page 39 et (3.10) page 95) qui permet, pour comparaison de produire un contrôle de l'ordre de $\sigma^2\sqrt{p \log n}$ quand $\log n \lesssim p$, tandis qu'en l'absence de correction on aurait eu :

$$|\Gamma|_\infty = p \times \sigma^2. \tag{3.11}$$

Dès lors on pourra observer les bénéfices en terme de vitesse de ce type de correction dès que $p \gtrsim \log n$. Empiriquement on en montre un effet Figure 1.3.

Où l'on introduit deux nouveaux estimateurs

En combinant les relaxations de l'estimateur des K-moyennes et une technique de correction comme ci-dessus, on produit et propose deux nouveaux estimateurs génériques pour le partitionnement, réalisant les programmes suivants :

Définition 3.1 (PECOK : Penalized, convex K-means)**.**

$$\underset{B \text{ sym.}}{\operatorname{argmin}} \langle \widehat{\Gamma} - \widehat{\Lambda}, B \rangle \text{ s.t. } \begin{array}{l} \bullet\ \operatorname{tr}(B) = K, B.1 = 1 \\ \bullet\ B \geqslant 0 \\ \bullet\ B \succcurlyeq 0 \end{array} \tag{3.12}$$

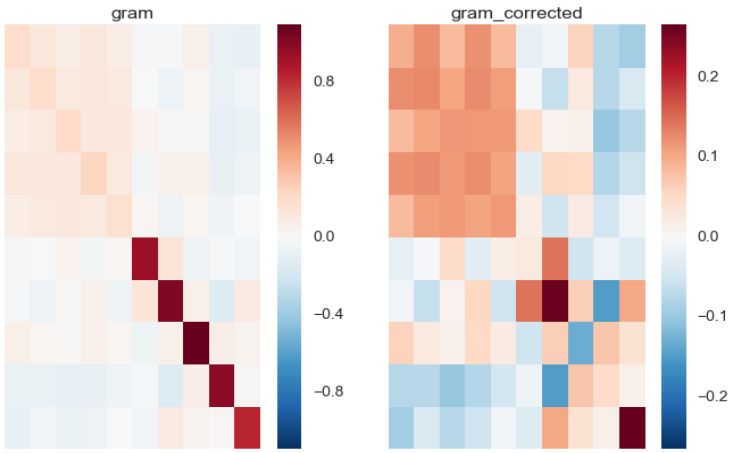


FIGURE 1.3 – Matrice de Gram pour deux groupes de 5 variables hétéroscédastiques, avec correction (à droite) et sans (à gauche) dans $\mathbb{R}^{100}$. La matrice de gauche souffrira du biais des K-moyennes pour le partitionnement, i.e. elle ne sera probablement pas bien partitionnée.

Définition 3.2 (CSC : Corrected spectral clustering)**.**

$$\underset{B \text{ sym.}}{\operatorname{argmin}} \langle \widehat{\Gamma} - \widehat{\Lambda}, B \rangle \; s.t. \; \begin{array}{l} \bullet \; \operatorname{tr}(B) = K \\ \bullet \; 1 \succcurlyeq B \succcurlyeq 0 \end{array} \tag{3.13}$$

On montre en Figure 1.4 l'importance de la correction dans les performances des deux algorithmes présentés. Ces deux estimateurs sont pareillement dérivés dans le contexte du partitionnement de variables.

3.2 Résultats théoriques

On présente à présent les comportements théoriques des estimateurs qu'on vient d'introduire, à travers des nouveaux modèles statistiques élaborés pour répondre au problème. Dans les deux cadres statistiques étudiés, on a introduit des modèles identifiables, génériques et robustes aux mis-spécifications. Ils postulent l'existence d'une structure par blocs dont l'analyse à travers les données permet de retrouver la partition $\mathcal{G}$. Ces modèles englobent ou sont à mettre en correspondance avec des modèles classiques comme les modèles à variables latentes, le modèle d'Ising par blocs (Berthet, Rigollet et Srivastava, 2018), le cluster model (Chrétien, Dombry et Faivre, 2016) ou les modèles de mélanges présentés plus haut.

Un modèle pour le partitionnement de variables

Pour construire un modèle de partitionnement, on propose l'approche suivante : les entités d'un même groupe devraient se ressembler dans leurs rapports intra et inter-groupes, de façon à être

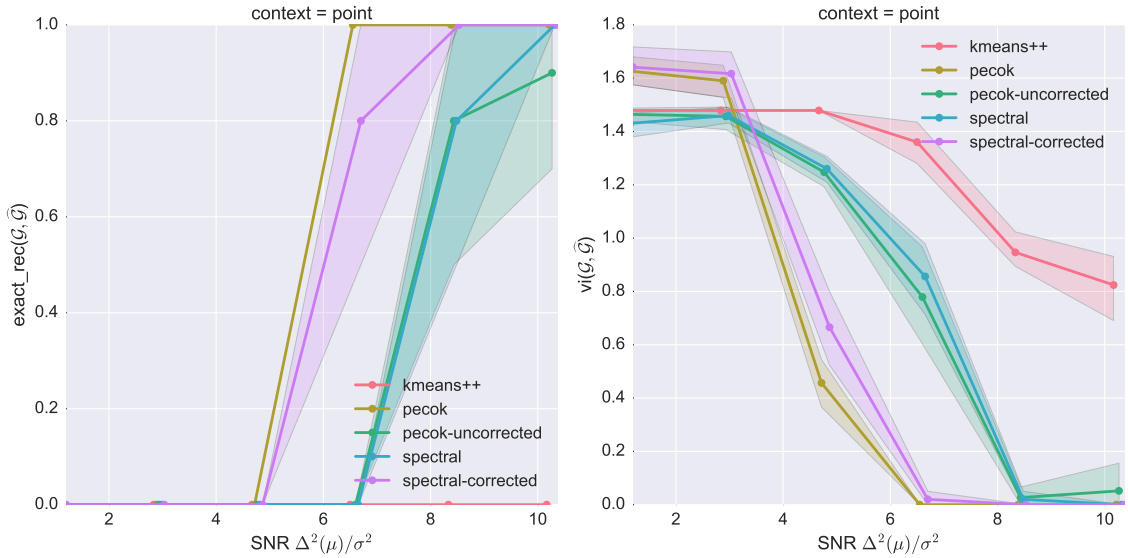


FIGURE 1.4 – Importance de la correction pour les algorithmes des K-moyennes relâchées et spectral, sur des données simulées : problème du partitionnement de 80 points en 4 groupes dans $\mathbb{R}^{500}$. En abscisse le ratio signal-bruit est augmenté, pour comparer les performances relatives des estimateurs. En ordonnées, à gauche le % de partitions exactement retrouvées, à droite une distance variationnelle entre vraie partition et estimée

interchangeables, de telle sorte que la matrice de covariance empirique $\widehat{\Sigma}$ devrait avoir une certaine forme d'invariance à une permutation qui préserverait les groupes d'une partition $\mathcal{G}$. Dès lors, l'espérance de cette matrice doit exhiber une structure de blocs à une permutation près des lignes et des colonnes. Par ailleurs, on est instruit par les modèles plus simples présentés plus haut (1.6) et (1.7) : dans les deux cas, la matrice relationnelle moyenne se compose alors d'une structure de groupe, comme attendu, mais aussi d'une structure de bruit diagonale. Par exemple partant de (1.7), on a en effet :

$$\mathrm{Cov}(X^{(i)}, X^{(j)}) = \mathrm{Cov}(Z^{(k)}, Z^{(l)}) + \mathbf{1}_{i=j}\,\mathrm{Var}(E^{(i)}) \tag{3.14}$$

où $Z^{(k)}, Z^{(l)}$ sont les variables latentes respectivement associées à $X^{(i)}, X^{(j)}$. Aussi pour $\mathcal{G} = \{G_1, \ldots, G_K\}$, on définit la matrice d'appartenance $A := (1_{a \in G_k})_{a,k} = \left[\begin{array}{c|c|c} 1_{G_1} & \ldots & 1_{G_K} \end{array}\right] \in \mathbb{R}^{p \times K}$ et on utilise le modèle suivant de partitionnement de variables :

Hypothèse 3.1 (Modèle G-block, introduit dans Bunea, Giraud et Luo, 2015)**.** *On suppose que la matrice de covariance de $\mathcal{X}$ se décompose comme :*

$$\Sigma = ACA^T + \Gamma \tag{3.15}$$

où $C \in \mathbb{R}^{K \times K}$ est une matrice symétrique et $\Gamma \in \mathbb{R}^{p \times p}$ une matrice diagonale. On demande aussi $\Delta(C) := \min_{j<k}(C_{kk} + C_{jj} - 2C_{kj}) > 0$.

La métrique pour le partitionnement $\Delta(C)$ peut être ré-écrite, dans le cas du modèle latent (1.7), de cette façon :

$$\Delta(C) = \min_{j<k} \mathbb{E}[(Z_j - Z_k)^2], \tag{3.16}$$

et s'interprète donc comme le plus petit écart quadratique moyen entre les variables latentes des différents groupes. Aussi $\Delta(C)$ quantifie plus généralement la séparation entre les groupes portée par la covariance, et est à mettre en correspondance avec la quantité $\min_{j<k} |\mu_j - \mu_k|^2$ dans le cas du cluster model pour le partitionnement de points. Pour ce qui est de Γ, on a vu qu'elle représente une perturbation de la matrice de covariance de source stochastique, et on peut définir $\sigma^2 := |\Gamma|_\infty$ pour mesure de son importance. Enfin de ce modèle G-block on peut identifier une unique partition minimale (au sens de la sous-division) donnée par les classes d'équivalences de la relation

$$a \sim b \text{ si et seulement si } \max_{c \neq a,b} |\Sigma_{ac} - \Sigma_{bc}| = 0, \tag{3.17}$$

et cette partition est $\mathcal{G}$.

Résultats partitionnement variables

Pour le partitionnement de variables, le lemme de Fano permet de montrer qu'il est impossible de retrouver la partition sous-jacente lorsque :

$$\Delta(C) \lesssim \sqrt{\frac{\log p}{nm}} + \frac{\log p}{n}, \tag{3.18}$$

(voir Théorème 3.2 page 33), où $m = \min_{G \in \mathcal{G}} |G|$ décrit la plus petite taille des groupes considérés. On observe que m aura donc de l'influence sur la capacité à discriminer de n'importe quel algorithme tant qu'il sera borné par $n/\log p$. Pour comparaison, la vitesse d'estimation en norme infinie pour la covariance empirique est de $\sqrt{\log p/n} + \log p/n$, ce qui illustre une différence de nature des deux problèmes.

L'estimateur PECOK a un comportement quasi-optimal vis-à-vis de la vitesse minimax (1.2). On a employé le mécanisme de preuve suivant : on étudie l'objectif primal directement qu'on découpe à l'aide de plusieurs inégalités de normes duales. En contrôlant les termes ainsi obtenus par des inégalités de déviations – certaines nouvelles – pour les formes linéaires et quadratiques gaussiennes, on arrive à montrer que la matrice B^* associée à la partition $\mathcal{G}$ recherchée est celle qui minimise le produit scalaire $\langle B, \widehat{\Gamma} - \widehat{\Sigma} \rangle$, dans des régimes non-asymptotiques qu'on présente par la suite.

On précise ici qu'on pourrait aussi obtenir des résultats en partitionnement approché, via une adaptation des techniques de Guédon et Vershynin, 2016 reposant sur l'inégalité de Grotendieck dans le modèle de SBM, à la manière de Mixon, Villar et Ward, 2016, Chrétien, Dombry et Faivre, 2016 ou en s'inspirant des travaux Fei et Chen, 2017 qui analysent des SDPs similaires.

De ces dérivations on a montré que l'estimateur retrouve la bonne partition avec grande probabilité dès lors que :

$$\Delta(C) \gtrsim \sigma^2 \left[\frac{K + \log p}{n} + \sqrt{\frac{K + \log p}{nm}} \right] \tag{3.19}$$

(voir Théorème 5.1, page 40). Ce résultat est donc optimal dès lors que K ne grandit pas plus vite que $\log p$. Dans le cas contraire on a de bonnes raisons de penser qu'il peut exister une écart entre la meilleure performance d'estimateurs polynomiaux et ce que la théorie statistique prévoit comme vitesse optimale pour le partitionnement exact (on renvoit à la discussion détaillée de Chen et Xu, 2016). Par ailleurs, ce théorème permet de retrouver la vitesse optimale de Berthet, Rigollet et Srivastava, 2018 pour le modèle d'Ising par blocs où $K = 2$ et les groupes considérés sont de même taille ; il en constitue donc une extension pour des valeurs de K plus générales et à des partitions de tailles quelconques.

Résultats partitionnement points

Les résultats pour le partitionnement de points ont des caractéristiques similaires, bien qu'ils s'appliquent à un problème fondamentalement différent. Les vitesses théoriques de séparation sont suggérées dans Lesieur et al., 2016 et confirmées dans Banks et al., 2018, on ne pourra détecter (c'est-à-dire être capable d'affirmer qu'il existe un mélange) des partitions si :

$$\min_{j \neq k} |\mu_j - \mu_k|^2 \lesssim \sqrt{\frac{pK \log K}{n}} \tag{3.20}$$

De plus on peut montrer facilement que le partitionnement exact requiert une condition de séparabilité moyenne, qui ne pourra être vérifiée si

$$\min_{j \neq k} |\mu_j - \mu_k|^2 \lesssim \sigma^2 \log n. \tag{3.21}$$

Ces deux limites sont à comparer avec les performances de l'estimateur PECOK dans le cadre du partitionnement de points. On emploie un mécanisme de preuve similaire à celui utilisé précédemment, pour établir un comportement quasi-minimax : l'estimateur trouve la partition sous-jacente dès lors que

$$\min_{j < k} |\mu_j - \mu_k|^2 \gtrsim \sigma^2 K \left[(1 + \frac{\log n}{K}) + \sqrt{\frac{p}{n}(1 + \frac{\log n}{K})} \right], \tag{3.22}$$

voir Théorème 3.1 page 96 – on s'est ici placé dans le cas où m grandit comme n/K pour faciliter la discussion. On voit ainsi que lorsque que $K \lesssim \log n$, les performances d'estimation de la partition sont optimale en faible dimension et ne diffèrent en grande dimension de la borne inférieure que d'un facteur $\log K / \log n$, facteur qui semble attendu dès lors qu'on passe du problème de détection au problème du partitionnement exact. Cette proximité montre qu'il n'y a pas de différence dans ces régimes entre performances optimales pour des partitionnements exact ou approchés. A nouveau lorsque $K \gtrsim \log n$, on constate un écart entre les performances de notre estimateur et le seuil de détection, qui pointe encore à l'existence d'un régime intermédiaire, comme discuté ci-dessus.

Que se passe-t-il si on veut se comparer aux estimateurs spectraux ? Le schéma de preuve utilisé pour PECOK ne fonctionne plus. Néanmoins on peut utiliser une technique développée par Lei et Rinaldo, 2015 pour les modèles à blocs stochastiques qui relie l'erreur de partitionnement à la quantité $|\widehat{\Lambda} - \widehat{\Gamma} - A\mu\mu^T A^T|_{op}$ où $\mu \in \mathbb{R}^{K \times p}$ est la matrice des centres des groupes. Le résultat obtenu est démontré en annexe, Proposition A.1 page 19 de ces travaux. Pour des groupes de tailles

égales, on obtient avec grande probabilité un contrôle pour tout $\rho < 1$ de la proportion d'erreur si :

$$\min_{j<k} |\mu_j - \mu_k|^2 \gtrsim \frac{\sigma^2 K}{\sqrt{\rho}} \Big(1 + \sqrt{\frac{p}{n}} + \max_{k\in[K]} |\mu_k|_2/\sigma\Big). \tag{3.23}$$

Le terme contenant du signal $\max_{k\in[K]} |\mu_k|_2$ est le plus problématique : il impose $\min_{j<k} |\mu_j - \mu_k|^2 \gtrsim \sigma^2 K^2/\rho$. On remarque que le résultat n'est pas optimal pour $\rho = 1/n$, qui assurerait un partitionnement exact. Précisons aussi qu'en l'absence de correction, on aurait en plus dans la parenthèse un terme en p/n, dominant en grandes dimensions.

Pour comparaison avec la littérature, avant la parution de Royer, 2017 le meilleur résultat venait de Li et al., 2017 qui à l'aide d'une condition de proximité et de certificats duaux, obtient un résultat de partitionnement exact si $\min_{j<k} |\mu_j - \mu_k|^2 \gtrsim \sigma^2(K + \log n)$ (avec une exigence sur la dimension : $n \gtrsim p^2 K^3 \log K$). Pour l'estimation de mélange gaussien, Mixon, Villar et Ward, 2016 qui utilise la même formulation SDP, exige une séparation de $\min_{j<k} |\mu_j - \mu_k|^2 \gtrsim \sigma^2 K^2$ (avec une condition sur la dimension : $n \gtrsim p$). Pour le partitionnement approché récemment Lu et Zhou, 2016 montre des vitesses de convergences exponentielles dans le cadre d'un mélange sous-gaussien à partir d'une initialisation spectrale, qui requièrent $\min_{j<k} |\mu_j - \mu_k|^2 \gtrsim \sigma^2 K^2(1 + Kp/n)$. Enfin dernièrement Giraud et Verzelen, 2018 donne un résultat de partitionnement approché pour notre formulation SDP à la condition que :

$$\min_{j<k} |\mu_j - \mu_k|^2 \gtrsim \sigma^2 K \Big(1 + \sqrt{\frac{p}{n}}\Big) \tag{3.24}$$

Par ailleurs on a montré (Théorème 4.1 page 97) qu'il est possible en optimisant sur un paramètre $\widehat{\kappa}$ de s'affranchir de la connaissance a priori de K, sans perdre le caractère optimal du résultat précédent. Par ailleurs, dans les deux résultats qu'on vient de citer, on établit une propriété supplémentaire d'adaptativité à la dimension effective de l'espace : c'est-à-dire que le p apparaissant dans (3.22) peut être remplacé par la quantité $r^* := \max_a \operatorname{tr}(\Sigma_a)/\max_a |\Sigma_a|_{op} \leqslant p$ (c'est d'ailleurs aussi le cas pour (3.23) et pour une quantité équivalente dans (3.24) de Giraud et Verzelen, 2018), quantité qui s'interprète grossièrement comme une dimension de l'espace dans lequel évoluent effectivement les observations. Ceci implique que les performances optimales de l'estimateur PECOK s'adaptent bien à la dimension effective du problème.

Extension à des modèles de partitionnement perturbés

Les modèles présentés jusqu'ici méritaient d'être relâchés, et on en présente des versions perturbées auxquelles tous les résultats présentés précédemment s'étendent naturellement.

Pour le partitionnement de variables on peut perturber le modèle G-block en admettant qu'il existe des coefficients non-nuls hors diagonale de la matrice de covariance, qui soient suffisamment petits, soit (voir 1.4 page 26) :

Hypothèse 3.2 (Modèle G-block avec perturbation)**.** *On suppose que la matrice de covariance se décompose de la façon suivante :*

$$\Sigma_{ab} = C_{jk} + \delta_{ab} D_a + (1 - \delta_{ab}) R_{ab} \tag{3.25}$$

où on demande $\Delta(C) > 8|R|_\infty$.

On a ainsi étendu la matrice Γ du modèle G-block en une matrice diagonale D plus une matrice de perturbation hors-diagonale R, une telle situation pouvant survenir par exemple dans le cas d'un modèle latent un peu perturbé où l'on aurait

$$X_a = (1 + \delta_a)Z_k + E_a \tag{3.26}$$

avec Z_k variable latente associée à X_a et $\delta_a = o(1)$. Cette décomposition apporte plus de réalisme et de souplesse à la modélisation, où les variables d'un même groupe n'ont plus qu'approximativement le même rapport aux autres. Il faut noter que l'identifiabilité de $\mathcal{G}$ exige alors des conditions supplémentaires pour être établie.

Pour le partitionnement de points, on propose le modèle de séparation sphérique suivant (voir Définition 2.1 page 93) :

Hypothèse 3.3 (Modèle (G, μ, δ)-clustered)**.** *Soit $\mathcal{G}$ partitionnant $\mathcal{X}$, $K = |\mathcal{G}|$, $\mu_1, \ldots, \mu_K \in \mathbb{R}^p$ et $\delta \geqslant 0$, on a $\forall a \in G_k$,*

$$|\mathbb{E}[X_a] - \mu_k|_2 \leqslant \delta \tag{3.27}$$

autrement dit chaque point a une moyenne dans la boule de centre μ_k et rayon δ, et on demande aussi que les moyennes soient sphériquement séparables soit que

$$\min_{k \neq l} |\mu_k - \mu_l|_2 > 4\delta. \tag{3.28}$$

Ce modèle étend largement le cluster model présenté plus haut, il est robuste aux mis-spécification : les observations d'un même groupe n'ont plus besoin d'être pareillement centrées, seulement suffisamment concentrées pour les besoins du partitionnement. Ce modèle sphérique est cohérent avec la mesure de similarité des observations, qui ne considère finalement les observations qu'à travers leurs produits scalaires.

3.3 Partitionnement : optimisation et applications

On a voulu compléter ces résultats théoriques pour PECOK par un ensemble de commentaires sur les qualités pratique de l'estimateur, qu'on présente au Chapitre 4.

Divers problèmes sont soulevés lorsque l'on veut résoudre le PECOK 3.12. D'abord, si le domaine de l'optimisation semi-définie positive a fait des découvertes importantes depuis l'introduction des méthodes de points intérieurs Nesterov et Nemirovskii, 1994, les programmes SDPs de grande dimension restent durs à résoudre en pratique : la difficulté d'optimiser selon la contrainte conique exige généralement des temps de calculs au moins cubiques. De plus parmi les programmes SDPs, PECOK est un programme très lourdement contraint du fait de la condition de positivité. Enfin le calcul du $\widehat{\Gamma}$ théorique utilisé aux Chapitres 2 et 3 est polynomial quartique, donc tout à fait prohibitif pour la grande dimension.

On a présenté des méthodes compétitives de résolution de SDPs, en particulier une approche de factorisation issue de Burer et Monteiro, 2003 traitant le problème à l'aide d'une reformulation non-linéaire, non-convexe qui semble pourtant avoir de bonnes propriétés de convergence. On a implémenté des algorithmes de partitionnement inspirés de nos analyses théoriques, dont une méthode modulaire du premier ordre Boyd et al., 2011, qui marie les méthodes d'ascension duale et des multiplicateurs, et permet d'évaluer le comportement empirique du SDP PECOK et de sa

reformulation non-convexe. Enfin on a introduit des substituts $\widehat{\Gamma}$ pour estimer les volumes Γ, qui s'inspirent de l'estimateur théorique ci-dessus mais ne demandent plus que des complexités polynomiales quadratiques ou cubiques. On produit des résultats de simulations qui illustrent forces et faiblesses des estimateurs ci-présentés. Les deux cadres du partitionnement de points et de variables montrent aussi des comportements bien spécifiques, et on verra empiriquement quels algorithmes sont les plus adaptés à ces deux contextes.

Enfin tandis que les travaux précédents abordent le problème du partitionnement au sens classique du terme, le chapitre 5 prend pour cadre l'étude le partitionnement mixte ("overlapping clustering") qui cherche à séparer $\mathcal{X}$ en parties distinctes plutôt que disjointes. Cette formulation est plus réaliste : elle admet que les parties de $\mathcal{X}$ considérées se chevauchent, elle est donc plus souple, par exemple au prix de difficultés supplémentaires pour l'identification du problème. Un article Bing, Das et Royer, 2018 synthétisant ces travaux est en cours de rédaction.

Ils prennent pour point de départ les travaux de Bing et al., 2017, dans le cadre du partitionnement de variables où les auteurs considèrent le problème du partitionnement mixte à partir d'un modèle à variables latentes (1.7) et une méthode pour résoudre le partitionnement mixte LOVE. On évalue cette méthode dans plusieurs contextes pour la biologie. Premièrement un problème d'indentification de gènes dans le cycle cellulaire humain et leur expressions dans les tumeurs, où la méthode proposée trouve une partition enrichie pour différent processus biologiques, et se compare favorablement à d'autres méthodes standard de partitionnement mixte. Deuxièmement la méthode est employée pour une partitionner une cohorte de contrôleurs HIV et progresseurs chroniques, et parvient à distinguer deux phénotypes cliniques distinct dans un cadre de partitionnement classique.

Appendice

A Contrôle de l'estimateur spectral corrigé

On montre ici comment contrôler les capacités d'estimation de CSC (3.13) dans le contexte du partitionnement de points. On rappelle qu'on peut travailler sur une formulation équivalente, c'est-à-dire qu'appliquer une η-approximation des K-moyennes sur (3.13) est équivalent à l'algorithme suivant :

1. Corriger la matrice de Gram : $\widetilde{\Lambda} := \widehat{\Lambda} - \widehat{\Gamma}$
2. Extraire les K premiers vecteurs propres de la matrice $\widetilde{\Lambda}$, assemblés en matrice $\widetilde{U}$
3. Appliquer une η-approximation des K-moyennes sur les lignes de $\widetilde{U}$

Soit $\widehat{\mathcal{G}} = \{\widehat{G}_1, \dots, \widehat{G}_K\}$ l'estimateur de partition produit, en notant S_k l'ensemble des points de G_k que l'algorithme ne classifie pas correctement, on peut définir et contrôler l'erreur de partitionnement :

$$l(\widehat{\mathcal{G}}, \mathcal{G}) := \sum_{k=1}^{K} |S_k|. \tag{A.1}$$

Enfin soient $m_k := |G_k|$, et $m := \min_{k\in[K]} m_k$.

Proposition A.1. *On se place dans le modèle simplifié* (1.6) *où les points sont groupés en moyenne selon* $\mathcal{G} = \{G_1, \dots, G_K\}$*, et en considérant des observations sous-gaussiennes de covariances dominées par* $\sigma^2 I_p$*. On suppose de plus que* $m > 2$*, et on note* $r_* := \max_{a\in[n]} \mathrm{tr}(\Sigma_a)/\max_{a\in[n]} |\Sigma_a|_{op}$ *(où* $\Sigma_a := \mathrm{Cov}(X_a)$*) la dimension effective du problème.*

Il existe c_η *constante ne dépendant que de* η *et* c *constante numérique telles que pour tout* $0 < \rho < 1$*, si*

$$\lambda_K(\mu\mu^T) \geqslant c_\eta \frac{\sigma^2}{\sqrt{\rho m}} \sqrt{K}\Big(\sqrt{n} + \sqrt{r_*} + \max_k \sqrt{m_k} \times |\mu|_{op}/\sigma\Big), \tag{A.2}$$

alors avec probabilité plus grande que $1 - c/n$ *on a* $l(\widehat{\mathcal{G}}, \mathcal{G}) \leqslant \rho n$*.*

Pour prouver ce résultat, on va d'abord établir un lien entre l'erreur totale $\sum_{k=1}^{K} |S_k|$ et la différence en norme de Frobénius entre les matrices de vecteurs propres de $\widetilde{\Lambda}$ et Λ (pour rappel, $\Lambda = A\mu\mu^T A^T$ contient le signal). Ensuite, par l'inégalité de Davis et Kahan, 1970 nous établissons un lien entre cette dernière quantité et la même différence en norme opérateur cette fois. Enfin nous contrôlons la norme opérateur à l'aide d'un résultat de Bunea, She et Wegkamp, 2011 et des contrôles développés dans ces travaux. Ce schéma de preuve est dû à Lei et Rinaldo, 2015 dans le contexte des modèles à blocs stochastiques, et est aussi utilisé au chapitre 2 pour le partitionnement de variables. Essentiellement, seuls les contrôles stochastiques diffèrent.

Démonstration. Soit $\mathcal{O}_K$ l'ensemble des matrices orthogonales carrées de dimension K.

On va utiliser le lemme suivant directement dérivé de Lei et Rinaldo, 2015 :

Lemma A.1. *Pour $\eta > 2$, $\widetilde{U}, U \in \mathbb{R}^{n\times K}$ où $U = AR$ avec A matrice d'appartenance associée à $\mathcal{G}$ et $R \in \mathbb{R}^{K\times K}$, soit $\widehat{\mathcal{G}}$ la partition obtenue d'une η-approximation des K-moyennes appliquée à $\widetilde{U}$. Si pour δ la distance minimale entre deux lignes de R, on a*

$$8\eta|\widetilde{U} - U|_F^2/\delta^2 \leqslant m, \tag{A.3}$$

alors il existe une permutation σ de $\{1, \dots, K\}$ telle que, en notant $S_k := G_k \setminus \widehat{G}_{\sigma(k)}$, on a :

$$\sum_{k=1}^{K} |S_k| \leqslant 8\eta|\widetilde{U} - U|_F^2/\delta^2. \tag{A.4}$$

Appliquons ce lemme au partitionnement spectral : on appelle $U_0 \in \mathbb{R}^{n\times K}$ la matrice des vecteurs propres de $A\mu\mu^T A^T$. Soit $\Delta := \mathrm{diag}(\sqrt{m_k})_k$, soit $Q' \in \mathcal{O}_K$ matrice des vecteurs propres de $\Delta\mu\mu^T\Delta^T$ et D matrice diagonale de ses valeurs propres. On a

$$A\mu\mu^T A^T = (A\Delta^{-1}Z)D(A\Delta^{-1}Z)^T, \tag{A.5}$$

donc il existe $Q \in \mathcal{O}_K$ telle que $U_0 = A\Delta^{-1}Q$. Soit $Q'' \in \mathcal{O}_K$ (à déterminer), on peut donc appliquer le lemme précédent à $\widetilde{U}$ et $U := U_0Q''$ et on remarque que $\delta = \min_{l\neq k}\sqrt{\frac{1}{m_k} + \frac{1}{m_l}} \leqslant \sqrt{2/m}$.

Pour contrôler l'erreur commise sur les vecteurs propres, on applique l'inégalité de Davis et Kahan, 1970 : il existe $Q'' \in \mathcal{O}_K$ telle que

$$|\widetilde{U} - UQ''|_F \leqslant \frac{\sqrt{8K}|\widetilde{\Lambda} - A\mu\mu^T A^T|_{op}}{m\lambda_K(\mu\mu^T)}. \tag{A.6}$$

Il s'agit enfin de contrôler la quantité

$$|\widetilde{\Lambda} - A\mu\mu^T A^T|_{op} \leqslant |\widehat{\Gamma} - \Gamma|_{op} + |EE^T - \Gamma|_{op} + 2|A\mu E^T|_{op}. \tag{A.7}$$

Dans ce qui suit, c_0, c_1, c_2, c_3 sont des constantes numériques. En utilisant les résultats du chapitre 3, la Proposition 3.3 page 95 et le Lemme A.1 page 99, on peut majorer avec probabilité supérieure à $1 - c_0/n$

$$|\widehat{\Gamma} - \Gamma|_{op} + |EE^T - \Gamma|_{op} \leqslant c_1\sigma^2(n + \sqrt{r_* n}). \tag{A.8}$$

Enfin pour la dernière quantité, on a :

$$|A\mu E^T|_{op} = |A\mu\mu^T(\mu\mu^T)^{-1}\mu E^T|_{op} \tag{A.9}$$

$$\leqslant |A\mu|_{op}|PE^T|_{op}, \tag{A.10}$$

où $P := \mu^T(\mu\mu^T)^{-1}\mu$ projecteur de rang K. On va utiliser le résultat suivant :

Proposition A.2. *Soit $W \in \mathbb{R}^{p\times n}$ matrice aux colonnes indépendantes, sous-gaussiennes de covariances dominées par $\sigma^2 I_p$, $Y \in \mathbb{R}^{p\times d}$ une matrice de rang K telle que $P = Y(Y^TY)^\dagger Y^T$ projecteur sur l'image de Y. Alors $\forall t > 0$ on a*

$$\mathbb{P}\left[|PW|_{op}^2 \geqslant 32\sigma^2\Big((n + K)\ln 5 + t\Big)\right] \leqslant 2\exp(-t). \tag{A.11}$$

Démonstration. Adaptation directe de la Proposition 15 de Bunea, She et Wegkamp, 2011. □

On applique cette proposition à la matrice $W = E^T$ pour obtenir qu'avec probabilité supérieure à $1 - c_2/n$, on a :

$$|PE|_{op} \leqslant c_3 \sigma \sqrt{n} \tag{A.12}$$

De plus $|A\mu|_{op} \leqslant \max_k \sqrt{m_k} \times |\mu|_{op}$.

On obtient une condition suffisante par union sur ces bornes, il existe c_η et c constante numérique telles que si :

$$m \geqslant c_\eta \frac{\sigma^2}{\sqrt{m}\lambda_K(\mu\mu^T)} \sqrt{K}\Big(n + \sqrt{nr_*} + \max_k \sqrt{m_k} \times |\mu|_{op}/\sigma\sqrt{n}\Big), \tag{A.13}$$

alors avec probabilité supérieure à $1 - c/n$:

$$\sum_{k=1}^{K} |S_k| \leqslant \Big[c_\eta \frac{\sigma^2}{\sqrt{m}\lambda_K(\mu\mu^T)} \sqrt{K}\Big(n + \sqrt{nr_*} + \max_k \sqrt{m_k}\sqrt{n} \times |\mu|_{op}/\sigma\Big)\Big]^2. \tag{A.14}$$

Ce qui permet de prouver la proposition pour $0 < \rho < 1$. □

Chapter 2

Model Assisted Variable Clustering: Minimax-Optimal Recovery and Algorithms

Contents

Ce chapitre présente des travaux en collaboration avec Florentina Bunea, Christophe Giraud, Xi Luo et Nicolas Verzelen, qui reposent sur les travaux antécédents de Bunea, Giraud, and Luo, 2015 et de Bunea et al., 2016. Ceux-ci ont été fusionnés, étendus et soumis pour publication dans Bunea et al., 2018a (et supplément Bunea et al., 2018b) à The Annals of Statistics. Les parties en rapport avec l'estimateur COD proviennent essentiellement de Bunea, Giraud, and Luo, 2015.

Abstract

The problem of variable clustering is that of estimating groups of similar components of a p-dimensional vector $X = (X_1, \ldots, X_p)$ from n independent copies of X. There exists a large number of algorithms that return data-dependent groups of variables, but their interpretation is limited to the algorithm that produced them. An alternative is model-based clustering, in which one begins by defining population level clusters relative to a model that embeds notions of similarity. Algorithms tailored to such models yield estimated clusters with a clear statistical interpretation. We take this view here and introduce the class of G-block covariance models as a background model for variable clustering. In such models, two variables in a cluster are deemed similar if they have similar associations will all other variables. This can arise, for instance, when groups of variables are noise corrupted versions of the same latent factor. We quantify the difficulty of clustering data generated from a G-block covariance model in terms of cluster proximity, measured with respect to two related, but different, cluster separation metrics. We derive minimax cluster separation thresholds, which are the metric values below which no algorithm can recover the model-defined clusters exactly, and show that they are different for the two metrics. We therefore develop two algorithms, COD and PECOK, tailored to G-block covariance models, and study their minimax-optimality with respect to each metric. Of independent interest is the fact that the analysis of the PECOK algorithm, which is based on a corrected convex relaxation of the popular K-means algorithm, provides the first statistical analysis of such algorithms for variable clustering. Additionally, we contrast our methods with another popular clustering method, spectral clustering, specialized to variable clustering, and show that ensuring exact cluster recovery via this method requires clusters to have a higher separation, relative to the minimax threshold. Extensive simulation studies, as well as our data analyses, confirm the applicability of our approach.

***Keywords*—** Variable clustering, latent models, block covariance matrix, minimax lower bound, consistent partition estimation, convex algorithms, SDP, K-means, high dimensional models

1 Introduction

The problem of variable clustering is that of grouping similar components of a p-dimensional vector $X = (X_1, \ldots, X_p)$. These groups are referred to as clusters. In this work we investigate the problem of cluster recovery from a sample of n independent copies of X. Variable clustering has had a long history in a variety of fields, with important examples stemming from gene expression data Zaag et al., 2015; Frey et al., 2014; Jiang, Tang, and Zhang, 2004 or protein profile data Bernardes et al., 2015. The solutions to this problem are typically algorithmic and entirely data based. They include applications of K-means, hierarchical clustering, spectral clustering, or versions of them. The statistical properties of these procedures have received a very limited amount of investigation. It is not currently known what probabilistic cluster models on X can be estimated by these popular techniques, or by their modifications. More generally, model-based variable clustering has received a limited amount of attention. One net advantage of model-based clustering is that population-level clusters are clearly defined, offering both interpretability of the clusters and a benchmark against which one can check the quality of a particular clustering algorithm.

In this work we propose the G-block covariance model as a flexible model for variable clustering and show that the clusters given by this model are uniquely defined. We then motivate and develop two algorithms tailored to the model, COD and PECOK, and analyze their respective performance in terms of exact cluster recovery, for minimally separated clusters, under appropriately defined cluster separation metrics.

1.1 The G-block covariance model

Our proposed model for variable clustering subsumes that the covariance matrix Σ of a centered random vector $X \in \mathbb{R}^p$ follows a block, or near-block, decomposition, with blocks corresponding to a partition $G = \{G_1, \ldots, G_K\}$ of $\{1, \ldots, p\}$. This structure of the covariance matrix has been observed to hold, empirically, in a number of very recent studies on the parcelation of the human brain, for instance Kong et al., 2018; Glasser et al., 2016; Craddock et al., 2012; Yeo et al., 2011. We further support these findings in Section 8, where we apply the clustering methods developed in this paper, tailored to G-block covariance models, for the clustering of brain regions.

To describe our model, we associate, to a partition G, a membership matrix $A \in \mathbb{R}^{p \times K}$ defined by $A_{ak} = 1$ if $a \in G_k$, and $A_{ak} = 0$ otherwise.

(A) The exact G-block covariance model. In view of the above discussion, clustering the variables $(X_1, \ldots, X_p)$ amounts to find a minimal (i.e. coarsest partition) G^*, such that two variables belong to the same cluster if they have the same covariance with all other variables. This implies that the covariance matrix Σ of X decomposes as

$$\Sigma = AC^*A^t + \Gamma, \tag{1.1}$$

where A is relative to G^*, C^* is a symmetric $K \times K$ matrix, and Γ a diagonal matrix. When a such a decomposition exists with the partition G^*, we say that $X \in \mathbb{R}^p$ follows an (exact) G^*-block covariance model.

(i) G-Latent Model. Such a structure arises, for instance, when components of X that belong to the same group can be decomposed into the sum between a common latent variable and an uncorrelated random fluctuation. Similarity within group is therefore given by association with

the same unobservable source. Specifically, the exact block-covariance model (1.1) holds, with a diagonal matrix Γ, when

$$X_a = Z_{k(a)} + E_a, \tag{1.2}$$

with $\mathrm{Cov}(Z_{k(a)}, E_a) = 0$, $\mathrm{Cov}(Z) = C^*$, and the individual fluctuations E_a are uncorrelated, and thus E has diagonal covariance matrix Γ. The index assignment function $k : \{1, \dots, p\} \to \{1, \dots, K\}$ is defined by $G_k = \{a : k(a) = k\}$. In practice, this model is used to justify the construction of a single variable that represents a cluster, the average of X_a, $a \in G_k$, viewed as an observable proxy of $Z_{k(a)}$. For example, a popular analysis approach for fMRI data, called region-of-interest (ROI) analysis Poldrack, 2007, requires averaging the observations from multiple voxels (a imaging unit for a small cubic volume of the brain) within each ROI (or cluster of voxels) to produce new variables, each representing a larger and interpretable brain area. These new variables are then used for downstream analyses. From this perspective, model (1.2) can be used in practice, see, for example Bellec et al., 2006, as a building block in a data analysis based on cluster representatives, which in turn requires accurate cluster estimation. Indeed, data-driven methods for clustering either voxels into regions or regions into functional systems, especially based on the covariance matrix of X, is becoming increasingly important, see for example Glasser et al., 2016; Yeo et al., 2011; Craddock et al., 2012; Power et al., 2011. Accurate data-driven clustering methods also enable studying the cluster differences across subjects Chong et al., 2017 or experimental conditions James, Hazaroglu, and Bush, 2016.

(ii) The Ising Block Model. The Ising Block Model has been proposed in Berthet, Rigollet, and Srivastava, 2018 for modelling social interactions, for instance political affinities. Under this model, the joint distribution of $X \in \{-1, 1\}^p$, a p-dimensional vector with binary entries, is given by

$$f(x) = \frac{1}{\kappa_{\alpha,\beta}} \exp\Big[\frac{\beta}{2p} \sum_{a \sim b} x_a x_b + \frac{\alpha}{2p} \sum_{a \nsim b} x_a x_b\Big], \tag{1.3}$$

where the quantity $\kappa_{\alpha,\beta}$ is a normalizing constant, and the notation $a \sim b$ means that the elements are in the same group of the partition. The variables X_a may for instance represent the votes of U.S. senators on a bill Banerjee, Ghaoui, and d'Aspremont, 2008. For parameters $\alpha > \beta$, the density (1.3) models the fact that senators belonging to the same political group tend to share the same vote. By symmetry of the density f, the covariance matrix Σ of X decomposes as an exact block covariance model $\Sigma = AC^*A^t + \Gamma$ where Γ is diagonal. When all groups G_k^* have identical size, we have $C^* = (\omega_{in} - \omega_{out})I_K + \omega_{out}J$ and $\Gamma = (1 - \omega_{in})I$, where the $K \times K$ matrix J has all entries equal to 1, and I_K denotes the $K \times K$ identity matrix, and the quantities $\omega_{in}, \omega_{out}$ depend on α, β, p.

(B) The approximate G-block model. In many situations, it is more appealing to group variables that *nearly* share the same covariance with all the other variables. In that situation, the covariance matrix Σ would decompose as

$$\Sigma = ACA^t + \Gamma, \text{ where } \Gamma \text{ has small off-diagonal entries.} \tag{1.4}$$

Such a situation can arise, for instance when $X_a = (1 + \delta_a)Z_{k(a)} + E_a$, with $\delta_a = o(1)$ and the individual fluctuations E_a are uncorrelated, $1 \leqslant a \leqslant p$.

1.2 Our contribution

We assume that the data consist in i.i.d. observations $X^{(1)}, \ldots, X^{(n)}$ of a random vector X with mean 0 and covariance matrix Σ. This work is devoted to the development of computationally feasible methods that yield estimates $\widehat{G}$ of G^*, such that $\widehat{G} = G^*$, with high probability, when the clusters are minimally separated, and to characterize the minimal value of the cluster separation from a minimax perspective. The separation between clusters is a key element in quantifying the difficulty of a clustering task as, intuitively, well separated clusters should be easier to identify. We consider two related, but different, separation metrics, that can be viewed as canonical whenever Σ satisfies (1.4). Although all our results allow, and are proved, for small departures from the diagonal structure of Γ in (1.1), our main contribution can be best seen when Γ is a diagonal matrix. We focus on this case below, for clarity of exposition. The case of Γ being a perturbation of a diagonal matrix is treated in Section 6.

When Γ is diagonal, our target partition G^* can be easily defined. It is the unique minimal (with respect to partition refinement) partition G^* for which there is a decomposition $\Sigma = AC^*A^t + \Gamma$, with A associated to G^*. We refer to Section 2 for details. We observe in particular, that $\max_{c \neq a,b} |\Sigma_{ac} - \Sigma_{bc}| > 0$ if and only if X_a and X_b belong to different clusters in G^*.

This last remark motivates our first metric MCOD based on the following COvariance Difference (COD) measure

$$\text{COD}(a,b) := \max_{c \neq a,b} |\Sigma_{ac} - \Sigma_{bc}| \quad \text{for any } \ a, b = 1, \ldots, p. \tag{1.5}$$

We use the notation $a \overset{G^*}{\sim} b$ whenever a and b belong to the same group G^*_k, for some k, in the partition G^*, and similarly $a \overset{G^*}{\nsim} b$ means that there does not exist any group G^*_k of the partition G^* that contains both a and b. We define the MCOD metric as

$$\text{MCOD}(\Sigma) := \min_{a \overset{G^*}{\nsim} b} \text{COD}(a,b). \tag{1.6}$$

The measure $\text{COD}(a,b)$ quantifies the similarity of the covariances that X_a and X_b have, respectively, with all other variables. From this perspective, the size of $\text{MCOD}(\Sigma)$ is a natural measure for the difficulty of clustering when analyzing clusters with components that are similar in this sense. Moreover, note that this metric is well defined even if C^* of model (1.1) is not semi-positive definite.

Another cluster separation metric appears naturally when we view model (1.1) as arising via model (1.2), or via small deviations from it. Then, clusters in (1.1) are driven by the latent factors, and intuitively they differ when the latent factors differ. Specifically, we define the "within-between group" covariance gap

$$\Delta(C^*) := \min_{j<k} \left(C^*_{kk} + C^*_{jj} - 2C^*_{jk} \right) = \min_{j<k} \mathbf{E}\left[(Z_j - Z_k)^2 \right], \tag{1.7}$$

where the second equality holds whenever (1.2) holds. In the latter case, the matrix C^*, which is the covariance matrix of the latent factors, is necessarily semi-positive definite. Further, we observe that $\Delta(C^*) = 0$ implies $Z_j = Z_k$ a.s. Conversely, we prove in Corollary 2.1 of Section 2 that if the decomposition (1.1) holds with $\Delta(C^*) > 0$, then the partition related to A is the partition G^* described above. An instance of $\Delta(C^*) > 0$ corresponds to having the within group covariances

stronger than those between groups. This suggests the usage of this metric $\Delta(C^*)$ for cluster analysis whenever, in addition to the general model formulation (1.1), we also expect clusters to have this property, which has been observed, empirically, to hold in applications. For instance, it is implicit in the methods developed by Craddock et al., 2012 for creating a human brain atlas by partitioning appropriate covariance matrices. We also present a neuroscience-based data example in 8.

Formally, the two metrics are connected via the following chain of inequalities, proved in Lemma B.1 page 66, and valid as soon as the size of the smallest cluster is larger than one, Γ and C^* is semi-positive definite (for the last inequality)

$$2\lambda_K(C^*) \leqslant \Delta(C^*) \leqslant 2\text{MCOD}(\Sigma) \leqslant 2\sqrt{\Delta(C^*)} \max_{k=1,\ldots,K} \sqrt{C^*_{kk}}. \tag{1.8}$$

The first inequality shows that conditions on $\Delta(C^*)$ are weaker than conditions on the minimal eigenvalue $\lambda_K(C^*)$ of C^*. In order to preserve the generality of our model, we do not necessarily assume that $\lambda_K(C^*) > 0$, as we show that, for model identifiability, it is enough to have the weaker condition $\Delta(C^*) > 0$, when the two quantities differ.

The second inequality in (1.8) shows that $\Delta(C^*)$ and $\text{MCOD}(\Sigma)$ can have the same order of magnitude, whereas the third inequality shows that they can also differ in order, and $\Delta(C^*)$ can be as small as $\text{MCOD}^2(\Sigma)$, for small values of these metrics, which is our main focus. This suggests that different statistical assessments, and possibly different algorithms, should be developed for estimators of clusters defined by (1.1), depending on the cluster separation metric. To substantiate this intuition, we first derive, for each metric, the rate below which no algorithm can recover exactly the clusters defined by (1.1). We call this the minimax optimal threshold for cluster separation, and prove that it is different for the two metrics. We call an algorithm that can be proved to recover exactly clusters with separation above the minimax threshold a minimax optimal algorithm.

Theorem 3.1 in Section 3 shows that no algorithm can estimate consistently clusters defined by (1.1) if

$$\text{MCOD}(\Sigma) \lesssim \sqrt{\frac{\log(p)}{n}}. \tag{1.9}$$

Here and throughout this paper the symbol $\lesssim$ is used whenever an inequality holds up to multiplicative constants, which are made precise in the statements of the theorems where such inequalities are proved. Theorem 3.2 in Section 3 shows that optimal separation distances with respect to the metric $\Delta(C^*)$ are sensitive to the size of the smallest cluster, $m^* = \min_{1\leqslant k\leqslant K} |G^*_k|$, in that no algorithm can estimate consistently clusters defined by (1.1) when

$$\Delta(C^*) \lesssim \left(\sqrt{\frac{\log(p)}{nm^*}} \bigvee \frac{\log(p)}{n} \right). \tag{1.10}$$

The first term will be dominant whenever the smallest cluster has size $m^* < n/\log(p)$, which will be the case in most situations. The second term in (1.10) becomes dominant whenever $m^* > n/\log(p)$, which can also happen when p scales as n, and we have a few balanced clusters.

The PECOK algorithm is tailored to the $\Delta(C^*)$ metric, and is shown in Theorem 5.1 to be near-minimax optimal. For instance, for balanced clusters, exact recovery is guaranteed when $\Delta(C^*) \gtrsim \sqrt{\frac{K\vee\log p}{m^* n}} + \frac{K\vee\log(p)}{n}$. This differs by factors in K from the $\Delta(C^*)$-minimax threshold,

for general K, whereas it is of optimal order when K is a constant, or grows as slowly as $\log p$. A similar discrepancy between minimax lower bounds and the performance of polynomial-time estimators has also been pinpointed in network clustering via the stochastic block model Chen and Xu, 2016 and in sparse PCA Berthet and Rigollet, 2013. It has been conjectured that, when K increases with n, there exists a gap between the statistical boundary, i.e. the minimal cluster separation for which a statistical method achieves perfect clustering with high probability, and the polynomial boundary, i.e. the minimal cluster separation for which there exists a polynomial-time algorithm that achieves perfect clustering. Further investigation of this computational trade-off is beyond the scope of this paper and we refer to Chen and Xu, 2016 and Berthet and Rigollet, 2013 for more details.

However, if we consider directly the metric MCOD(Σ), and its corresponding, larger, minimax threshold, we derive the COD algorithm, which is minimax optimal with respect to MCOD(Σ). In view of (1.8), it is also minimax optimal with respect to $\Delta(C^*)$, whenever there exist small clusters, the size of which does not change with n. The description of the two algorithms and theoretical properties are given in Sections 4 and 5, respectively, for exact block covariance models. Companions of these results, regarding the performance of the algorithms for approximate block covariance models are given in Section 6, in Theorem 6.1 and Theorem 6.2, respectively.

Table 2.1 below gives a snap-shot of our results, which for ease of presentation, correspond to the case of balanced clusters, with the same number of variables per cluster. We stress that neither our algorithms, nor our theory, is restricted to this case, but the exposition becomes more transparent in this situation.

Metric	Minimax thresh.	PECOK	COD
$d_1 =: \Delta(C^*)$	$\sqrt{\frac{\log p}{mn}} + \frac{\log p}{n}$	Always near-minimax optimal w.r.t. d_1.	Minimax optimal w.r.t. d_1 when m is constant.
$d_2 =:$ MCOD(Σ)	$\sqrt{\frac{\log p}{n}}$	Minimax optimal w.r.t. d_2 when $m > n/\log(p)$, $K = O(\log p)$.	Always minimax optimal w.r.t. d_2.

Table 2.1 – Algorithm performance relative to minimax thresholds of each metric

In this table m denotes the size of the smallest cluster in the partition. The performance of COD under d_1 follows from the second inequality in (1.8), whereas the performance of PECOK under d_2 follows from the last inequality in (1.8). The overall message transmitted by Table 1 and our analysis is that, irrespective of the separation metric, the COD algorithm will be most powerful whenever we expect to have at least one, possibly more, small clusters, a situation that is typically not handled well in practice by most of the popular clustering algorithms, see Bouveyron and Brunet-Saumard, 2014 for an in-depth review. The PECOK algorithm is expected to work best for larger clusters, in particular when there are no clusters of size one. We defer more comments on the relative numerical performance of the methods to the discussion Section 9.3.

We emphasize that both our algorithms are generally applicable, and our performance analysis is only in terms of the most difficult scenarios, when two different clusters are almost indistinguishable and yet, as our results show, consistently estimable. Our extensive simulation results confirm these theoretical findings.

We summarize below our key contributions.
(1) An identifiable model for variable clustering and metrics for cluster separation. We advocate model-based variable clustering, as a way of proposing objectively defined and interpretable clusters. We propose identifiable G-block covariance models for clustering, and prove cluster identifiability in Proposition 2.1 of Section 2.
(2) Minimax lower bounds on cluster separation metrics for exact partition recovery. Two of our main results are Theorem 3.2 and Theorem 3.1, presented in Section 3, in which we establish, respectively, minimax limits on the size of the $\Delta(C^*)$-cluster separation and MCOD(Σ)-cluster separation below which no algorithm can recover clusters defined by (1.1) consistently, from a sample of size n on X. To the best of our knowledge these are the first results of this type in variable clustering.
(3) Variable clustering procedures with guaranteed exact recovery of minimally separated clusters. The results of (**1**) and (**2**) provide a much needed framework for motivating variable clustering algorithm development and for clustering algorithm assessments.

In particular, they motivate a correction of a convex relaxation of the K-means algorithm, leading to our proposed PECOK procedure, based on Semi-Definite Programing (SDP). Theorem 5.1 shows it to be near-minimax optimal with respect to the $\Delta(C^*)$ metric. The PECOK - $\Delta(C^*)$ pairing is natural, as $\Delta(C^*)$ measures the difference of the "within cluster" signal relative to the "between clusters" signal, which is the idea that underlies K-means type procedures. To the best of our knowledge, this is the first work that explicitly shows what model-based clusters of variables can be estimated via K-means style methods, and assesses theoretically the quality of estimation. Moreover, our work shows that the results obtained in Berthet, Rigollet, and Srivastava, 2018, for the block Ising model, can be generalized to arbitrary values of K and unbalanced clusters.

The COD procedure is a companion of PECOK for clusters given by model (1.1), and is minimax optimal with respect to the MCOD(Σ) cluster separation, as established in Theorem 3.1. Another advantage of COD is of computational nature, as SDP-based methods, although convex, can be computationally involved.
(4) Comparison with corrected spectral variable clustering methods. In Section 5.4, we connect PECOK with another popular algorithm, spectral clustering. We show that although it may be less computationally involved than PECOK, good cluster recovery can only be theoretically guaranteed for very well separated clusters, well above the minimax optimal threshold.

1.3 Organization of the paper

The rest of the paper is organized as follows:

Sections 1.4 and 1.5 contain the notation and distributional assumptions used throughout the paper.

For clarity of exposition, Sections 2 - 5 contain results established for model (1.1), when is Γ a diagonal matrix. Extensions to the case when Γ has small off-diagonal entries are presented in Section 6.

Section 2 shows that we have a uniquely defined target of estimation, the partition G^*.

Section 3 derives the minimax thresholds on the separation metrics $\Delta(C^*)$ and $MCOD(\Sigma)$, respectively, for estimating G^* consistently.

Section 4 is devoted to the COD algorithm, and its analysis.

Section 5 is devoted to the PECOK algorithm and its analysis.

Section 5.4 analyses spectral clustering for variable clustering, and compares it with PECOK.

Section 6 contains extensions to approximate G-block covariance models.

Section 7 explores the numerical performance of our methods, and Section 8 presents their application to the clustering of putative brain areas using a real fMRI data.

Section 9 contains a discussion of our results and overall recommendations regarding the usage of our methods.

1.4 Notation

We denote by $\mathbf{X}$ the $n \times p$ matrix with rows corresponding to observations $X^{(i)} \in \mathbb{R}^p$, for $i = 1, \ldots, n$. The sample covariance matrix $\widehat{\Sigma}$ is defined by

$$\widehat{\Sigma} = \frac{1}{n}\mathbf{X}^t\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}(X^{(i)})^t. \tag{1.11}$$

Given a vector v and $q \geqslant 1$, $|v|_q$ stands for the ℓ_q norm. For a generic matrix M: $|M|_q$ denotes its the entry-wise ℓ_q norm, $\|M\|_{op}$ denotes its operator norm, and $\|M\|_F$ refers to the Frobenius norm. We use $M_{:a}$, $M_{b:}$, to denote the a-th column or, respectively, b-th row of a generic matrix M. The bracket $\langle ., . \rangle$ refers to the Frobenius scalar product. Given a matrix M, we denote $\mathrm{supp}(M)$ its support, that is the set of indices (i, j) such that $M_{ij} \neq 0$. I denotes the identity matrix. We define the variation semi-norm of a diagonal matrix D as $|D|_V := \max_a D_{aa} - \min_a D_{aa}$. We use $B \succcurlyeq 0$ to denote a symmetric and positive semidefinite matrix.

Throughout this paper will make use of the notation $c_1, c_2, \cdots$ to denote positive constants independent of n, p, K, m. The same letter, for instance c_1 may be used in different statements and may denote different constants, which are made clear within each statement, when there is no possibility for confusion.

We use $[p]$ to denote the set $\{1, \ldots, p\}$. We use the notation $a \overset{G}{\sim} b$ whenever $a, b \in G_k$, for the same k. Also, $m = \min_k |G_k|$ stands for the size of the smallest group of the partition G.

The notation $\gtrsim$ and $\lesssim$ is used for whenever the inequalities hold up to multiplicative numerical constants.

1.5 Distributional assumptions

For a p-dimensional random vector Y, its Orlicz norm is defined by $\|Y\|_{\psi_2} = \sup_{t \in \mathbb{R}^p:\ \|t\|_2 = 1} \inf\{s > 0 :\ \mathbb{E}[e^{(Z^t t)/s^2} \leqslant 2]\}$. Throughout the paper we will assume that X follows a sub-Gaussian distribution. Specifically, we use:

Assumption 1. (sub-Gaussian distributions) The exists $L > 0$ such that random vector $\Sigma^{-1/2}X$ satisfies $\|\Sigma^{-1/2}X\|_{\psi_2} \leqslant L$, where

Our class of distributions includes, in particular, that of bounded distributions, which may be of independent interest, as example (ii) illustrates. We will therefore also specialize some of our results to this case, in which case we will use directly.

Assumption 1-bis. (Bounded distributions) There exists $M > 0$ such that $\max_{i=1,\ldots,p} |X_i| \leqslant M$ almost surely.

Gaussian distributions satisfy Assumption 1 with $L = 1$. A bounded distribution is also sub-Gaussian, but the corresponding quantity L can be much larger than M, and sharper results can be obtained if Assumption 1-bis holds.

2 Cluster identifiability in G-block models

To keep the presentation focused, we consider in sections 2–5 the model (1.1) with Γ diagonal. We treat the case corresponding to a diagonally dominant Γ in Section 6 below. In the sequel, it is assumed that $p > 2$.

We observe that if the decomposition (1.1) holds for a partition G, it also holds for any sub-partition of G. It is natural therefore to seek the smallest (coarsest) of such partitions, that is the partition with the least number of groups for which (1.1) holds. Since the partition ordering is a partial order, the smallest partition is not necessarily unique. However, the following Lemma shows that uniqueness is guaranteed for our model class.

Lemma 2.1. *Consider any covariance matrix Σ.*

(a) There exists a unique minimal partition G^* such that $\Sigma = ACA^t + \Gamma$ for some diagonal matrix Γ, some membership matrix A associated to G^* and some matrix C.

(b) The partition G^* is given by the equivalence classes of the relation

$$a \equiv b \text{ if and only if } COD(a,b) := \max_{c \neq a,b} |\Sigma_{ac} - \Sigma_{bc}| = 0. \tag{2.1}$$

Proof of Lemma 2.1. If decomposition $\Sigma = ACA^t + \Gamma$ holds with A related to a partition G, then we have $COD(a,b) = 0$ for any a, b belonging to the same group of G. Hence, each group G_k of G is included in one of the equivalence class of $\equiv$. As a consequence, G is a finer partition than G^* as defined in (b). Hence, G^* is the (unique) minimal partition such that decomposition $\Sigma = ACA^t + \Gamma$ holds. □

As a consequence, the partition G^* is well-defined and is identifiable. Next, we discuss the definitions of MCOD and Δ metrics.

For any partition G, we let $\text{MCOD}(\Sigma, G) := \min_{a \overset{G}{\not\sim} b} COD(a,b)$, where we recall that the notation $a \overset{G}{\not\sim} b$ means that a and b are not in a same group of the partition G. By definition of G^*, we notice that $\text{MCOD}(\Sigma, G^*) > 0$ and the next proposition shows that G^* is characterized by this property.

Proposition 2.1. *Let G be any partition such that $\text{MCOD}(\Sigma, G) > 0$ and the decomposition $\Sigma = ACA^t + \Gamma$ holds with A associated to G. Then $G = G^*$.*

The proofs of this proposition and the following corollary are in appendix Section B. In what follows, we use the notation $\text{MCOD}(\Sigma)$ for $\text{MCOD}(\Sigma, G^*)$.

In general, without further restrictions on the model parameters, the decomposition $\Sigma = ACA^t + \Gamma$ with A relative to G^* is not unique. If, for instance Σ is the identity matrix I, then G^* is the complete partition (with p groups) and the decomposition (1.1) holds for any $(C, \Gamma) = (\lambda I, (1-\lambda)I)$ with $\lambda \in \mathbf{R}$.

Recall that $m^* := \min |G_k^*|$ stands for the size of the smallest cluster. If we assume that $m^* > 1$ (no singleton), then Γ is uniquely defined. Besides, the matrix C in (1.1) is only defined up to a permutation of its rows and columns. In the sequel, we denote C^* any of these matrices C. When the partition contains singletons ($m^* = 1$), the matrix decomposition $\Sigma = ACA^t + \Gamma$ is made unique (up to a permutation of row and columns of C) by putting the additional constraint that the entries Γ_{aa} corresponding to singletons are equal to 0. Since the definition of $\Delta(C)$ is invariant with respect to permutation of rows and columns, this implies that $\Delta(C^*)$ is well-defined for any covariance matrix Σ.

For arbitrary Σ, $\Delta(C^*)$ is not necessarily positive. Nevertheless, if $\Delta(C^*) > 0$, then G^* is characterized by this property.

Corollary 2.1. *Let G be a partition such that $m = \min_k |G_k| \geqslant 2$, the decomposition $\Sigma = ACA^t + \Gamma$ holds with A associated to G and $\Delta(C) > 0$. Then $G = G^*$.*

As pointed in (1.7), in the latent model (1.2), $\Delta(C^*)$ is equal to the square of the minimal L^2-norm between two latent variables. So, in this case, the condition $\Delta(C^*) > 0$ simply requires that all latent variables are distincts.

3 Minimax thresholds on cluster separation for perfect recovery

Before developing variable clustering procedures, we begin by assessing the limits of the size of each of the two cluster separation metrics below which no algorithm can be expected to recover the clusters perfectly. We denote by $m^* = \min_k |G_k^*|$ the size of the smallest cluster of the target partition G^* defined above. For $1 \leqslant m \leqslant p$ and $\eta, \tau > 0$, we consider the following sets of covariance matrices : $\mathcal{M}(m, \eta) := \{\Sigma : \text{MCOD}(\Sigma) > \eta |\Sigma|_\infty,\ m^* > m\}$ and $\mathcal{D}(m, \tau) := \{\Sigma : \Delta(C^*) > \tau |\Gamma|_\infty,\ m^* > m\}$. We use the notation $\mathbb{P}_\Sigma$ to refer to the normal distribution with covariance Σ.

Theorem 3.1. *There exists a positive constant c_2 such that, for any $1 \leqslant m \leqslant p/3$ and any η such that*

$$0 \leqslant \eta < \eta^* := c_2 \sqrt{\frac{\log(p)}{n}} \tag{3.1}$$

we have $\inf_{\widehat{G}} \sup_{\Sigma \in \mathcal{M}(m,\eta)} \mathbb{P}(\widehat{G} \neq G^*) \geqslant 1/7$, where the infimum is taken over all possible estimators.

We also have:

Theorem 3.2. *There exists a positive constant c_1 such that, for any $2 \leqslant m \leqslant p/2$ and any τ such that*

$$0 \leqslant \tau < \tau^* := c_1 \left[\sqrt{\frac{\log(p)}{n(m-1)}} \bigvee \frac{\log(p)}{n} \right] \tag{3.2}$$

then $\inf_{\widehat{G}} \sup_{\Sigma \in \mathcal{D}(m,\tau)} \mathbb{P}_\Sigma \big[\widehat{G} \neq G^*\big] \geqslant 1/7$, where the infimum is taken over all estimators.

Theorems 3.2 and 3.1 show that if either metric falls below the thresholds in (3.2) or (3.1), respectively, the estimated partition $\widehat{G}$, irrespective of the method of estimation, cannot achieve perfect recovery with high-probability uniformly over the set $\mathcal{M}(m, \eta)$ or $\mathcal{D}(m, \tau)$.

The proofs are given in appendix Section C. We note that $\Delta(C^*)$ minimax threshold takes into account the size m^* of the smallest cluster, and therefore the required cluster separation becomes smaller for large clusters. This is not the case for the second metric. The proof of (3.1) shows that even when we have $K = 3$ clusters, that are very large, of size $m^* = p/3$ each, the MCOD(Σ) threshold does not decrease with m^*.

4 COD for variable clustering

4.1 COD Procedure

We begin with a procedure that can be viewed as natural for model (1.1). It is based on the following intuition. Two indices a and b belong to the same cluster of G^*, if and only if COD$(a, b) = 0$, with COD defined in (2.1). Equivalently, a and b belong to the same cluster when

$$s\text{COD}(a,b) =: \max_{c\neq a,b} \frac{|\mathbf{cov}(X_a - X_b, X_c)|}{\sqrt{\mathbf{var}(X_b - X_a)\mathbf{var}(X_c)}} = \max_{c\neq a,b} |\mathbf{cor}(X_a - X_b, X_c)| = 0, \tag{4.1}$$

where sCOD stands for $s$caled COvariance Differences. In the following we work with this quantity, as it is scale invariant. It is natural to place a and b in the same cluster when the estimator $\widehat{s\text{COD}}(a, b)$ is below a certain threshold, where

$$\widehat{s\text{COD}}(a,b) := \max_{c\neq a,b} |\widehat{\mathbf{cor}}(X_a - X_b, X_c)| = \max_{c\neq a,b} \left| \frac{\widehat{\Sigma}_{ac} - \widehat{\Sigma}_{bc}}{\sqrt{\left(\widehat{\Sigma}_{aa} + \widehat{\Sigma}_{bb} - 2\widehat{\Sigma}_{ab}\right)\widehat{\Sigma}_{cc}}} \right|. \tag{4.2}$$

We estimate the partition $\widehat{G}$ according to the simple COD algorithm explained below. The algorithm does not require as input the specification of the number K of groups, which is automatically estimated by our procedure. Step 3(c) of the algorithm is called the "or" rule, and can be replaced with the "and" rule below, without changing the theoretical properties of our algorithm,

$$\widehat{G}_l = \left\{ j \in S : \widehat{s\text{COD}}(a_l, j) \vee \widehat{s\text{COD}}(b_l, j) \leqslant \alpha \right\}. \tag{4.3}$$

The numerical performance of these two rules are also very close through simulation studies, same as we reported on a related COD procedure on correlations Bunea, Giraud, and Luo, 2015. Due to these small differences, we will focus on the "or" rule for the sake of space.

The algorithmic complexity for computing $\widehat{\Sigma}$ is $O(p^2 n)$ and the complexity of COD is $O(p^3)$, so the overall complexity of our estimation procedure is $O\big(p^2(p \vee n)\big)$. The procedure is also valid when Γ has very small off-diagonal entries, and the results are presented in Section 6.

The COD Algorithm

- Input: $\widehat{\Sigma}$ and $\alpha > 0$
- Initialization: $S = \{1, \ldots, p\}$ and $l = 0$
- Repeat: while $S \neq \emptyset$
  1. $l \leftarrow l + 1$
  2. If $|S| = 1$ Then $\widehat{G}_l = S$
  3. If $|S| > 1$ Then
     (a) $(a_l, b_l) = \underset{a,b \in S,\ a \neq b}{\operatorname{argmin}} \widehat{s\mathrm{COD}}(a, b)$
     (b) If $\widehat{s\mathrm{COD}}(a_l, b_l) > \alpha$ Then $\widehat{G}_l = \{a_l\}$
     (c) If $\widehat{s\mathrm{COD}}(a_l, b_l) \leqslant \alpha$ Then
     $\widehat{G}_l = \left\{ j \in S : \widehat{s\mathrm{COD}}(a_l, j) \wedge \widehat{s\mathrm{COD}}(b_l, j) \leqslant \alpha \right\}$
  4. $S \leftarrow S \setminus \widehat{G}_l$
- Output: the partition $\widehat{G} = (\widehat{G}_l)_{l=1,\ldots,k}$

4.2 Perfect cluster recovery with COD for MCOD-minimax cluster separation

Theorem 4.1 shows that the partition $\widehat{G}$ produced by the COD algorithm has the property that $\widehat{G} = G^*$, with high probability, as soon as the separation $\mathrm{MCOD}(\Sigma)$ between clusters exceeds its minimax optimal threshold established in Theorem 3.1 of the previous section.

Theorem 4.1. *Under the distributional Assumption 1, there exists numerical constants $c_1, c_2 > 0$ such that, if*

$$\alpha \geqslant c_1 L^2 \sqrt{\frac{\log(p)}{n}}$$

and $MCOD(\Sigma) > 3\alpha|\Sigma|_\infty$, then we have exact cluster recovery with probability $1 - c_2/p$.

We recall that for Gaussian data, the constant $L = 1$. The proof is given in appendix Section D.

4.3 A Data-driven Calibration procedure for COD

The performance of the COD algorithm depends on the value of the threshold parameter α. Whereas Theorem 4.1 ensures that a good value for α is the order of $\sqrt{\log p/n}$, its optimal value depends on the actual distribution (at least through the subGaussian norm) and is unknown to the statistician. We propose below a new, fully data dependent, criterion for selecting α, and the corresponding partition $\widehat{G}$, from a set of candidate partitions $\mathcal{G}$. This criterion is based on data splitting: the estimators are built from a *training* sample, and then the selection involves an

independent *test* sample (Hold-Out sample). The main task is to design a meaningful selection criterion.

Let us consider two independent sample sets indexed by $i = 1, 2$, each of size $n/2$. The sample (1) will be a *training* dataset, and we denote by $\widehat{\mathcal{G}}^{(1)}$ a collection of partitions computed from sample (1), for instance via the COD algorithm with a varying threshold α. For any $a < b$, we set $\widehat{\Delta}_{ab}^{(i)} = \left[\widehat{Cor}^{(i)}(X_a - X_b, X_c)\right]_{c \neq a,b}$; $i = 1, 2$. Since $\Delta_{ab} := [Cor(X_a - X_b, X_c)]_{c \neq a,b}$ equals zero if and only if $a \overset{G}{\sim} b$, we want to select a partition G such that $\widehat{\Delta}_{ab}^{(2)} \mathbf{1}_{a \overset{G}{\nsim} b}$ is a good predictor of Δ_{ab}. To implement this principle, it remains to evaluate Δ_{ab} independently of $\widehat{\Delta}_{ab}^{(2)}$. For this evaluation, we propose to re-use the *training* sample (1) which has already been used to build the family of partitions $\widehat{\mathcal{G}}^{(1)}$. More precisely, we select $\widehat{G} \in \widehat{\mathcal{G}}^{(1)}$ by minimizing

$$\widehat{G} \in \operatorname*{argmin}_{G \in \widehat{\mathcal{G}}^{(1)}} CV(G) \quad \text{with} \quad CV(G) = \sum_{a<b} \left[|\widehat{\Delta}_{ab}^{(2)} \mathbf{1}_{a \overset{G}{\nsim} b} - \widehat{\Delta}_{ab}^{(1)}|_\infty^2 \right]. \tag{4.4}$$

An unusual feature of the above criterion is that the *training* sample (1) is involved both in the *training* stage and in the *test* stage.

The following proposition assesses the performance of $\widehat{G}$. We need the following additional assumption.

$(\mathbf{P1})$ If $Cor(X_a - X_b, X_c) = 0$ then $\mathbb{E}\,\widehat{Cor}(X_a - X_b, X_c) = 0$.

In general, the sample correlation is not an unbiased estimator of the population level correlation. Still, $(\mathbf{P1})$ is satisfied when the data are normally distributed or in a latent model (1.2) when the noise variables E_a have a symmetric distribution. The next proposition provides guaranties for the CV criterion averaged over the Hold-Out sample $\mathbb{E}^{(2)}[CV(G)]$. The proof is given in appendix Section D.2.

Proposition 4.1. *Assume that the distributional Assumption 1 and (**P1**) hold. Then, there exists a constant $c_1 > 0$ such that, when $MCOD(\Sigma) > c_1 |\Sigma|_\infty L^2 \sqrt{\log(p)/n}$, we have*

$$\mathbb{E}^{(2)}[CV(G^*)] \leqslant \min_{G \in \widehat{\mathcal{G}}^{(1)}} \mathbb{E}^{(2)}[CV(G)], \tag{4.5}$$

both with probability larger than $1 - 4/p$ and in expectation with respect to $\mathbb{P}^{(1)}$.

Under the condition $MCOD(\Sigma) > c_1 |\Sigma|_\infty L^2 \sqrt{\log(p)/n}$, Theorem 4.1 ensures that G^* belongs to $\widehat{\mathcal{G}}^{(1)}$ with high probability, whereas (4.5) suggests that the CV criterion is minimized at G^*.

If we consider a CV algorithm based on $\widehat{\text{COD}}(a, b)$ instead of $\widehat{\text{sCOD}}(a, b)$, then we can obtain a counterpart of Proposition 4.1 without requiring the additional assumption $(\mathbf{P1})$. Still, we favor the procedure based on $\widehat{\text{sCOD}}(a, b)$ mainly for its scale-invariance property.

5 Penalized convex K-means: PECOK

5.1 PECOK Algorithm

Motivated by the fact that the COD algorithm is minimax optimal with respect to the MCOD(Σ) metric, but not necessarily with respect to the $\Delta(C^*)$ metric (unless the size of the smallest cluster

is constant), we propose below an alternative procedure, that adapts to this metric. Our second method is a natural extension of one of the most popular clustering strategies. When we view the G-block covariance model as arising via the latent factor representation in **(i)** in the Introduction, the canonical clustering approach would be via the K-means algorithm Lloyd, 1982, which is NP-hard Awasthi et al., 2015. Following Peng and Wei Peng and Wei, 2007, we consider a convex relaxation of it, which is computationally feasible in polynomial time. We argue below that, for estimating clusters given by (1.1), one needs to further tailor it to our model. The statistical analysis of the modified procedure is the first to establish consistency of variable clustering via K-means type procedures, to the best of our knowledge.

The estimator offered by the standard K-means algorithm, with the number K of groups of G^* known, is

$$\widehat{G} \in \operatorname*{argmin}_{G} \operatorname{crit}(\mathbf{X}, G) \quad \text{with} \quad \operatorname{crit}(\mathbf{X}, G) = \sum_{a=1}^{p} \min_{k=1,\dots,K} \|\mathbf{X}_{:a} - \bar{\mathbf{X}}_{G_k}\|^2, \tag{5.1}$$

and $\bar{\mathbf{X}}_{G_k} = |G_k|^{-1} \sum_{a \in G_k} \mathbf{X}_{:a}$.

For a partition G, let us introduce the corresponding partnership matrix B by

$$B_{ab} = \begin{cases} \frac{1}{|G_k|} & \text{if } a \text{ and } b \text{ are in the same group } G_k, \\ 0 & \text{if } a \text{ and } b \text{ are in a different groups.} \end{cases} \tag{5.2}$$

we observe that $B_{ab} > 0$ if and only if $a \overset{G}{\sim} b$. In particular, there is a one-to-one correspondence between partitions G and their corresponding partnership matrices. It is shown in Peng and Wei Peng and Wei, 2007 that the collection of such matrices B is described by the collection $\mathcal{O}$ of orthogonal projectors fulfilling $\operatorname{tr}(B) = K$, $B1 = 1$ and $B_{ab} \geqslant 0$ for all a, b.

Theorem 2.2 in Peng and Wei Peng and Wei, 2007 shows that solving the K-means problem is equivalent to finding the global maximum

$$\bar{B} = \operatorname*{argmax}_{B \in \mathcal{O}} \langle \widehat{\Sigma}, B \rangle \tag{5.3}$$

and then recovering $\widehat{G}$ from $\bar{B}$.

The set of orthogonal projectors is not convex, so, following Peng and Wei Peng and Wei, 2007, we consider a convex relaxation $\mathcal{C}$ of $\mathcal{O}$ obtained by relaxing the condition "B orthogonal projector", by "B positive semi-definite", leading to

$$\mathcal{C} := \left\{ B \in \mathbb{R}^{p \times p} : \begin{array}{l} \bullet\ B \succcurlyeq 0 \ \text{ (symmetric and positive semidefinite)} \\ \bullet\ \sum_a B_{ab} = 1, \ \forall b \\ \bullet\ B_{ab} \geqslant 0, \ \forall a, b \\ \bullet\ \operatorname{tr}(B) = K \end{array} \right\}. \tag{5.4}$$

Thus, the (uncorrected) convex relaxation of K-means is equivalent with finding

$$\widetilde{B} = \operatorname*{argmax}_{B \in \mathcal{C}} \langle \widehat{\Sigma}, B \rangle. \tag{5.5}$$

To assess the relevance of this estimator, we first study its behavior at the population level, when $\widehat{\Sigma}$ is replaced by Σ in (5.5). Indeed, if the minimizer of our criterion does not recover the true partition at the population level, we cannot expect it to be consistent, even in a large sample asymptotic context (fixed p, n goes to infinity). We recall that $|\Gamma|_V := \max_a \Gamma_{aa} - \min_a \Gamma_{aa}$.

Proposition 5.1. *Assume that* $\Delta(C^*) > 2|\Gamma|_V/m^*$. *Then,* $B^* = \operatorname{argmax}_{B\in\mathcal{O}}\langle\Sigma, B\rangle$. *If* $\Delta(C^*) > 7|\Gamma|_V/m^*$, *then* $B^* = \operatorname{argmax}_{B\in\mathcal{C}}\langle\Sigma, B\rangle$.

For $\Delta(C^*)$ large enough, the population version of convexified K-means recovers B^*. The next proposition illustrates that the condition $\Delta(C^*) > 2|\Gamma|_V/m^*$ for population K-means is in fact necessary.

Proposition 5.2. *Consider the model* (1.1) *with*

$$C^* = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \beta & \beta-\tau \\ 0 & \beta-\tau & \beta \end{bmatrix} \qquad \Gamma = \begin{bmatrix} \gamma_+ & 0 & 0 \\ 0 & \gamma_- & 0 \\ 0 & 0 & \gamma_- \end{bmatrix}, \tag{5.6}$$

and $|G_1^*| = |G_2^*| = |G_3^*| = m^*$. *The population maximizer* $B_\Sigma = \operatorname{argmax}_{B\in\mathcal{O}}\langle\Sigma, B\rangle$ *is not equal to* B^* *as soon as* $2\tau = \Delta(C^*) < \frac{2}{m^*}|\Gamma|_V$.

The two propositions above are proved in Appendix A.1. As a consequence, when Γ is not proportionnal to the identity matrix, the population minimizers based on K-means and convexified K-means do not necessary recover the true partition even when the "within-between group" covariance gap is strictly positive. This undesirable behavior of K-means is not completely unexpected as K-means is a quantization algorithm which aims to find for clusters of similar width, instead of "homogeneous" clusters. Hence, we need to modify it for our purpose.

This leads us to suggesting a population level correction in Proposition 5.1. Indeed, as a direct Corollary of Proposition 5.1, we have

$$B^* = \operatorname*{argmin}_{B\in\mathcal{C}}\langle\Sigma - \Gamma, B\rangle \tag{5.7}$$

as long as $\Delta(C^*) > 0$. This suggests the following **Pe**nalized **Co**nvex **K**-means (PECOK) algorithm, in three steps. The main step 2 produces an estimator $\widehat{B}$ of B from which we derive the estimated partition $\widehat{G}$. We summarize this below.

The PECOK algorithm

1. Estimate Γ by $\widehat{\Gamma}$.
2. Estimate B^* by $\widehat{B} = \operatorname{argmax}_{B\in\mathcal{C}}\left(\langle\widehat{\Sigma}, B\rangle - \langle\widehat{\Gamma}, B\rangle\right)$.
3. Estimate G^* by applying a clustering algorithm to the columns of $\widehat{B}$

The required inputs for Step 2 of our algorithm are: (i) $\widehat{\Sigma}$, the sample covariance matrix; (ii) $\widehat{\Gamma}$, the estimator produced at Step 1; and (iii) K, the number of groups. Our only requirement on the clustering algorithm applied in Step 3 is that it succeeds to recover the partition G^* when applied to true partnership matrix B^*. The standard K-means algorithm Lloyd, 1982 seeded with K distinct centroids, kmeans++ Arthur and Vassilvitskii, 2007, or any approximate K-means as defined in (5.18) in Section 5.4, fulfill this property. This step is done at no additional statistical accuracy cost, as shown in Corollary 5.1 below.

We view the term $\langle \widehat{\Gamma}, B \rangle$ as a penalty term on B, with data dependent weights $\widehat{\Gamma}$. Therefore, the construction of an accurate estimator $\widehat{\Gamma}$ of Γ is a crucial step for guaranteeing the statistical optimality of the PECOK estimator.

5.2 Construction of $\widehat{\Gamma}$

Estimating Γ before estimating the partition itself is a non-trivial task, and needs to be done with care. We explain our estimation below and analyze it in Proposition A.1 in Appendix A.4. We show that this estimator of Γ is appropriate whenever Γ is a diagonal matrix (or diagonally dominant, with small off-diagonal entries). For any $a, b \in [p]$, define

$$V(a,b) := \max_{c,d \in [p]\setminus\{a,b\}} \frac{\left|(\widehat{\Sigma}_{ac} - \widehat{\Sigma}_{ad}) - (\widehat{\Sigma}_{bc} - \widehat{\Sigma}_{bd})\right|}{\sqrt{\widehat{\Sigma}_{cc} + \widehat{\Sigma}_{dd} - 2\widehat{\Sigma}_{cd}}}\,, \tag{5.8}$$

with the convention $0/0 = 0$. Guided by the block structure of Σ, we define

$$b_4(a) := \operatorname*{argmin}_{b \in [p]\setminus\{a\}} V(a,b) \quad \text{and} \quad b_{4'}(a) := \operatorname*{argmin}_{b \in [p]\setminus\{a, b_4(a)\}} V(a,b), \tag{5.9}$$

to be two elements "close" to a, that is two indices $b_4(a)$ and $b_{4'}(a)$ such that the empirical covariance difference $\widehat{\Sigma}_{bc} - \widehat{\Sigma}_{bd}$, for $b \in \{b_4(a), b_{4'}(a)\}$, is most similar to $\widehat{\Sigma}_{ac} - \widehat{\Sigma}_{ad}$, for all variables c and d not equal to a or either b's. It is expected that both b's either belong to the same group as a, or belong to some "close" groups. Then, our estimator $\widehat{\Gamma}$ is a diagonal matrix, defined by

$$\widehat{\Gamma}^{(4)}_{aa} = \widehat{\Sigma}_{aa} + \widehat{\Sigma}_{b_4(a)b_{4'}(a)} - \widehat{\Sigma}_{ab_4(a)} - \widehat{\Sigma}_{ab_{4'}(a)}, \quad \text{for } a = 1 \dots p \tag{5.10}$$

Intuitively, $\widehat{\Gamma}_{aa}$ should be close to $\Sigma_{aa} + \Sigma_{b_4(a)b_{4'}(a)} - \Sigma_{ab_4(a)} - \Sigma_{ab_{4'}(a)}$, which is equal to Γ_{aa} in the favorable event where both $b_4(a)$ and $b_{4'}(a)$ belong to the same group as a.

In general, the b's cannot be guaranteed to belong to the same group as a. Nevertheless, these two surrogates are close enough to a so that $|\widehat{\Gamma}_{aa} - \Gamma_{aa}|$ to be at most of the order of $|\Gamma|_\infty \sqrt{\log(p)/n}$ in ℓ^∞-norm, as shown in Proposition A.1 in Appendix A.4. In the next subsection, we show that $\widehat{\Gamma}^{(4)}$ is good enough to ensure that PECOK perfectly recovers G^* under minimal separation condition.

Note that PECOK requires the knowledge of the true number K of groups. When the number K of groups itself is unknown, we can modify the PECOK criterion by adding a penalty term as explained in a previous version of our work Bunea et al., 2016, Sec. 4. Alternatively, we propose in Section 7 a simple cross-validation procedure.

5.3 Perfect cluster recovery with PECOK for near-minimax Δ-cluster separation

We show in this section that the PECOK estimator recovers the clusters exactly, with high probability, at a near-minimax separation rate with respect to the $\Delta(C^*)$ metric.

Theorem 5.1. *There exist c_1, c_2, c_3 three positive constants such that the following holds. Let $\widehat{\Gamma}$ be any estimator of Γ, such that $|\widehat{\Gamma} - \Gamma|_V \leqslant \delta_{n,p}$ with probability $1 - c_1/p$. Then, under Assumption 1, and when $L^4 \log(p) \leqslant c_3 n$ and*

$$\Delta(C^*) \geqslant c_L \left[\|\Gamma\|_\infty \left\{ \sqrt{\frac{\log p}{m^* n}} + \sqrt{\frac{p}{n m^{*2}}} + \frac{\log(p)}{n} + \frac{p}{n m^*} \right\} + \frac{\delta_{n,p}}{m^*} \right] , \tag{5.11}$$

then $\widehat{B} = B^*$, with probability higher than $1 - c_1/p$. Here, c_L is a positive constant that only depends on L in Assumption 1. In particular, if $\widehat{\Gamma}$ is the estimator (5.10)*, the same conclusion holds with probability higher than $1 - c_2/p$ when*

$$\Delta(C^*) \geqslant c_L \|\Gamma\|_\infty \left\{ \sqrt{\frac{\log p}{m^* n}} + \sqrt{\frac{p}{n m^{*2}}} + \frac{\log(p)}{n} + \frac{p}{n m^*} \right\}. \tag{5.12}$$

The proof is given in Appendix A.3.

Remark 1. We left the term $\delta_{n,p}$ explicit in (5.11) in order to make clear how the estimation of Γ affects the cluster separation $\Delta(C^*)$ metric. Without a correction (i.e. taking $\widehat{\Gamma} = 0$), the term $\delta_{n,p}/m^*$ equals $|\Gamma|_V/m^*$ which is non zero (and does not decrease in a high-sample asymptotic) unless Γ has equal diagonal entries. This phenomenon is consistent with the population analysis in the previous subsection. Display (5.12) shows that the separation condition can be much decreased with the correction. In particular, for balanced clusters, exact recovery is guaranteed when

$$\Delta(C^*) \geqslant c_L \left[\sqrt{\frac{K \vee \log p}{m^* n}} + \frac{K \vee \log p}{n} \right] , \tag{5.13}$$

for an appropriate constant $c_L > 0$. In view of Theorem 3.2 the rate is minimax optimal when the number of clusters $K \leqslant \log(p)$. For an even larger number of clusters ($K \geqslant \log(p)$), the rate is only minimax up to some loss. For instance, if the clusters are balanced, we possibly lose a factor $K/\log(p)$ relative to the optimal rate. As discussed in the introduction, this gap is possibly due to a computational barrier and we refer to Chen and Xu, 2016 for more details.

Bounded variables X also follow subGaussian distribution. Nevertheless, the corresponding subGaussian norm L may be large and Theorem 5.1 can sometimes be improved, as in Theorem 5.2 below, proved in Appendix A.3

Theorem 5.2. *There exist c_1, c_2, c_3 three positive constants such that the following holds. Let $\widehat{\Gamma}$ be any estimator of Γ, such that $|\widehat{\Gamma} - \Gamma|_V \leqslant \delta_{n,p}$ with probability $1 - c_1/p$. Then, under Assumption 1-bis, and*

$$\Delta(C^*) \geqslant \qquad c_2 \left[M \|\Gamma\|_\infty^{1/2} \sqrt{\frac{p \log(p)}{n m^{*2}}} + M^2 \frac{p \log(p)}{n m^*} + \frac{\delta_{n,p}}{m^*} \right]. \tag{5.14}$$

then $\widehat{B} = B^*$, with probability higher than $1 - c_1/p$.

When we choose $\widehat{\Gamma}$ as in (5.10), the term $\delta_{n,p}/m^*$ can be simplified as under Assumption 1, see Proposition A.1 in Appendix A.4. For balanced clusters, Condition (5.14) can be simplified in

$$\Delta(C^*) \geqslant c_2 \left[M \|\Gamma\|_\infty^{1/2} \sqrt{\frac{K \log(p)}{n m^*}} + M^2 \frac{K \log(p)}{n} + \frac{\delta_{n,p}}{m^*} \right]. \tag{5.15}$$

In comparison to (5.13), the condition does not depend anymore on the subGaussian nom L, but the term $K \vee \log(p)$ has been replaced by $K \log(p)$.

Remark 2. For the Ising Block Model (1.3) with K balanced groups, we have $M = 1$ and $p = m^* K$, $C^* = (\omega_{in} - \omega_{out})I_K + \omega_{out} J$ and $\Gamma = (1 - \omega_{in})I_K$. As a consequence, no diagonal correction is needed, that is we can take $\widehat{\Gamma} = 0$, and since $|\Gamma|_V = 0$, we have $\delta_{n,p} = 0$. Then, for K balanced groups, condition (5.14) simplifies to

$$(\omega_{in} - \omega_{out}) \gtrsim K \sqrt{\frac{\log(p)}{np}} + \frac{K \log(p)}{n} \tag{5.16}$$

In the specific case $K = 2$, we recover (up to numerical multiplicative constants) the optimal rate proved in Berthet, Rigollet, and Srivastava, 2018. Our procedure and analysis provide a generalization of these results, as they are valid for general K and Theorem 5.2 also allows for unequal groups.

5.4 A comparison between PECOK and Spectral Clustering

In this section we discuss connections between the clustering methods introduced above and spectral clustering, a method that has become popular in network clustering. When used for variable clustering, uncorrected spectral clustering consists in applying a clustering algorithm, such as K-means, on the rows of the $p \times K$-matrix obtained by retaining the K leading eigenvectors of $\widehat{\Sigma}$.

SC algorithm

1. Compute $\widehat{V}$, the matrix of the K leading eigenvectors of $\widehat{\Sigma}$
2. Estimate G^* by applying a (rotation invariant) clustering method to the rows of $\widehat{V}$.

First, we recall the premise of spectral clustering, adapted to our context. For G^*-block covariance models as (1.1), we have $\Sigma - \Gamma = AC^*A^t$. Let U be the $p \times K$ matrix collecting the K leading eigenvectors of $\Sigma - \Gamma$. It has been shown, see e.g. Lemma 2.1 in Lei and Rinaldo, 2015, that a and b belong to the same cluster if and only if $U_{a:} = U_{b:}$ and if and only if $[UU^t]_{a:} = [UU^t]_{b:}$. Arguing as in Peng and Wei, 2007, we have the following.

Lemma 5.1. *SC algorithm is equivalent to the following algorithm:*

1. *Find* $\overline{B} = \operatorname{argmax}\{\langle \widehat{\Sigma}, B \rangle \ : \ tr(B) = K, \ I \succcurlyeq B \succcurlyeq 0\}$.
2. *Estimate G^* by applying a (rotation invariant) clustering method to the rows of* $\overline{B}$.

The connection between (unpenalized) PECOK and spectral clustering now becomes clear. The (unpenalized) PECOK estimator $\widetilde{B}$ (5.5) involves the calculation of

$$\widetilde{B} = \operatorname*{argmax}_B \{\langle \widehat{\Sigma}, B \rangle \ : \ B1 = 1, \ B_{ab} \geqslant 0, \ tr(B) = K, \ B \succcurlyeq 0\}. \tag{5.17}$$

Since the matrices B involved in (5.17) are doubly stochastic, their eigenvalues are smaller than 1 and hence (5.17) is equivalent to $\widetilde{B} = \mathrm{argmax}_B\{\langle\widehat{\Sigma}, B\rangle \,:\, B1 = 1,\, B_{ab} \geqslant 0,\, tr(B) = K,\, I \succcurlyeq B \succcurlyeq 0\}$. Note then that $\overline{B}$ can be viewed as a less constrained version of $\widetilde{B}$, in which $\mathcal{C}$ is replaced by $\overline{\mathcal{C}} = \{B :\, tr(B) = K,\, I \succcurlyeq B \succcurlyeq 0\}$, where we have dropped the $p(p+1)/2$ constraints given by $B1 = 1$, and $B_{ab} \geqslant 0$. We show in what follows that the possible computational gains resulting from such a strategy may result in severe losses in the theoretical guarantees for exact partition recovery. In addition, the proof of Lemma 5.1 shows that $\overline{B} = \widehat{V}\widehat{V}^t$, so, contrary to $\widehat{B}$, the estimator $\overline{B}$ is (almost surely) never equal to B^*.

In view of this connection between Spectral clustering and unpenalized PECOK and on the fact that the population justification of spectral clustering deals with the spectral decomposition of $\Sigma - \Gamma$, this leads to propose the following corrected version of the algorithm based on $\widetilde{\Sigma} := \widehat{\Sigma} - \widehat{\Gamma}$.

CSC algorithm

1. Compute $\widehat{U}$, the matrix of the K leading eigenvectors of $\widetilde{\Sigma} := \widehat{\Sigma} - \widehat{\Gamma}$
2. Estimate G^* by clustering the rows of $\widehat{U}$, via an η-approximation of K-means (5.18).

For $\eta > 1$, an η-approximation of K-means is a clustering algorithm producing a partition $\widehat{G}$ such that

$$\mathrm{crit}\left(\widehat{U}^t, \widehat{G}\right) \leqslant \eta \min_G \mathrm{crit}\left(\widehat{U}^t, G\right), \tag{5.18}$$

with $\mathrm{crit}(\cdot,\cdot)$ the K-means criterion (5.1). Although solving K-means is NP-Hard Awasthi et al., 2015, there exist polynomial time approximate K-means algorithms, see Kumar, Sabharwal, and Sen, 2004. As a consequence of the above discussion, the first step of CSC can be interpreted as a relaxation of the program associated to PECOK estimator $\widehat{B}$.

In the sequel, we provide some results for CSC procedure. To simplify the presentation, we assume in the following that all the groups have the same size $|G_1^*| = \ldots = |G_K^*| = m = p/K$. We emphasize that this information is not required by either PECOK or CSC, or in the proof of Proposition 5.3 below. We only use it here to illustrate the issues associated with CSC in a way that is not cluttered by unnecessary notation. We denote by $\mathcal{S}_K$ the set of permutations on $\{1, \ldots, K\}$ and we denote by

$$\overline{L}(\widehat{G}, G^*) = \min_{\sigma \in \mathcal{S}_K} \sum_{k=1}^{K} \frac{|G_k^* \setminus \widehat{G}_{\sigma(k)}|}{m} \tag{5.19}$$

the sum of the ratios of miss-assigned variables with indices in G_k^*. In the previous sections, we studied perfect recovery of G^*, which would correspond to $\overline{L}(\widehat{G}, G^*) = 0$. We give below conditions under which $\overline{L}(\widehat{G}, G^*) \leqslant \rho$, for an appropriate quantity $\rho < 1$, and we show that very small values of ρ require large cluster separation, possibly much larger than the minimax optimal rate. We begin with a general theorem pertaining to partial partition recovery by CSC, under restrictions on the smallest eigenvalue $\lambda_K(C^*)$ of C^*.

Proposition 5.3. *We let $Re(\Sigma) = tr(\Sigma)/\|\Sigma\|_{op}$ denote the effective rank of Σ. There exist $c_{\eta,L} > 0$ only depending on η and L and a numerical constant c_1 such that the following holds under Assumption 1. For any $0 < \rho < 1$, if*

$$\lambda_K(C^*) \geqslant \frac{c_{\eta,L}\sqrt{K}\|\Sigma\|_{op}}{m^*\sqrt{\rho}}\sqrt{\frac{Re(\Sigma) \vee \log(p)}{n}}, \tag{5.20}$$

then $\overline{L}(\widehat{G}, G^*) \leqslant \rho$, with probability larger than $1 - c_1/p$.

The proof extends the arguments of Lei and Rinaldo, 2015, initially developed for clustering procedures in stochastic block models, to our context. Specifically, we relate the error $\overline{L}(\widehat{G}, G^*)$ to the noise level, quantified in this problem by $\|\widetilde{\Sigma} - AC^*A^t\|_{op}$. We then employ the results of Koltchinskii and Lounici, 2017 to show that this operator norm can be controlled, with high probability, which leads to the conclusion of the theorem.

We observe that $\Delta(C^*) \geqslant 2\lambda_K(C^*)$, so the lower bound (5.20) on $\lambda_K(C^*)$ enforces the same lower-bound on $\Delta(C^*)$. As n goes to infinity, the right hand side of Condition (5.20) goes to zero, and CSC is therefore consistent in a large sample asymptotic. In contrast, we emphasize that (uncorrected) SC algorithm is not consistent as can be shown by a population analysis similar to that of Proposition 5.2.

To further facilitate the comparison between CSC and PECOK, we discuss both the conditions and the conclusion of this theorem in the simple setting where $C^* = \tau I_K$ and $\Gamma = I_p$. Then, the cluster separation measures coincide up to a factor 2, $\Delta(C^*) = 2\lambda_K(C^*) = 2\tau$.

Corollary 5.1 (Illustrative example: $C^* = \tau I_K$ and $\Gamma = I_p$)**.** *There exist three positive numerical constants $c_{\eta,L}$, $c'_{\eta,L}$ and c_3 such that the following holds under Assumption 1. For any $0 < \rho < 1$, if*

$$\rho \geqslant c_{\eta,L}\Big[\frac{K^2}{n} + \frac{K\log(p)}{n}\Big] \qquad \textit{and} \qquad \tau \geqslant c'_{\eta,L}\Big[\frac{K^2}{\rho n} \vee \frac{K}{\sqrt{\rho nm}}\Big], \tag{5.21}$$

then $\overline{L}(\widehat{G}, G^*) \leqslant \rho$, with probability larger than $1 - c_3/p$.

Recall that, as a benchmark, Theorem 5.1 above states that, when $\widehat{G}$ is obtained via the PECOK algorithm, and if $\tau \gtrsim \sqrt{\frac{K\vee\log p}{mn}} + \frac{\log(p)\vee K}{n}$, then $\overline{L}(\widehat{G}, G^*) = 0$, or equivalently, $\widehat{G} = G^*$, with high probability. We can therefore provide the following comparison.

- If we consider ρ as a user specified small value, independent of n or p, and if the number of groups K is either a constant or grows at most as $\log p$, then the size of the cluster separation given by either Condition (5.21) or by PECOK are essentially the same, up to unavoidable $\log p$ factors. The difference is that, in this regime, CSC guarantees recovery up to a fixed, small, fraction of mistakes, whereas PECOK guarantees exact recovery.
- Although perfect recovery, with high probability, cannot be guaranteed for CSC, we could be close to it by requiring ρ to be close to zero. In this case, the distinctions between Conditions (5.21) and that for PECOK become much more pronounced.

- When we move away from the case $C^* = \tau I_K$, the comparison becomes even less favorable to CSC. For instance, when $\Gamma = I$ and $C^* = \tau I_K + \alpha J$, with J being the matrix with all entries equal to one, as in the Ising Block model discuss page 41. Notice that in this case we continue to have $\Delta(C^*) = 2\lambda_K(C^*) = 2\tau$. Then, for a given, fixed, value of ρ and K fixed, condition (5.20) requires the cluster separation

$$\tau \gtrsim \frac{\alpha\sqrt{\log(p)}}{\sqrt{n\rho}}\,, \tag{5.22}$$

  which is independent of m, unlike the minimax cluster separation rate that we established in Theorem 3.2 above. Therefore, the correction strategies employed in SBM are not directly transferable to variable clustering, which further supports the merits our PECOK method.

All the results of this section are proved in appendix Section F.

6 Approximate G-block covariance models

In the previous sections, we have proved that under some separation conditions, COD and PECOK procedures are able to exactly recover the partition G^*. However, in practical situations, the separation conditions may not be met. Besides, if the entries of Σ have been modified by an infinitesimal perturbation, then the corresponding partition G^* would consist of p singletons.

As a consequence, it may be more realistic and more appealing from a practical point of view to look for a partition $G[K]$ with $K < |G^*|$ groups such that Σ is close to a matrix of the form $ACA^t + \Gamma$ where Γ is diagonal and A is associated to $G[K]$. This is equivalent to considering a decomposition $\Sigma = ACA^t + \Gamma$ with Γ non-diagonal, where the non-diagonal entries of Γ are small. In the sequence, we write $R = \Gamma - \mathrm{Diag}(\Gamma)$ for the matrix of the off-diagonal elements of Γ and $D = \mathrm{Diag}(\Gamma)$ for the diagonal matrix given by the diagonal of Γ.

In the next subsection, we discuss under which conditions the partition $G[K]$ is identifiable and then, we prove that COD and PECOK are able to recover these partitions.

6.1 Identifiability of approximate G-block covariance models

When Γ is allowed to be not exactly equal to a diagonal matrix, we encounter a further identifiability issue, as a generic matrix Σ may admit many decompositions $\Sigma = ACA^t + \Gamma$. In fact, such a decomposition holds for any membership matrix A and any matrix C if we define $\Gamma = \Sigma - ACA^t$. So we need to specify the kind of decomposition that we are looking for. For K being fixed, we would like to consider the partition G with K clusters that maximizes the distance between goups (e.g. $\mathrm{MCOD}(\Sigma, G)$) while having the smallest possible noise term $|R|_\infty$. Unfortunately, such a partition G does not necessarily exist and is not necessarily unique. Let us illustrate this situation with a simple example.

Example. Assume that Σ is given by $\Sigma = \begin{bmatrix} 2r & 0 & 0 \\ 0 & 2r & 0 \\ 0 & 0 & 2r \end{bmatrix} + I_p$, with $r > 0$, with the convention that each entry corresponds to a block of size 2. Considering partitions with 2 groups and allowing Γ

to be non diagonal, we can decompose Σ using different partitions. For instance

$$\Sigma = \underbrace{\begin{bmatrix} 2r & 0 & 0 \\ 0 & r & r \\ 0 & r & r \end{bmatrix}}_{=A_1 C_1 A_1^t} + \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & r & -r \\ 0 & -r & r \end{bmatrix} + I_p}_{=\Gamma_1} = \underbrace{\begin{bmatrix} r & r & 0 \\ r & r & 0 \\ 0 & 0 & 2r \end{bmatrix}}_{=A_2 C_2 A_2^t} + \underbrace{\begin{bmatrix} r & -r & 0 \\ -r & r & 0 \\ 0 & 0 & 0 \end{bmatrix} + I_p}_{=\Gamma_2}. \tag{6.1}$$

Importantly, the two decompositions correspond to two different partitions G_1 and G_2 and both decompositions have $|R_i|_\infty = r$ and $\text{MCOD}(\Sigma, G_i) = 2r = 2|R|_\infty$, for $i = 1, 2$. In addition, no decomposition $\Sigma = ACA^t + D + R$ with associated partition in 2 groups, satisfies $\text{MCOD}(\Sigma, G) > 2r$ or $|R|_\infty < r$. As a consequence, there is no satisfying way to define a unique partition maximizing $\text{MCOD}(\Sigma, G)$, while having $|R|_\infty$ as small as possible. We show below that the cutoff $\text{MCOD}(\Sigma, G) > 2|R|_\infty$ is actually sufficient for partition identifiability.

For this, let us define $\mathcal{P}_j(\Sigma, K), \quad j \in \{1, 2\}$ as the set of quadruplets (A, C, D, R) such that $\Sigma = ACA^t + D + R$, with A a membership matrix associated to a partition G with K groups with $\min_k |G_K| \geqslant j$, and D and R defined as above. Hence $\mathcal{P}_1$ corresponds to partitions without restrictions on the minimum group size. For instance, singletons are allowed. In contrast $\mathcal{P}_2$ only contains partitions without singletons. We define

$$\rho_1(\Sigma, K) = \max\{\text{MCOD}(\Sigma, G)/|R|_\infty : (A, C, D, R) \in \mathcal{P}_1(\Sigma, K) \text{ and } G \text{ associated to } A\}, \tag{6.2}$$

$$\rho_2(\Sigma, K) = \max\{\Delta(C)/|R|_\infty : (A, C, D, R) \in \mathcal{P}_2(\Sigma, K)\}. \tag{6.3}$$

We view ρ_1 and ρ_2 as respective measures of "purity" of the block structure of Σ.

Proposition 6.1. *(i) Assume that $\rho_1(\Sigma, K) > 2$. Then, there exists a unique partition G such that there exists a decomposition $\Sigma = ACA^t + \Gamma$, with A associated to G and $\text{MCOD}(\Sigma, G) > 2|R|_\infty$. We denote by $G_1[K]$ this partition.*

(ii) Assume that $\rho_2(\Sigma, K) > 8$. Then, there exists a unique partition G with $\min_k |G_k| \geqslant 2$, such that there exists a decomposition $\Sigma = ACA^t + \Gamma$, with A associated to G and $\Delta(C) > 8|R|_\infty$. We denote by $G_2[K]$ this partition.

(iii) In addition, if both $\rho_1(\Sigma, K) > 2$ and $\rho_2(\Sigma, K) > 8$, then $G_1[K] = G_2[K]$.

The conditions $\rho_1(\Sigma, K) > 2$ and $\rho_2(\Sigma, K) > 8$ are minimal for defining uniquely the partition $G_1[K]$. For ρ_1, this has been illustrated in the example above the proposition. For ρ_2, we provide a counter example when $\rho_2(\Sigma, K) = 8$ in appendix Section B.3. The proof of Proposition 6.1 is given in appendix Section B.2.

The conclusion of Proposition 6.1 does essentially revert to that of Proposition 2.1 of Section 2 as soon as $|R|_\infty$ is small enough respective to the cluster separation sizes. Denoting K^* the number of groups of G^*, we observe that $G_1[K^*] = G^*$ and $G_2[K^*] = G^*$ if $m^* \geqslant 2$. Besides, $\rho_1(\Sigma, K) = \rho_2(\Sigma, K) = 0$ for $K > K^*$. For $K < K^*$ and when $G_1[K]$ (resp. $G_2[K]$) are well defined, then the partition $G_1[K]$ (resp. $G_2[K]$) is coarser than G^*. In other words, $G_1[K]$ is derived from G^* by merging groups G_k^* thereby increasing $MCOD(\Sigma, G)$ (resp. $\Delta(C)$) while requiring $|R|_\infty$ to be small enough.

We point out that, in general, there is no unique decomposition $\Sigma = ACA^t + \Gamma$ with A associated to $G_2[K]$, even when $\min_k |G_2[K]_k| \geqslant 2$. Actually, it can be possible to change some entries of C and R, while keeping $C + R$, $\Delta(C)$ and $|R|_\infty$ unchanged.

6.2 The COD algorithm for approximate G-block covariance models

We show below that the COD algorithm is still applicable if Σ has small departures from a block structure. We set $\lambda_{min}(\Sigma)$ for the smallest eigenvalue of Σ.

Theorem 6.1. *Under the distributional Assumption 1, there exist numerical constants $c_1, c_2 > 0$ such that the following holds for all $\alpha \geqslant c_1 L^2 \sqrt{\frac{\log p}{n}}$. If, for some partition G and decomposition $\Sigma = ACA^t + R + D$, we have*

$$|R|_\infty \leqslant \frac{\lambda_{min}(\Sigma)}{2\sqrt{2}}\alpha \quad \textit{and} \quad \textit{MCOD}(\Sigma, G) > 3\alpha|\Sigma|_\infty \,, \tag{6.4}$$

then COD *recovers G with probability higher than $1 - c_2/p$.*

The proof is given in appendix Section D. If G satisfies the assumptions of Theorem 6.1, then it follows from Proposition 6.1 that $G = G_1[K]$ for some $K > 0$. First, consider the situation where the tuning parameter α is chosen to be of the order $\sqrt{\log(p)/n}$. If $\text{MCOD}(\Sigma, G^*) \geqslant 3\alpha|\Sigma|_\infty$, then COD selects G^* with high probability. If $\text{MCOD}(\Sigma, G^*)$ is smaller than this threshold, then no procedure is able to recover G^* with high probability (Theorem 3.1). Nevertheless, COD is able to recover a coarser partition $G_1[K]$ whose corresponding MCOD metric $\text{MCOD}(\Sigma, G)$ is higher than the threshold $3\alpha|\Sigma|_\infty$ and whose matrix R is small enough. For larger α, then COD recovers a coarser partition G (corresponding to $G_1[K]$ with a smaller K) whose corresponding approximation $|R|_\infty$ is allowed to be larger.

6.3 The PECOK algorithm for approximate G-block covariance models

In this subsection, we investigate the behavior of PECOK under the approximate G-block models. The number K of groups being fixed, we assume that $\rho_2(\Sigma, K) > 8$ so that $G_2[K]$ is well defined. We shall prove that PECOK recovers $G_2[K]$ with high probability. By abusing the notation, we denote in this subsection G^* for the target partition $G_2[K]$, B^* for the associated partnership matrix and $(A, C^*, D, R) \in \mathcal{P}_2(\Sigma, K)$ any decomposition of Σ maximizing $\Delta(C)/|R|_\infty$.

Similarly to Proposition 5.1, we first provide sufficient conditions on C^* under which a population version of PECOK can recover the true partition.

Proposition 6.2. *If $\Delta(C^*) > \frac{7|D|_V + 2\|R\|_{op}}{m} + 3|R|_\infty$, then $B^* = \operatorname{argmin}_{B \in \mathcal{C}} \langle \Sigma, B \rangle$.*

Corollary 6.1. *If $\Delta(C^*) > 3|R|_\infty + \frac{2\|R\|_{op}}{m}$, then $B^* = \operatorname{argmin}_{B \in \mathcal{C}} \langle \Sigma - D, B \rangle$.*

In contrast to the exact G-block model, the cluster distance $\Delta(C^*)$ now needs to be larger than $|R|_\infty$ for the population version to recover the true partition. The $|R|_\infty$ condition is fact necessary as discussed in subsection 6.1. In comparison to the necessary conditions discussed in subsection 6.1, there is an additional $\|R\|_{op}/m$ term. The proofs are given in Appendix A.2.

We now examine the behavior of PECOK when we specify the estimator $\widehat{\Gamma}$ to be as in (5.10). Note that in this approximate block covariance setting, the diagonal estimator $\widehat{\Gamma}$ is in fact an estimator of the diagonal matrix D. In order to derive deviation bounds for our estimator $\widehat{\Gamma}$, we need the following diagonal dominance assumption.

Assumption 2: (diagonal dominance of Γ) The matrix $\Gamma = D + R$ fulfills

$$\Gamma_{aa} \geqslant 3 \max_{c:c\neq a} |\Gamma_{ac}| \quad \text{(or equivalently} \quad D_{aa} \geqslant 3 \max_{c:c\neq a} |R_{ac}|). \tag{6.5}$$

The next theorem states that PECOK estimator $\widehat{B}$ recovers the groups under similar conditions to that of Theorem 5.1 if R is small enough. The proof is given in appendix Section A.3, with proofs of intermediate results given in appendix Section E.

Theorem 6.2. *There exist c_1, c_2, c_L, c'_L four positive constants such that the following holds. Under Assumptions 1 and 2, and when $L^4 \log(p) \leqslant c_1 n$ and*

$$|R|_\infty + \frac{\sqrt{|R|_\infty |D|_\infty} + \|R\|_{op}}{m} \leqslant c_L \|\Gamma\|_{op} \left\{ \sqrt{\frac{\log p}{mn}} + \sqrt{\frac{p}{nm^2}} + \frac{\log(p)}{n} + \frac{p}{nm} \right\} \tag{6.6}$$

we have $\widehat{B} = B^*$, with probability higher than $1 - c_2/p$, as soon as

$$\Delta(C^*) \geqslant c'_L \left[\|\Gamma\|_{op} \left\{ \sqrt{\frac{\log p}{mn}} + \sqrt{\frac{p}{nm^2}} + \frac{\log(p)}{n} + \frac{p}{nm} \right\} \right] , \tag{6.7}$$

So, as a long as $|R|_\infty$ and $\|R\|_{op}$ are small enough so that (6.6) are satisfied, textscPecok algorithm will correctly identify the target partition G^* at the Δ-(near) optimal minimax level (6.7). A counterpart of Theorem 6.2 for Assumption 1-bis is provided in Appendix A.3.

7 Simulation results

In this section we verify numerically our theoretical findings and also illustrate the finite sample performance of our methods. The implementation of PECOK can be found at github.com/martinroyer/pecok/ and that of COD at CRAN.R-project.org/package=cord.

7.1 Simulation design

Recall our G-latent covariance $\Sigma = ACA' + \Gamma$. Under various scenarios of A and Γ to be described momentarily, we consider the following models for C:

- Model 1: $C = B^T B$ where B is a random $(K-1) \times K$ matrix with independent entries. Each entry takes the value $+1$ and -1 with equal probability $0.5 \times K^{-1/2}$, and the value 0 with probability $1 - K^{1/2}$.
- Model 2: $C = C' - 0.001I$ where C' is generated by Model 1.

The matrix C is positive semi-definite in Model 1 and negative definite in Model 2. In the first two simulation scenarios (referred to as M1 and M2 thereafter), we set C derived from Models 1–2 respectively, and specify A to correspond to $K = 10$ equal-size groups of variables (or equivalently $m = p/K$).

In a third scenario (M1S), we specify A such that it corresponds to the existence of 5 singletons, which are variables that form their own groups of size 1, respectively, and the remaining $K - 5$ groups have equal-size, while C is the same as M1.

In these first three scenarios, we employ diagonal $\Gamma = D$ where the p diagonal entires of D are random permutations of $\{0.5, 0.5 + 1.5/(p-1), \ldots, 2\}$.

In the fourth scenario (M1P), we consider the approximate G-block model by setting $\Gamma = D + R$ where $R = 0.1 \cdot U^t U / \max(U^t U)$ and U is a $p \times p$ matrix with iid random entries from a uniform distribution on $[-1, 1]$. We run these four scenarios for two representative $p = 200$ and $p = 1600$, and for sample sizes $n = 100, 300, \ldots, 900$ unless noted otherwise. All simulations are repeated 100 times.

The goal of our methods is to create sub-groups of vectors of dimension n, from a given collection of p vectors of observations, each of dimension n. This task can be viewed as that of clustering p objects in $\mathbb{R}^n$. The existing data clustering algorithms are not tailored to recovering groups with this structure, but they can serve as comparative methods. We thus compare the performance of COD and PECOK with three popular clustering algorithms: K-means, Hierarchical Clustering (HC) and spectral clustering. We apply K-means on the columns of the $n \times p$ matrix of n observations, and use the negative correlation as distance matrix in HC. The spectral clustering algorithm is discussed in Section 5.4, which does not correct for Γ. We use the standard K-means and HC algorithms in R, and we also implement our COD algorithm in R. We also include a variant of the COD algorithm suggested by a reviewer. This variant applies the connected component algorithm (as implemented in CRAN package igraph available at `https://CRAN.R-project.org/package=igraph`) to a graph converted from our proposed COD metric matrix thresholded at level α, which will be referred to as COD-CC. The spectral clustering algorithm is based on the widely used Python package scikit-learn, and we implement our PECOK algorithm also using Python.

The three competing methods require specification of the number of groups K, and we will use the true $K = 10$ to evaluate their oracle performance. For our proposed methods, we use the CV approach proposed in Section 4.3 to determine K in a data-adaptive fashion. We select either K from a grid in PECOK or the K-related threshold α also from a grid in COD, using the two independent datasets of size n proposed in Section 4.3. The grid for PECOK is $2, 4, \ldots, 40$. Since the theoretical choice of α is proportional to $n^{-1/2} \log^{1/2} p$, we use a grid of $\alpha / \big(n^{-1/2} \log^{1/2} p\big) = 0.25, 0.5, \ldots, 5$ in COD.

7.2 Exact recovery performance and comparison

Figure 2.1 shows the average percentages of exact recovery across 100 runs by K-means, HC, COD, COD-CC, and PECOK when n varies. Under one setting with a very large $p = 1600$, PECOK did not complete computation within two weeks, and thus its numerical performance for large $p = 1600$ was not reported in the figure. COD and COD-CC clearly outperform all other competing methods (K-means, HC and spectral clustering) when n is about 300 or larger in all the scenarios. K-means and HC, even with the oracle choice of K and large $n = 900$, fail to recover the true groups exactly. COD-CC is better than COD for small $n = 100$, there are almost no finite sample differences between COD and COD-CC for $n = 900$, which is consistent with our theory showing that they share the same rates described before. COD-CC and COD have similar performance across almost all models, and COD-CC achieves close to 100% recovery for smaller samples than COD under the model with singletons (M1S). Except for the model containing singletons (M1S), PECOK has the best performance for small $n = 100$ and $p = 200$, and achieves close to 100% as COD and COD-CC for larger n. Under the singleton scenario M1S, COD-CC has the best performance for all

n, while the difference between COD-CC and COD vanishes when $n = 900$. Under this model, the recovery percentages of PECOK increase with n but only reaches about 40% when $n = 900$ and $p = 200$. This is consistent with our theoretical results that PECOK is not expected to work well in the presence of singletons, while COD adapts to this situation. We also note that the competing methods used for comparison are able to recover clusters very close to the truth (measured by the adjusted rand index, or ARI), see appendix Section H for this partial recovery comparison.

7.3 The importance of correcting for Γ in PECOK

The step 1 of our PECOK algorithm is to estimate and correct for Γ. We illustrate the importance of this step by comparing its performance with two closely related methods, Γ-uncorrected PECOK and K-means, neither of which corrects for Γ. The Γ-uncorrected PECOK algorithm simply replaces the estimated $\hat{\Gamma}$ in the step 1 of PECOK by a zero matrix. We use the true K as input to these two methods to assess their oracle performance under the true K, while the parameter K in PECOK is selected by CV as described before. To fix ideas, we use scenario M1 described before and set $p = 200$. Figure 2.2(a) shows that the exact recovery percentages of K-means are close to zero, and those of Γ-uncorrected PECOK are smaller than 30%, across all n. After correcting for Γ, PECOK yields close to 100% exact recovery when n increases to 300.

7.4 Comparison under varying m

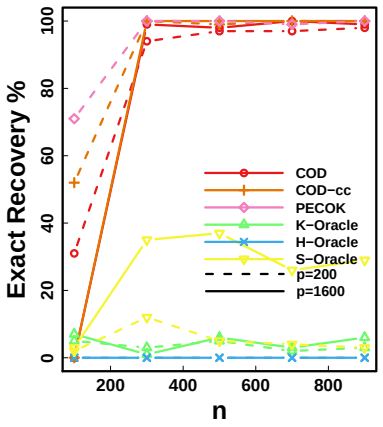
In this section we illustrate the finite sample performance of PECOK when m varies. We use the simulation scenario M1 under $p = 200$, and consider $m = 50$ and $m = 20$, for $n = 60, 80, 100, 150$, to compare the increasing trend in performance before reaching 100% when n reaches 200 as shown in Figure 2.1. The rest of the simulation set-up is the same as that of the previous section. As predicted by our theory, Figure 2.2(b) shows the percentages of exact recovery is better for a larger $m = 50$, compared with $m = 20$, for the same n.

8 Data analysis

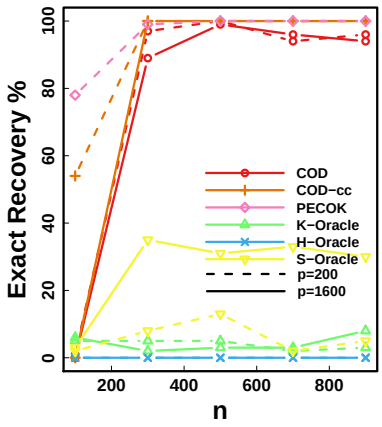
Using functional MRI data, Power et al., 2011 found that the human brain putative areas are organized into clusters, sometimes referred to as networks or functional systems. We use a publicly available fMRI dataset to illustrate the clusters recovered by different methods. The dataset was originally published in Xue, Aron, and Poldrack, 2008 and is publicly available from Open fMRI (`https://openfmri.org/data-sets`) under the accession number ds000007. We will focus on analyzing two scan sessions from subject 1 under a visual-motor stop/go task (task 1). Before performing the analysis, we follow the preprocessing steps suggested by Xue, Aron, and Poldrack, 2008, and we follow Power et al., 2011 to subsample the whole brain data using $p = 264$ putative areas, see appendix Section I for details. This subject was also scanned in two separate sessions, and each session yielded $n = 180$ samples for each putative area.

We apply our CV approach described in Section 4.3 to these two session data. Using the first scan session data only, we first estimate $\hat{G}$ using COD and COD-CC on a fine grid of $\alpha = c\sqrt{\log(p)/n}$ where $c = 0.5, 0.6, \ldots, 3$. For a fair comparison, we set K in PECOK to be the same as the resulting K's found by COD. We then use the second session data to evaluate the CV loss $CV(G)$ given in Section 4.3. Among our methods (COD, COD-CC, and PECOK), COD yields the smallest CV

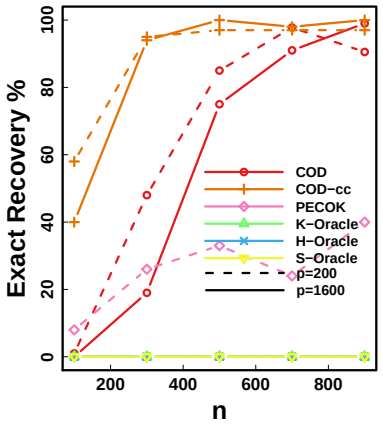
Figure 2.1 – Percentages of exact recovery by K-means (K-Oracle, medium green lines, triangle points), HC (H-Oracle, light blue lines, cross points), spectral clustering (S-Oracle, light yellow lines, up-side-down triangle points), COD (dark red lines, circle points), COD-CC (light orange lines, plus points), PECOK (light pink lines, diamond points) across 100 runs of 4 scenarios described in the main text, when $p = 200$ (solid lines) and $p = 1600$ (dashed lines). All standard errors are smaller than 5%.



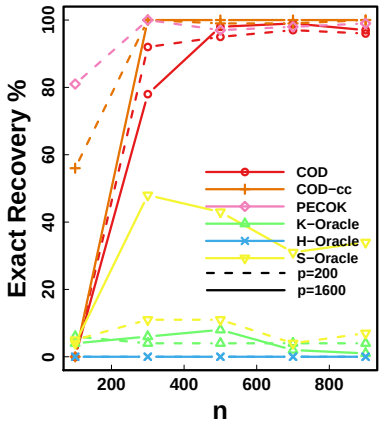
(a) M1



(b) M2

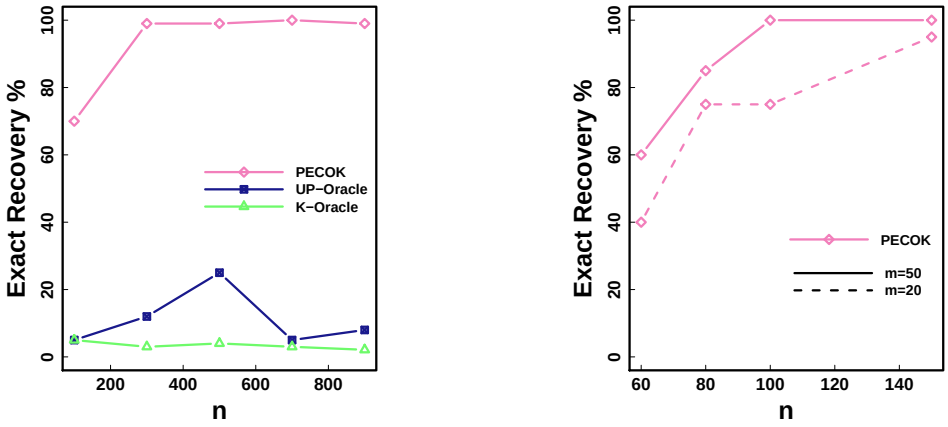


(c) M1S



(d) M1P

Figure 2.2 – Comparison of Exact recovery percentages across 100 runs. (a) The parameter K in K-means (K-Oracle, medium green lines, triangle points) and Γ-uncorrected PECOK (UP-Oracle, navy blue lines, square points) are set to the true K while PECOK (light pink lines, diamond points) selects K using our CV criterion. (b) The exact recovery percentages of PECOK are shown under $m = 50$ (solid line) and $m = 20$ (dash line). All standard errors are smaller than 5%.



(a) With/without correcting for Γ (b) Varying m

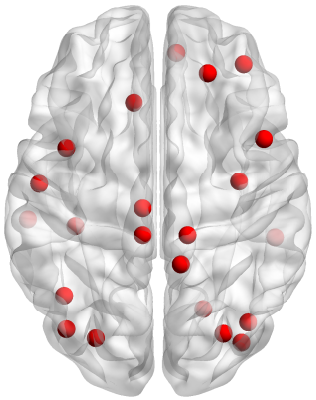
loss when $K = 142$. We thus first focus on illustrating the COD clusters here. Table 2.2 lists the largest cluster of putative areas recovered by COD and their functional classification based on prior knowledge. Most of these areas are classified to be related to visual, motor, and task functioning, which is consistent with the implication of our experimental task that requires the subject to perform motor responses based on visual stimuli. Figure 2.3(a) plots the locations of these coordinates on a standard brain template. It shows that our COD cluster appears to come mostly from approximately symmetric locations from the left and right hemisphere, though we do not enforce this brain function symmetry in our algorithm. Note that the original coordinates in Power et al., 2011 are not sampled with exact symmetry from both hemispheres of the brain, and thus we do not expect exact symmetric locations in the resulting clusters based on these coordinates.

Because there are no gold standards for partitioning the brain, we follow common practice and use a prediction criterion to further compare the clustering performance of different methods. For a fair comparison, we also estimate $\hat{G}$ using K-means, HC, and spectral clustering on the same resulting K's found by COD. The prediction criterion is as follows. We first compute the covariance matrices $\hat{S}_1$ and $\hat{S}_2$ from the first and second session data respectively. For a grouping estimate $\hat{G}$, we use the following loss to evaluate its performance

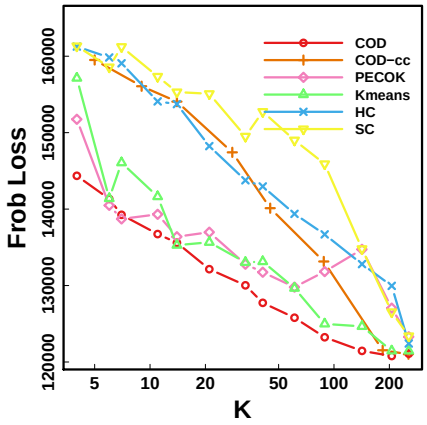
$$\|\hat{S}_2 - \Upsilon\left(\hat{S}_1, \hat{G}\right)\|_F \tag{8.1}$$

where block averaging operator $\Upsilon(R, G)$ produces a G-block structured matrix based on $\hat{G}$. For

Figure 2.3 – (a) Plot of the coordinates of the largest COD cluster overplayed over a standard brain template. The coordinates are shown as red balls. (b) Comparison of COD, COD-CC, PECOK, K-means, HC, and SC using the Frobenius prediction loss criterion (8.1) where the groups are estimated by these methods respectively.



(a) A brain cluster by COD



(b) Prediction loss

any $a \in G_k$ and $b \in G_{k'}$, the output matrix entry $[\Upsilon(R,G)]_{ab}$ is given by

$$[\Upsilon(R,G)]_{ab} = \begin{cases} |G_k|^{-1}(|G_k|-1)^{-1}\sum_{i,j\in G_k, i\neq j} R_{ij} & \text{if } a \neq b \text{ and } k = k' \\ |G_k|^{-1}|G_{k'}|^{-1}\sum_{i\in G_k, j\in G_{k'}} R_{ij} & \text{if } a \neq b \text{ and } k \neq k' \\ 1 & \text{if } a = b. \end{cases} \tag{8.2}$$

In essence, this operator smooths over the matrix entries with indices in the same group, and one may expect that such smoothing over variables in the true cluster will reduce the loss (8.1) while smoothing over different clusters will increase the loss.

Figure 2.3(b) compares the prediction loss values under different group sizes for each method. This shows that our CV approach for COD indeed selects a value $K = 142$ that is immediately next to a slightly larger one ($K = 206$), the latter having the smallest prediction loss, near the bottom plateau. However, the differences are almost negligible. This suggests that our CV criterion, which comes with theoretical guarantees, also provides good prediction performance in this real data example, while selecting a slightly smaller K, as desired, since this makes the resulting clusters easier to describe and interpret.

Regardless of the choice of K or α, Figure 2.3(b) also shows that COD almost always yields the smallest prediction loss for a wide range of K, while PECOK does slightly better when K is between 5 and 10. Though COD-CC has large losses for medium or small K, its performance is very close to the best performer COD near $K = 146$. Kmeans in this example is the closest competing method, while the other two methods (HC and SC) yield larger losses across the choices of K.

Table 2.2 – MNI coordinates (x, y, z, in mm) of the largest COD group and their functioning classification.

X	Y	Z	Function	X	Y	Z	Function
40	-72	14	visual	-7	-21	65	motor
-28	-79	19	visual	-7	-33	72	motor
20	-66	2	visual	13	-33	75	motor
29	-77	25	visual	10	-46	73	motor
37	-81	1	visual	36	-9	14	motor
47	10	33	task	-53	-10	24	motor
-41	6	33	task	-37	-29	-26	uncertain
38	43	15	task	52	-34	-27	uncertain
-41	-75	26	default	-58	-26	-15	uncertain
8	48	-15	default	-42	-60	-9	attention
22	39	39	default	-11	26	25	saliency

9 Discussion

In this section, we discuss some related models and give an overall recommendation on the usage of our methods.

9.1 Comparison with Stochastic Block Model

The problem of variable clustering that we consider in this work is fundamentally different from that of variable clustering from *network data*. The latter, especially in the context of the Stochastic Block Model (SBM), has received a large amount of attention over the past years, for instance Guédon and Vershynin, 2016; Lei and Rinaldo, 2015; Chen and Xu, 2016; Lei and Zhu, 2014; Abbe and Sandon, 2015; Mossel, Neeman, and Sly, 2014; Le, Levina, and Vershynin, 2014. The most important difference stems from the nature of the data: the data analyzed via the SBM is a $p \times p$ binary matrix $\mathbf{A}$, called the adjacency matrix, with entries assumed to have been generated as independent Bernoulli random variables; its expected value is assumed to have a block structure. In contrast, the data matrix $\mathbf{X}$ generated from a G-block covariance is a $n \times p$ matrix with real entries, and rows viewed as i.i.d copies of a p-dimensional vector X with mean zero and dependent entries. The covariance matrix Σ of X is assumed to have (up to the diagonal) a block structure.

Need for a correction. Even though the analysis of the methods in our setting would differ from the SBM setting, we could have applied available clustering procedures tailored for SBMs to the empirical covariance matrix $\widehat{\Sigma} = \mathbf{X}^t\mathbf{X}/n$ by treating it as some sort of weighted adjacency matrix. It turns out that applying verbatim the spectral clustering procedure of Lei and Rinaldo Lei and Rinaldo, 2015 or the SDP such as the ones in Amini and Levina, 2018 would lead to poor results. The main reason for this is that, in our setting, we need to **correct** both the spectral algorithm and the SDP to recover the correct clusters (Section 5). Second, the SDPs studied in the SBM context (such as those of Amini and Levina, 2018) do not handle properly groups with different and

unknown sizes, contrary to our SDP. To the best of our knowledge, our SDP (without correction) has only been independently studied by Mixon et al. Mixon, Villar, and Ward, 2016 in the context of Gaussian mixtures.

Analysis of the SDP. As for the mathematical arguments, our analysis of the SDP in our on covariance-type model differs from that in mean-type models partly because of the the presence of non-trivial cross-product terms. Instead of relying on dual certificates arguments as in other work such as Perry and Wein, 2017, we directly investigate the primal problem and combine different duality-norm bounds. The crucial step is the Lemma A.3 which allows to control the Frobenius inner product by a (unusual) combination of ℓ^1 and spectral control. In our opinion, our approach is more transparent than dual certificates techniques, especially in the presence of a correction $\widehat{\Gamma}$ and allows for the attainment of optimal convergence rates.

9.2 Extension to other Models

The general strategy of correcting a convex relaxation of K-means can be applied to other models. In Royer, 2017, one of the authors has adapted the PECOK algorithm to the clustering problem of mixture of subGaussian distributions. In particular, in the high-dimensional setting where the correction plays a key role, Royer, 2017 obtains sharper separation conditions dependencies than in state-of-the-art clustering procedures Mixon, Villar, and Ward, 2016. Extensions to model-based overlapping clustering are beyond the scope of this paper, but we refer to Bing et al., 2017 for recent results.

9.3 Practical recommendations

Based on our extensive simulation studies, we conclude this section with general recommendations on the usage of our proposed algorithms.

If p is moderate in size, and if there are reasons to believe that no singletons exist in a particular application, or if they have been removed in a pre-processing step, we recommend the usage of the PECOK algorithm, which is numerically superior to existing methods: exact recovery can be reached for relatively small sample sizes. COD is also very competitive, but requires a slightly larger sample size to reach the same performance as PECOK. The constraint on the size of p reflects the existing computational limits in state-of-the art algorithms for SDP, not the statistical capabilities of the procedure, the theoretical analysis of which being one of the foci of this work.

If p is large, we recommend COD-type algorithms. Since COD is optimization-free, it scales very well with p, and only requires a moderate sample size to reach exact cluster recovery. Moreover, COD adapts very well to data that contains singletons and, more generally, to data that is expected to have many inhomogeneous clusters.

Appendix

A Results for the PECOK estimator

In order to avoid notational clutter, we write G for G^* and m for m^* for the entirety of this section.

A.1 The motivation for a K-means correction: proof of Propositions 5.1 and 5.2

Proofs of Proposition 5.1

The basis of this proof is the following Lemma.

Lemma A.1. *The collection $\mathcal{C}$ contains only one matrix whose support is included in* $\mathrm{supp}(B^*)$*, that is*

$$\mathcal{C} \cap \left\{B,\ \mathrm{supp}(B) \subset \mathrm{supp}(B^*)\right\} = \{B^*\}\,. \tag{A.1}$$

Proof. Consider any matrix $B \in \mathcal{C}$ whose support is included in $\mathrm{supp}(B^*)$. Since $B1 = 1$, it follows that each submatrix $B_{G_kG_k}$ is symmetric doubly stochastic. Since $B_{G_kG_k}$ is also positive semidefinite, we have

$$tr(B_{G_kG_k}) \geqslant \|B_{G_kG_k}\|_{op} \geqslant 1^t B_{G_kG_k} 1/|G_k| = 1 \tag{A.2}$$

As $B \in \mathcal{C}$, we have $\mathrm{tr}(B) = K$, so all the submatrices $B_{G_kG_k}$ have a unit trace. Since $\|B_{G_kG_k}\|_{op} \geqslant 1$, this also enforces that $B_{G_kG_k}$ contains only one non-zero eigenvalue and that a corresponding eigenvector is the constant vector 1. As a consequence, $B_{G_kG_k} = 11^t/|G_k|$ for all $k = 1, \ldots, K$ and $B = B^*$. □

As a consequence of Lemma A.1, in order to prove Proposition 5.1 we only need to prove that

$$\langle \Sigma, B^* - B\rangle > 0, \ \text{ for all } B \in \mathcal{C} \text{ (resp. } \mathcal{O}) \text{ such that } \mathrm{supp}(B) \nsubseteq \mathrm{supp}(B^*). \tag{A.3}$$

We have

$$\langle \Sigma, B^* - B\rangle = \langle AC^*A^t, B^* - B\rangle + \langle \Gamma, B^* - B\rangle. \tag{A.4}$$

Define the p-dimensional vector v by $v = diag(AC^*A^t)$. Since $B1 = 1$ for all $B \in \mathcal{C}$, we have $\langle v1^t + 1v^t, B^* - B\rangle = 0$. Hence, we have

$$\begin{aligned}
\langle AC^*A^t, B^* - B\rangle &= \langle AC^*A^t - \frac{1}{2}(v1^t + 1v^t), B^* - B\rangle && \text{(A.5)}\\
&= \sum_{j,k} \sum_{a\in G_j,\, b\in G_k} \left(C^*_{jk} - \frac{C^*_{jj} + C^*_{kk}}{2}\right)(B^*_{ab} - B_{ab}) && \text{(A.6)}\\
&= \sum_{j\neq k} \sum_{a\in G_j,\, b\in G_k} \left(\frac{C^*_{jj} + C^*_{kk}}{2} - C^*_{jk}\right) B_{ab} && \text{(A.7)}\\
&= \sum_{j\neq k} \left(\frac{C^*_{jj} + C^*_{kk}}{2} - C^*_{jk}\right) |B_{G_jG_k}|_1, && \text{(A.8)}
\end{aligned}$$

where $B_{G_jG_k} = [B_{ab}]_{a\in G_j,\, b\in G_k}$. Next Lemma lower bounds $\langle \Gamma, B^* - B\rangle$ for $B \in \mathcal{O}$. It is stated below and proved at page 56.

Lemma A.2. *For any matrix B belonging to $\mathcal{O}$ and any diagonal matrix Γ,*

$$\langle \Gamma, B^* - B\rangle \geqslant -\frac{\|\Gamma\|_V}{m} \sum_{k\neq j} |B_{G_jG_k}|_1 \,. \tag{A.9}$$

Hence, combining (A.8) and Lemma A.2, we obtain

$$\langle \Sigma, B^* - B\rangle \geqslant \sum_{j\neq k} \left(\frac{C^*_{jj} + C^*_{kk}}{2} - C^*_{jk} - \frac{\|\Gamma\|_V}{m} \right) |B_{G_jG_k}|_1 \,, \tag{A.10}$$

for all $B \in \mathcal{O}$. The condition $\Delta(C^*) > \frac{2|\Gamma|_V}{m}$ enforces that if $\mathrm{supp}(B) \nsubseteq \mathrm{supp}(B^*)$ then $\langle \Sigma, B^* - B\rangle > 0$. This proves the first claim of Proposition 5.1.

To show the counterpart of this result that corresponds to replacing $\mathcal{O}$ by $\mathcal{C}$, we invoke the Lemma stated below and proved at page 57.

Lemma A.3. *For any $p \times p$ symmetric matrix S, we have for any $B \in \mathcal{C}$*

$$|\langle S, B^* - B\rangle| \leqslant 2 \left[\sum_{j\neq k} |B_{G_jG_k}|_1 \right] \left(\frac{\|S\|_{op}}{2m} + 3|B^* S|_\infty \right) \tag{A.11}$$

Define the diagonal matrix $D = (\max_a \Gamma_{aa} + \min_a \Gamma_{aa})I/2$. Since $tr(B) = tr(B^*)$, we have $\langle \Gamma, B^* - B\rangle = \langle \Gamma - D, B^* - B\rangle$. The matrix $S = \Gamma - D$ satisfies $\|S\|_{op} = \|\Gamma\|_V/2$ and $|B^* S|_\infty \leqslant \|\Gamma\|_V/(2m)$. Applying Lemma A.3 to S, we obtain

$$|\langle \Gamma, B^* - B\rangle| \leqslant \frac{7}{2m} \|\Gamma\|_V \Big[\sum_{j\neq k} |B_{G_jG_k}|_1 \Big] \,. \tag{A.12}$$

Hence, together with (A.8), we obtain

$$\langle \Sigma, B^* - B\rangle \geqslant \sum_{j\neq k} \left(\frac{C^*_{jj} + C^*_{kk}}{2} - C^*_{jk} - \frac{7\|\Gamma\|_V}{2m} \right) |B_{G_jG_k}|_1 \,, \tag{A.13}$$

for all $B \in \mathcal{C}$. The condition $\Delta(C^*) > \frac{7|\Gamma|_V}{m}$ enforces that if $\mathrm{supp}(B) \nsubseteq \mathrm{supp}(B^*)$ then $\langle \Sigma, B^* - B\rangle > 0$, which proves the second claim of Proposition 5.1. To complete the proof of Proposition 5.1 it remains to prove the two Lemmas stated above.

Proof of Lemma A.2. By definition of B^* and since $\mathrm{tr}(B) = \mathrm{tr}(B^*) = K$, we have

$$\langle \Gamma, B^* - B\rangle = \langle \Gamma - (\max_b \Gamma_{bb})I, B^* - B\rangle \tag{A.14}$$

$$= \sum_{a=1}^{p} (\Gamma_{aa} - (\max_b \Gamma_{bb})) \Big[\frac{1}{|G_{k(a)}|} - B_{aa} \Big] \tag{A.15}$$

$$\geqslant \sum_{a=1}^{p} -\|\Gamma\|_V \Big[\frac{1}{|G_{k(a)}|} - B_{aa} \Big]_+ \tag{A.16}$$

Since B belongs to $\mathcal{O}$, each row sums to one and each B_{ab} is either equal to 0 or to B_{aa}. Thus,

$$\sum_{b \notin G_{k(a)}} B_{ab} = 1 - \sum_{b \in G_{k(a)}} B_{ab} \geqslant [1 - |G_{k(a)}| B_{aa}]_+ \tag{A.17}$$

which implies

$$\Big[\frac{1}{|G_{k(a)}|} - B_{aa}\Big]_+ \leqslant \frac{1}{m} \sum_{b \notin G_{k(a)}} B_{ab}. \tag{A.18}$$

Coming back to (A.16), this gives us $\langle \Gamma, B^* - B \rangle \geqslant -\frac{\|\Gamma\|_V}{m} \sum_{k \neq j} |B_{G_j G_k}|_1$. □

Proof of Lemma A.3. Observe first that B^* is a projection matrix that induces the following decomposition of S.

$$S = B^* S + S B^* - B^* S B^* + (I - B^*) S (I - B^*). \tag{A.19}$$

By the definition of the inner product, followed by the triangle inequality, and since $(I - B^*)B^* = 0$, we further have

$$\begin{aligned} |\langle S, B^* - B \rangle| &\leqslant 3 |B^* S|_\infty |B^*(B^* - B)|_1 + |\langle (I - B^*) S (I - B^*), B^* - B \rangle| && \text{(A.20)} \\ &= 3 |B^* S|_\infty |B^*(B^* - B)|_1 + |\langle S, (I - B^*) B (I - B^*) \rangle|. && \text{(A.21)} \end{aligned}$$

Relying on the duality of the nuclear $\|\cdot\|_*$ and operator $\|\ \|_{op}$ norms, we have

$$|\langle S, (I - B^*) B (I - B^*) \rangle| \leqslant \|S\|_{op} \|(I - B^*) B (I - B^*)\|_*. \tag{A.22}$$

We begin by bounding the nuclear norm $\|(I - B^*) B (I - B^*)\|_*$. Since $(I - B^*) B (I - B^*) \in \mathcal{S}^+$, we have

$$\begin{aligned} \|(I - B^*) B (I - B^*)\|_* &= \operatorname{tr}((I - B^*) B (I - B^*)) && \text{(A.23)} \\ &= \langle I - B^*, B (I - B^*) \rangle && \text{(A.24)} \\ &= \langle I - B^*, B \rangle. && \text{(A.25)} \end{aligned}$$

Using the fact that the sum of each row of B is 1 and $\operatorname{tr}(B) = K$, we have

$$\begin{aligned} \|(I - B^*) B (I - B^*)\|_* = \langle I - B^*, B \rangle &= \operatorname{tr}(B) - \sum_{k=1}^{K} \sum_{a,b \in G_k} \frac{B_{ab}}{|G_k|} && \text{(A.26)} \\ &= K - K + \sum_{k \neq j} \sum_{a \in G_k,\ b \in G_j} \frac{B_{ab}}{|G_k|} && \text{(A.27)} \\ &\leqslant \frac{1}{m} \sum_{k \neq j} |B_{G_j G_k}|_1 . && \text{(A.28)} \end{aligned}$$

Next, we simplify the expression of $|B^*(B^* - B)|_1 = |B^*(I - B)|_1$.

$$|B^*(I-B)|_1 = \sum_{j\neq k} \sum_{a\in G_j,\ b\in G_k} |(B^*B)_{ab}| + \sum_{k=1}^{K} \sum_{a,b\in G_k} |[B^*(I-B)]_{ab}| \tag{A.29}$$

$$= \sum_{j\neq k} \sum_{a\in G_j,\ b\in G_k} \frac{1}{|G_j|} \sum_{c\in G_j} B_{cb} + \sum_{k=1}^{K} \sum_{a,b\in G_k} \frac{1}{|G_k|} \Big|1 - \sum_{c\in G_k} B_{cb}\Big| \tag{A.30}$$

$$= 2\sum_{j\neq k} |B_{G_jG_k}|_1 \ , \tag{A.31}$$

where we used again $B1 = 1$ and that the entries of B are nonnegative. Gathering the above bounds together with (A.21) yields the desired result. This completes the proof of this result and of Proposition 5.1. □

Proof of Proposition 5.2

By symmetry, we can assume that the true partition matrix B^* is diagonal block constant. Define the partition matrix $B_1 := \begin{bmatrix} 2/m & 0 & 0 \\ 0 & 2/m & 0 \\ 0 & 0 & 1/(2m) \end{bmatrix}$ where the first two blocks are of size $m/2$ and the the last block has size $2m$. The construction of the matrix B_1 amounts to merging groups G_2 and G_3, and to splitting G_1 into two groups of equal size. Then,

$$\langle \Sigma, B^* \rangle = \gamma_+ + 2\gamma_- + mtr(C^*), \quad \langle \Sigma, B_1 \rangle = 2\gamma_+ + \gamma_- + mtr(C^*) - m\tau. \tag{A.32}$$

As a consequence, $\langle \Sigma, B_1 \rangle < \langle \Sigma, B^* \rangle$ if and only if $\tau > \frac{\gamma_+ - \gamma_-}{m}$.

A.2 Analysis of the population version under the approximate model: proofs of Proposition 6.2 and Corollary 6.1

In this subsection, we prove Proposition 6.2 and Corollary 6.1. As a consequence of Lemma A.1 above, we only need to prove that,

$$\langle \Sigma, B^* - B \rangle > 0, \ \text{ for all } B \in \mathcal{C} \text{ such that } \ \mathrm{supp}(B) \nsubseteq \mathrm{supp}(B^*). \tag{A.33}$$

We have

$$\langle \Sigma, B^* - B \rangle = \langle AC^*A^t, B^* - B \rangle + \langle D, B^* - B \rangle + \langle R, B^* - B \rangle. \tag{A.34}$$

Define the p-dimensional vector v by $v = diag(AC^*A^t)$. Since $B1 = 1$ for all $B \in \mathcal{C}$, we have $\langle v1^t + 1v^t, B^* - B\rangle = 0$. Hence, we have

$$\langle AC^*A^t, B^* - B\rangle = \langle AC^*A^t - \frac{1}{2}(v1^t + 1v^t), B^* - B\rangle \tag{A.35}$$

$$= \sum_{j,k} \sum_{a\in G_j,\, b\in G_k} \left(C^*_{jk} - \frac{C^*_{jj} + C^*_{kk}}{2}\right)(B^*_{ab} - B_{ab}) \tag{A.36}$$

$$= \sum_{j\neq k} \sum_{a\in G_j,\, b\in G_k} \left(\frac{C^*_{jj} + C^*_{kk}}{2} - C^*_{jk}\right) B_{ab} \tag{A.37}$$

$$= \sum_{j\neq k} \left(\frac{C^*_{jj} + C^*_{kk}}{2} - C^*_{jk}\right) |B_{G_jG_k}|_1, \tag{A.38}$$

where $B_{G_jG_k} = [B_{ab}]_{a\in G_j,\, b\in G_k}$. From Lemma A.3 we get

$$|\langle D, B^* - B\rangle| \leqslant \frac{7}{2m}|D|_V \sum_{j\neq k} |B_{G_jG_k}|_1 \tag{A.39}$$

and

$$|\langle R, B^* - B\rangle| \leqslant \left(\frac{3}{2}|R|_\infty + \frac{\|R\|_{op}}{m}\right) \sum_{j\neq k} |B_{G_jG_k}|_1. \tag{A.40}$$

Combining the two last inequalities with (A.38) gives the Proposition. The Corollary follows.

A.3 Exact recovery with PECOK: approximate model. Proofs of Theorems 6.2, 5.1 and 5.2

The conclusion of Theorems 6.2 and 5.1 follows by combining the conclusion of Theorems A.1 and A.2, stated below and proved in the next subsection, with the conclusion of Proposition A.1 stated and proved in Appendix A.4 below.

Specifically: Theorem A.1, specialized to $R = 0$ and coupled with (i) of Proposition A.1 proves Theorem 5.1. Theorem A.2, specialized to $R = 0$ and coupled with (ii) of Proposition A.1 proves Theorem 5.2. Finally, Theorem A.1 coupled with (i) of Proposition A.1 proves Theorem 6.2. If we combine Theorem A.1 with (ii) of Proposition A.1 we obtained a version of Theorem 6.2 for bounded variables, which we do not state, for space limitations.

The following theorem examines the behavior of PECOK under the general model (1.4).

Theorem A.1. *There exist $c_1, \ldots, c_3$ three positive constants such that the following holds. Let $\widehat{\Gamma}$ be any estimator of D, such that $|\widehat{\Gamma} - D|_V \leqslant \delta_{n,p}$ with probability $1 - c_1/(2p)$. Then, under Assumption 1, and when $L^4 \log(p) \leqslant c_3 n$ and*

$$\Delta(C^*) \geqslant c_L \Big[\|\Gamma\|_{op} \left\{\sqrt{\frac{\log p}{mn}} + \sqrt{\frac{p}{nm^2}} + \frac{\log(p)}{n} + \frac{p}{nm}\right\} + \frac{\delta_{n,p} + \|R\|_{op}}{m} + |R|_\infty\Big], \tag{A.41}$$

we have $\widehat{B} = B^*$, with probability higher than $1 - c_1/p$.

The following theorem examines the behavior of PECOK under the general model (1.4) and when the variables are bounded.

Theorem A.2. *There exist c_1, c_2 two positive constants such that the following holds. Let $\widehat{\Gamma}$ be any estimator of D, such that $|\widehat{\Gamma} - D|_V \leqslant \delta_{n,p}$ with probability $1 - c_1/(2p)$. Then, under Assumption 1-bis, and when*

$$\Delta(C^*) \geqslant \qquad c_2 \left[M\|\Gamma\|_{op}^{1/2} \sqrt{\frac{p\log(p)}{nm^2}} + M^2 \frac{p\log(p)}{nm} + \frac{\delta_{n,p} + \|R\|_{op}}{m} + |R|_\infty \right], \tag{A.42}$$

we have $\widehat{B} = B^*$, with probability higher than $1 - c_1/p$.

Proofs of Theorems A.1 and A.2

In contrast to other SDP analyses performed for other models Mixon, Villar, and Ward, 2016, our proof does not rely on dual certificates techniques. Instead of that, we directly investigate the primal problem and combine different duality-norm bounds. In our opinion, this makes the arguments more transparent. The two key ingredients are Lemmas A.1 and A.3 above.

Given $k, l \in [K]$, we define $\Delta_{kl}(C^*) = C^*_{kk} + C^*_{ll} - 2C^*_{kl}$. As a consequence of Lemma A.1 page 55, we only need to prove that

$$\langle \widehat{\Sigma} - \widehat{\Gamma}, B^* - B \rangle > 0, \quad \text{for all } B \in \mathcal{C} \text{ such that } \operatorname{supp}(B) \nsubseteq \operatorname{supp}(B^*), \tag{A.43}$$

with high probability.

We begin by introducing some notation. For any $k, l \in [K]$, we denote $m_k = |G_k|$ the size of group G_k and

$$\gamma_{kl} = \frac{1}{m_k m_l} \sum_{a \in G_k, b \in G_l} \Gamma_{ab} \,. \tag{A.44}$$

Recall that $\mathbf{X}$ denotes the $n \times p$ matrix of observations and we set $\underline{\mathbf{Z}} = \mathbf{X}A^t(A^tA)^{-1}$. We have the decomposition

$$\mathbf{X} = \mathbf{X}B^* + \mathbf{X}(I - B^*) =: \underline{\mathbf{Z}}A^t + \underline{\mathbf{E}} \tag{A.45}$$

with $\operatorname{Cov}(A\underline{Z}, \underline{E}) = B^*\Gamma(I - B^*)$, $\operatorname{Cov}(\underline{E}) = (I - B^*)\Gamma(I - B^*)$, and $\operatorname{Cov}(\underline{Z}_k, \underline{Z}_l) = C^*_{kl} + \gamma_{kl}$. Note that in the latent model $X_a = Z_{k(a)} + E_a$, the random variables $\underline{Z}_k$ and $\underline{E}_a$ differ from Z_k and E_a.

Our first goal is to decompose $\widehat{\Sigma} - \widehat{\Gamma}$ in such a way that the distance $|\underline{\mathbf{Z}}_{:k} - \underline{\mathbf{Z}}_{:j}|_2^2$ becomes evident. To this end, recall that $n\widehat{\Sigma} = \mathbf{X}^t\mathbf{X}$ and let us define $\widetilde{\Gamma} = \frac{1}{n}\underline{\mathbf{E}}^{\mathbf{t}}\underline{\mathbf{E}}$. Hence, we have

$$n\widehat{\Sigma} = A\underline{\mathbf{Z}}^t\underline{\mathbf{Z}}A^t + n\widetilde{\Gamma} + A(\underline{\mathbf{Z}}^t\underline{\mathbf{E}}) + (\underline{\mathbf{E}}^t\underline{\mathbf{Z}})A^t. \tag{A.46}$$

Using the fact that for any vectors v_1 and v_2 we have $|v_1 - v_2|_2^2 = |v_1|_2^2 + |v_2|^2 - 2v_1^t v_2$, we can write

$$[A\underline{\mathbf{Z}}^t\underline{\mathbf{Z}}A^t]_{ab} = \frac{1}{2}|[A\underline{\mathbf{Z}}^t]_{a:}|_2^2 + \frac{1}{2}|[A\underline{\mathbf{Z}}^t]_{b:}|_2^2 - \frac{1}{2}|[A\underline{\mathbf{Z}}^t]_{a:} - [A\underline{\mathbf{Z}}^t]_{b:}|_2^2, \tag{A.47}$$

for any $1 \leqslant a, b \leqslant p$. We also observe that

$$[A(\underline{\mathbf{Z}}^t\underline{\mathbf{E}}) + (\underline{\mathbf{E}}^t\underline{\mathbf{Z}})A^t]_{ab} = [(A\underline{\mathbf{Z}}^t)_{a:} - (A\underline{\mathbf{Z}}^t)_{b:}][\underline{\mathbf{E}}_{b:} - \underline{\mathbf{E}}_{a:}] + [A\underline{\mathbf{Z}}^t\underline{\mathbf{E}}]_{aa} + [A\underline{\mathbf{Z}}^t\underline{\mathbf{E}}]_{bb}. \tag{A.48}$$

Define the $p \times p$ matrix W by

$$W_{ab} := n(\widehat{\Sigma}_{ab} - \widehat{\Gamma}_{ab}) - \frac{1}{2}|[A\underline{\mathbf{Z}}^t]_{a:}|_2^2 - \frac{1}{2}|[A\underline{\mathbf{Z}}^t]_{b:}|_2^2 - [A\underline{\mathbf{Z}}^t\underline{\mathbf{E}}]_{aa} - [A\underline{\mathbf{Z}}^t\underline{\mathbf{E}}]_{bb}. \tag{A.49}$$

Combining the four displays above we have

$$W = W_1 + W_2 + W_3 + n(\Gamma - \widehat{\Gamma}), \tag{A.50}$$

with

$$(W_1)_{ab} := -\frac{1}{2}|[A\underline{\mathbf{Z}}^t]_{a:} - [A\underline{\mathbf{Z}}^t]_{b:}|_2^2 - nB^*\Gamma B^*, \tag{A.51}$$

$$(W_2)_{ab} := [(A\underline{\mathbf{Z}}^t)_{a:} - (A\underline{\mathbf{Z}}^t)_{b:}][\underline{\mathbf{E}}_{b:} - \underline{\mathbf{E}}_{a:}] - n[B^*\Gamma(I - B^*) + (I - B^*)\Gamma B^*]_{ab}, \tag{A.52}$$

and

$$W_3 = n\widetilde{\Gamma} - n(I - B^*)\Gamma(I - B^*), \tag{A.53}$$

for any $1 \leqslant a, b \leqslant p$. Observe from (A.49) that $W - n(\widehat{\Sigma} - \widehat{\Gamma})$ is a sum of four matrices, two of which are of the type $1v_1^t$, and two of the type $v_2 1^t$, for some vectors $v_1, v_2 \in \mathcal{R}^p$. Since for any two matrices B_1 and B_2 in $\mathcal{C}$, we have $B_1 1 = B_2 1 = 1$, it follows that

$$\langle W - n(\widehat{\Sigma} - \widehat{\Gamma}), B_1 - B_2\rangle = 0\,. \tag{A.54}$$

As a consequence and using the decomposition (A.50), proving (A.43) reduces to proving

$$\langle W_1 + W_2 + W_3 + n(\Gamma - \widehat{\Gamma}), B^* - B\rangle > 0, \forall B \in \mathcal{C} \text{ s.t. } \mathrm{supp}(B) \nsubseteq \mathrm{supp}(B^*). \tag{A.55}$$

We will analyze the inner product between $B^* - B$ and each of the four matrices in (A.55) separately in the following Lemmas. Their proofs are given after the proof of this theorem.

The matrix W_1 contains the information about the clusters, as we explain below. Note that for two variables a and b belonging to the same group G_k, $(W_1)_{ab} = -n\gamma_{kk}$. As a consequence, $\langle W_1, B^*\rangle = -n\sum_k m_k\gamma_{kk}$. For two variables a and b belonging to two different groups G_j and G_k, $(W_1)_{ab} = -|\underline{\mathbf{Z}}_{:i} - \underline{\mathbf{Z}}_{:k}|_2^2/2 - n\gamma_{kj}$. In the sequel, we denote by B_{G_j,G_k} the submatrix $(B_{ab})_{a\in G_j,\, b\in G_k}$. Since all the entries of B are nonnegative, and $B1 = 1$,

$$-\langle W_1, B\rangle = \frac{1}{2}\sum_{j\neq k}|\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2|B_{G_jG_k}|_1 + \sum_{j,k} n\gamma_{jk}|B_{G_jG_k}|_1 \tag{A.56}$$

$$= \frac{1}{2}\sum_{j\neq k}\big[|\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2 + 2n\gamma_{jk}\big]|B_{G_jG_k}|_1 + n\sum_k \gamma_{kk}\Big(m_k - \sum_{j:j\neq k}|B_{G_jG_k}|_1\Big) \tag{A.57}$$

$$= \frac{1}{2}\sum_{j\neq k}\Big(|\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2 - n\gamma_{kk} - n\gamma_{jj} + 2n\gamma_{jk}\Big)|B_{G_jG_k}|_1 + n\sum_k m_k\gamma_{kk}. \tag{A.58}$$

Hence, we obtain

$$\langle W_1, B^* - B\rangle = \frac{1}{2}\sum_{j\neq k}\Big[|\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2 - n\gamma_{jj} - n\gamma_{kk} + 2n\gamma_{jk}\Big]|B_{G_jG_k}|_1. \tag{A.59}$$

Each of the random variables $|\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2$ is a quadratic form of independent random variables. As a consequence, we can apply Hanson-Wright inequalities, of the type stated in Lemma A.1 to simultaneously control all these quantities. This leads us to Lemmas A.4, A.5, A.6 and A.7, proved in Section E.

Lemma A.4. *Under either Assumption 1 and Condition* (A.41) *or Assumption 1-bis and Condition* (A.42)*, it holds with probability higher than* $1 - 1/p$*, that*

$$\langle W_1, B^* - B\rangle \geqslant \sum_{j\neq k} n\frac{\Delta_{jk}(C^*)}{4}|B_{G_jG_k}|_1\ , \tag{A.60}$$

simultaneously for all matrices $B \in \mathcal{C}$*.*

We will analyze below the three remaining cross products.

Lemma A.5. *Under Assumption 1, there exists an event of probability larger than* $1 - 2/p$ *such that the following holds simultaneously for all* $B \in \mathcal{C}$

$$|\langle W_2, B^* - B\rangle| \leqslant c_1 L^2 \sum_{j\neq k}\left[\sqrt{\Delta_{jk}(C^*)|\Gamma|_\infty} + \frac{|D|_\infty}{\sqrt{m}} + |R|_\infty + |R|_\infty^{1/2}|D|_\infty^{1/2}\right] \times \left[\sqrt{n\log(p)} + \log(p)\right]|B_{G_jG_k}|_1\ , \tag{A.61}$$

Under Assumption 1-bis, there exists an event of probability larger than $1 - 2/p$ *such that the following holds simultaneously for all* $B \in \mathcal{C}$

$$|\langle W_2, B^* - B\rangle| \leqslant c_1' M \sum_{j\neq k}\Big[\sqrt{n\log(p)\big[\Delta_{k(a)k(b)}(C^*) + \frac{|D|_\infty}{m} + |R|_\infty\big]} + M\log(p)\Big]|B_{G_jG_k}|_1\ . \tag{A.62}$$

It remains to control the term W_3 corresponding to the empirical covariance matrix of the noise $\underline{\mathbf{E}}$. This is the main technical difficulty in this proof.

Lemma A.6. *Under Assumption 1, it holds with probability higher than* $1 - 1/p$ *that*

$$|\langle W_3, B^* - B\rangle| \leqslant c_L\|\Gamma\|_{op}\left(\sqrt{\frac{np}{m^2}} + \frac{p}{m}\right)\sum_{j\neq k}|B_{G_jG_k}|_1\ , \tag{A.63}$$

simultaneously over all matrices $B \in \mathcal{C}$*. Here,* c_L *is a constant that only depends on* $L > 0$*.*

Under Assumption 1-bis, it holds with probability higher than $1 - 1/p$ *that*

$$|\langle W_3, B^* - B\rangle| \leqslant c_2 M\Big[\sqrt{\frac{np\|\Gamma\|_{op}\log(p)}{m^2}} + \frac{pM\log(p)}{m}\Big]\sum_{j\neq k}|B_{G_jG_k}|_1, \tag{A.64}$$

simultaneously over all matrices $B \in \mathcal{C}$*.*

Finally, we control the last term $\langle n(\Gamma - \widehat{\Gamma}), B^* - B\rangle$ with the next Lemma.

Lemma A.7. *It holds that*

$$|\langle \Gamma - \widehat{\Gamma}, B^* - B\rangle| \leqslant c_3\Big[\frac{|D - \widehat{\Gamma}|_V + \|R\|_{op}}{m} + |R|_\infty\Big] \sum_{j\neq k} |B_{G_jG_k}|_1 \tag{A.65}$$

simultaneously over all matrices $B \in \mathcal{C}$.

End of the proof of Theorem A.1. Under Assumption 1, we combine (A.60), (A.61), (A.63) and (A.65) and the assumption $L^4 \log(p) \leqslant c_4 n$ and obtain that, with probability larger than $1 - c/p$,

$$\frac{1}{n}\langle W, B^* - B\rangle \geqslant \sum_{j\neq k} \Bigg[\frac{\Delta_{jk}(C^*)}{4} - c_1 L^2 \sqrt{\Delta_{jk}(C^*)|\Gamma|_\infty \frac{\log(p)}{n}} \tag{A.66}$$

$$- c_2 L^2 \Big[\frac{|D|_\infty}{\sqrt{m}} + |R|_\infty + |R|_\infty^{1/2}|D|_\infty^{1/2}\Big]\sqrt{\frac{\log(p)}{n}} \tag{A.67}$$

$$- c_L \|\Gamma\|_{op}\Big[\sqrt{\frac{p}{nm^2}} + \frac{p}{nm}\Big] \tag{A.68}$$

$$- c_3\Big[\frac{|D - \widehat{\Gamma}|_V + \|R\|_{op}}{m} + |R|_\infty\Big]\Bigg] |B_{G_jG_k}|_1, \tag{A.69}$$

simultaneously for all $B \in \mathcal{C}$. Condition (A.41) enforces that, for each (j, k), the term in the bracket of (A.69) is positive. Hence, with probability at least $1 - c_1/p$, the Inequality (A.55) holds since any matrix $B \in \mathcal{C}$ whose support is not included in $\mathrm{supp}(B^*)$ satisfies $|B_{G_jG_k}|_1 > 0$ for some $j \neq k$. □

End of the proof of Theorem A.2. Let us now assume that Assumption 1-bis holds. Combining (A.59), (A.62), (A.64) and (A.65) we obtain that, with probability larger than $1 - c/p$,

$$\frac{1}{n}\langle W, B^* - B\rangle \geqslant \sum_{j\neq k} \Bigg[\frac{\Delta_{jk}(C^*)}{4} - c_1 M \sqrt{\Delta_{jk}(C^*)\frac{\log(p)}{n}} \tag{A.70}$$

$$- c_2 M \sqrt{\frac{|D|_\infty}{m} + |R|_\infty}\sqrt{\frac{\log(p)}{n}} \tag{A.71}$$

$$- c_3 M^2 \frac{p\log(p)}{nm} - c_4 M \|\Gamma\|_{op}^{1/2}\sqrt{\frac{p\log(p)}{nm^2}} \tag{A.72}$$

$$- c_5\Big[\frac{|D - \widehat{\Gamma}|_V + \|R\|_{op}}{m} + |R|_\infty\Big]\Bigg] |B_{G_jG_k}|_1, \tag{A.73}$$

simultaneously for all $B \in \mathcal{C}$. Condition (A.42) enforces that, for each (j, k), the term in the bracket of (A.73) is positive and as previously that (A.55) holds with probability at least $1 - c_3/p$. □

A.4 Guarantees for the estimator (5.10) of Γ

Proposition A.1 controls the estimation error of estimator $\widehat{\Gamma}$ defined by (5.10) under both the exact model and the approximate block G-block model (1.4). We set $v^2 = \min_{c\neq d} \mathrm{Var}\,(X_c - X_d)$.

Proposition A.1. *Assume that Γ either*

(a) is diagonal;

(b) or fulfills the diagonal dominance assumption (6.5).

Assume also that $\Delta(C^*) \geqslant 0$. Then, the two following results holds.

(i) *Under Assumption 1, there exist three numerical constants c_1–c_3 such that when $m \geqslant 3$ and $L^4 \log(p) \leqslant c_1 n$, with probability larger than $1 - c_3/p$, the estimator $\widehat{\Gamma}$ defined by (5.10) satisfies*

$$|\widehat{\Gamma} - \Gamma|_V \leqslant 2|\widehat{\Gamma} - \Gamma|_\infty \leqslant c_2 \left(\sqrt{|R|_\infty |\Gamma|_\infty} + |\Gamma|_\infty L^2 \sqrt{\frac{\log(p)}{n}} \right) . \tag{A.74}$$

(ii) *Under Assumption 1-bis, there exist three numerical constants c_1–c_3 such that when $m \geqslant 3$ and $\log(p) \leqslant c_1 (v/M)^2 n$, with probability larger than $1 - c_3/p$, the estimator $\widehat{\Gamma}$ defined by (5.10) satisfies*

$$|\widehat{\Gamma} - \Gamma|_V \leqslant 2|\widehat{\Gamma} - \Gamma|_\infty \leqslant c_2 \left(\sqrt{|R|_\infty |\Gamma|_\infty} + M \sqrt{\frac{|\Gamma|_\infty \log(p)}{n}} + M^2 \frac{\log(p)}{n} \right) . \tag{A.75}$$

Proof of Proposition A.1. To ease the presentation of this proof, we introduce the new notation

$$ne_1(a) := \operatorname*{argmin}_{b \in [p] \setminus \{a\}} V(a,b) \quad \text{and} \quad ne_2(a) := \operatorname*{argmin}_{b \in [p] \setminus \{a, ne_1(a)\}} V(a,b). \tag{A.76}$$

(i) We start with the first part of the proposition. Let a, b_1, b_2 be three different indices. Under Assumption 1, the Corollary G.1 of Hanson-Wright inequality gives that with probability at least $1 - p^{-4}$

$$\begin{aligned} &\Big| \frac{1}{n} \langle \mathbf{X}_{:a} - \mathbf{X}_{:b_1}, \mathbf{X}_{:a} - \mathbf{X}_{:b_2} \rangle - \operatorname{Cov}\left(X_a - X_{b_1}, X_a - X_{b_2}\right) \Big| \\ &\qquad \leqslant c L^2 \sqrt{\operatorname{Var}(X_a - X_{b_1}) \operatorname{Var}(X_a - X_{b_2})} \left(\sqrt{\frac{\log(p)}{n}} + \frac{\log(p)}{n} \right). \end{aligned} \tag{A.77}$$

Applying the inequality $2ab \leqslant a^2 + b^2$, and a union bound, we obtain that the inequalities

$$\begin{aligned} &\Big| \frac{1}{n} \langle \mathbf{X}_{:a} - \mathbf{X}_{:b_1}, \mathbf{X}_{:a} - \mathbf{X}_{:b_2} \rangle - \operatorname{Cov}\left(X_a - X_{b_1}, X_a - X_{b_2}\right) \Big| \\ &\qquad \leqslant c' L^2 \big[\operatorname{Var}\left(X_a - X_{b_1}\right) + \operatorname{Var}\left(X_a - X_{b_2}\right) \big] \left(\sqrt{\frac{\log(p)}{n}} + \frac{\log(p)}{n} \right) \end{aligned} \tag{A.78}$$

hold simultaneously over all triplets of different indices a, b_1, b_2, with probability $1 - 1/p$. Decomposing these variance and covariance terms, we obtain

$$\begin{aligned} \operatorname{Cov}(X_a - X_{b_1}, X_a - X_{b_2}) &= D_{aa} + R_{b_1 b_2} - R_{a b_1} - R_{a b_2} \\ &\quad + \frac{1}{2} \left(\Delta_{k(a)k(b_1)}(C^*) + \Delta_{k(a)k(b_2)}(C^*) - \Delta_{k(b_2)k(b_1)}(C^*) \right), \end{aligned} \tag{A.79}$$

and $\mathrm{Var}\,(X_a - X_b) = D_{aa} + D_{bb} + \Delta_{k(a)k(b)}(C^*) - 2R_{ab}$. Hence

$$\mathrm{Var}\,(X_a - X_{b_i}) \leqslant 2|D|_\infty + 2|R|_\infty + |\Delta_{k(a)k(b_i)}(C^*)|, \tag{A.80}$$

$$\Big|\mathrm{Cov}\,(X_a - X_{b_1}, X_a - X_{b_2}) - D_{aa}\Big| \leqslant 3|R|_\infty + \frac{|\Delta_{k(a)k(b_1)}(C^*)| + |\Delta_{k(a)k(b_2)}(C^*)| + |\Delta_{k(b_1)k(b_2)}(C^*)|}{2}. \tag{A.81}$$

For $i = 1, 2$, write $t_i := |\Delta_{k(a)k(ne_i(a))}(C^*)|$ and $t_{12} = |\Delta_{k(ne_1(a))k(ne_2(a))}(C^*)|$. Since $\log(p) \leqslant c_1 L^{-4} n$, the previous inequalities entail that

$$\Big|\widehat{\Gamma}_{aa} - D_{aa}\Big| \leqslant c\left(|R|_\infty + (t_1 + t_2 + t_{12}) + L^2|\Gamma|_\infty\sqrt{\frac{\log(p)}{n}}\right) \tag{A.82}$$

with probability at least $1 - 1/p$. As a consequence, we only have to prove that t_1, t_2 and t_{12} are smaller than $c\left(\sqrt{|R|_\infty|\Gamma|_\infty} + L^2|\Gamma|_\infty\sqrt{\log(p)/n}\right)$ with probability at least $1 - c/p$. We focus on t_1, the arguments for t_2 and t_{12} being similar.

First note that $t_1 = 0$ if $k(a) = k(ne_1(a))$ so that we may assume henceforth that $k(a) \neq k(ne_1(a))$. We have the following.

Lemma A.8. *Assume that Γ either*

(a) is diagonal;

(b) or fulfills the diagonal dominance assumption (6.5).

Assume also that $\Delta(C^*) \geqslant 0$. Then, there exists a numerical constant c_1, such that, outside an event of probability less than $1/p^2$, we have

(i) *under Assumption 1,*

$$\left|\langle \mathbf{X}_{:a} - \mathbf{X}_{:ne_1(a)}, \mathbf{X}_{:c} - \mathbf{X}_{:d}\rangle\right| \leqslant c_1\left(\sqrt{n|R|_\infty} + L^2|\Gamma|_\infty^{1/2}\sqrt{\log(p)}\right)|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2\,, \tag{A.83}$$

simultaneously over all $c, d \neq (a, ne_1(a))$;

(ii) *under Assumption 1-bis,*

$$\left|\langle \mathbf{X}_{:a} - \mathbf{X}_{:ne_1(a)}, \mathbf{X}_{:c} - \mathbf{X}_{:d}\rangle\right| \leqslant c_1\left(\sqrt{n|R|_\infty} + M\sqrt{\log(p)}\right)|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2\,, \tag{A.84}$$

simultaneously over all $c, d \neq (a, ne_1(a))$.

Similar bounds also hold for $ne_2(a)$ instead of $ne_1(a)$.

This Lemma is proved in Section E.2. Below, c' denotes a numerical constant, whose value may vary from line to line.

For any c and d, the variance of $X_c - X_d$ is less than $\Delta_{k(c)k(d)}(C^*) + 2|D|_\infty + 2|R|_\infty$. As a consequence, Hanson-Wright inequality together with an union bound over all $c, d \in [p]$ and the condition $L^4\log(p) \leqslant c_1 n$ leads to

$$|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2 \leqslant c'\sqrt{n[4|\Gamma|_\infty + \Delta_{k(c)k(d)}(C^*)]}\,, \tag{A.85}$$

simultaneously over all $c \neq d$, with probability $1 - 1/p^2$. Take c and d any two indices such that $k(a) = k(c)$ and $k(ne_1(a)) = k(d)$. So combined with Lemma A.8, we get with probability at least $1 - 2p^{-2}$

$$\left|\langle \mathbf{X}_{:a} - \mathbf{X}_{:ne_1(a)}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle\right| \leqslant c' \left(\sqrt{n|R|_\infty} + L^2 |\Gamma|_\infty^{1/2} \sqrt{\log(p)}\right) \sqrt{n[|\Gamma|_\infty + t_1]}. \tag{A.86}$$

Let us now lower bound the left hand-side of the above inequality. For any c in the same group as a and b in the same group as d, we have

$$\mathbb{E}\left[\langle X_a - X_b, X_c - X_d \rangle\right] = \Delta_{k(b)k(a)}(C^*) + R_{ac} + R_{bd} - R_{bc} - R_{ab}. \tag{A.87}$$

Therefore, Corollary G.1 of the Hanson-Wright inequality yields, with probability at least $1 - p^4$,

$$\begin{aligned}\left|\langle \mathbf{X}_{:a} - \mathbf{X}_{:b}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle\right| \geqslant & n\Delta_{k(b)k(a)}(C^*) - 4n|R|_\infty \\ & - c' L^2 \sqrt{n \log(p)} [\Delta_{k(b)k(a)}(C^*) + |\Gamma|_\infty].\end{aligned} \tag{A.88}$$

As a consequence, for any c and d such that $k(a) = k(c)$ and $k(ne_1(a)) = k(d)$, we get from $L^4 \log(p) \leqslant c_1 n$ and a union bound

$$\left|\langle \mathbf{X}_{:a} - \mathbf{X}_{:ne_1(a)}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle\right| \geqslant n t_1/2 - 4n|R|_\infty - c' L^2 |\Gamma|_\infty \sqrt{n \log(p)} \tag{A.89}$$

with probability $1 - 1/p^2$. Gathering the previous bound with (A.86), Condition (A.83), Assumption (6.5) and $L^4 \log(p) \leqslant n$, we conclude that

$$t_1 \leqslant c' \left[|R|_\infty^{1/2} |\Gamma|_\infty^{1/2} + L^2 |\Gamma|_\infty \sqrt{\frac{\log(p)}{n}}\right] \tag{A.90}$$

simultaneously for all a, with probability $1 - c_3/p$. Together with (A.82), this concludes the proof of the first part. □

B Proof of results concerning model Identifiability

B.1 Proofs of Sections 2 page 32 and 6.1 page 44

Proof of Proposition 2.1. We first observe that if G is such that $\Sigma = ACA^t + \Gamma$ holds, with A associated to G, then G is a sub-partition of G^*. To see this, note that for any a, b belonging to the same group of the partition G, we have $COD(a, b) = 0$. Hence a, b are in a same group of the partition G^*, therefore G is a sub-partition of G^*. Furthermore, for any sub-partition G of G^*, $MCOD(\Sigma, G) = 0$ unless $G = G^*$. Thus, $G = G^*$. □

Proof of Corollary 2.1. We start with the following Lemma that proves the inequalities in display (1.8).

Lemma B.1. *Assume that G is a partition, such that $\Sigma = ACA^t + \Gamma$ with Γ diagonal and A associated to G. Then, we have:*

(a) $2\lambda_K(C) \leqslant \Delta(C)$

(b) if the size m of the smallest group of G is larger than one, then

$$\Delta(C) \leqslant 2\mathrm{MCOD}(\Sigma, G). \tag{B.1}$$

(c) if C is semi-positive definite, then

$$\mathrm{MCOD}(\Sigma, G) \leqslant \sqrt{\Delta(C)} \max_{k=1,\dots,K} \sqrt{C_{kk}}. \tag{B.2}$$

The proof of Corollary 1 relies on the inequality (B.1) which holds since $m \geqslant 2$. The result then follows from Proposition 2.1. □

Proof of Lemma B.1. The first inequality holds because

$$\Delta(C) = \min_{j<k} (e_j - e_k)^t C (e_j - e_k) \geqslant 2\lambda_K(C). \tag{B.3}$$

Let us prove the second inequality. For any $k \neq j$, and for any pair of $a, a' \in G_k$ and $b, b' \in G_j$ we have

$$C_{kk} + C_{jj} - 2C_{jk} = \Sigma_{aa'} - \Sigma_{ba'} + \Sigma_{bb'} - \Sigma_{ab'} \leqslant 2COD(a, b). \tag{B.4}$$

Since the inequalities are valid for any a, b in distincts groups, the bound $\Delta(C) \leqslant 2\mathrm{MCOD}(\Sigma, G)$ follows.

For the last bound, we observe that for any $a \in G_k$, $b \in G_j$ and $c \in G_\ell$, we have

$$|\Sigma_{ac} - \Sigma_{bc}| = |C_{k\ell} - C_{j\ell}| = (e_k - e_j)^t C e_\ell \leqslant \sqrt{(e_k - e_j)^t C (e_k - e_j)} \sqrt{e_\ell^t C e_\ell}, \tag{B.5}$$

where the last inequality holds since C is positive semi-definite. The last inequality of the Lemma follows. □

B.2 Proof of Proposition 6.1

Consider any covariance matrix Σ that either satisfies $\rho_1(\Sigma, K) > 2$ or $\rho_2(\Sigma, K) > 8$. Thus, there either exists $(A, C, D, R) \in \mathcal{P}_1(\Sigma, K)$ with $MCOD(\Sigma, G) > 2|R|_\infty$, where G be the partition associated to A; or $(A, C, D, R) \in \mathcal{P}_2(\Sigma, K)$ with $\Delta(C) > 8|R|_\infty$. Let us consider another quadruplet $(A', C', D', R') \in \mathcal{P}_1(\Sigma, K)$ (resp. $\mathcal{P}_2(\Sigma, K)$) with $MCOD(\Sigma, G') > 2|R|_\infty$ (resp. $\Delta(C') > 8|R'|_\infty$) and with associated partition G'. We prove below $G = G'$, based on the following Lemma.

Lemma B.2. *Assume either that $(A, C, D, R) \in \mathcal{P}_1(\Sigma, K)$ with $MCOD(\Sigma, G) > 2|R|_\infty$ or that $(A, C, D, R) \in \mathcal{P}_2(\Sigma, K)$ with $\Delta(C) > 8|R|_\infty$. For any integer a between 1 and p set*

$$V(a) = \{a' : a' \neq a,\ COD(a, a') \leqslant 2|R|_\infty\}. \tag{B.6}$$

Writing $k(a)$ for the integer such that $a \in G_{k(a)}$, we have $V(a) = G_{k(a)} \setminus \{a\}$.

Before proving this Lemma, let us explain why it implies that $G = G'$. By symmetry, we can assume that $|R'|_\infty \leqslant |R|_\infty$. So $V'(a) \subset V(a)$ and hence $G'_{k'(a)} \subset G_{k(a)}$. To conclude, we observe that G' is then a sub-partition of G, with the same number K of groups, so the two partitions G and G' are equal. The three assertions (i), (ii) and (iii) of the Proposition 6 follows.

To conclude, it remains to prove the Lemma B.2.

Proof of Lemma B.2. Since $\Sigma = ACA^t + R + D$, writing $k(a)$ for the integer such that $a \in G_{k(a)}$, we have $\Sigma_{ab} = C_{k(a)k(b)} + R_{ab}$ for any $a \neq b$.

First, if $a' \in G_{k(a)} \setminus \{a\}$, then $k(a') = k(a)$, so

$$|\Sigma_{ac} - \Sigma_{a'c}| = |R_{ac} + C_{k(a)k(c)} - C_{k(a)k(c)} - R_{a'c}| \leqslant 2|R|_\infty \tag{B.7}$$

for any $c \neq a, a'$. So $a' \in V(a)$ and hence $G_{k(a)} \setminus \{a\} \subset V(a)$.

Conversely, let us prove that $V(a) \subset G_{k(a)} \setminus \{a\}$. If (A, C, D, R) belongs to $\mathcal{P}_1(\Sigma, K)$, then any two a and b in different groups satisfy $COD(a, b) > 2|R|_\infty$, implying that $V(a) \subset G_{k(a)} \setminus \{a\}$. Now assume that (A, C, D, R) belongs to $\mathcal{P}_2(\Sigma, K)$ and that there exists $a' \in V(a)$ such that $k(a') \neq k(a)$. Hence, $m > 1$ and $a' \notin G_{k(a)}$. We can therefore find $b \in G_{k(a)} \setminus \{a\}$ and $b' \in G_{k(a')} \setminus \{a'\}$. We have

$$C_{k(a)k(a)} - C_{k(a)k(a')} = \Sigma_{ab} - \Sigma_{a'b} - R_{ab} + R_{a'b} \tag{B.8}$$

$$C_{k(a)k(a')} - C_{k(a')k(a')} = \Sigma_{ab'} - \Sigma_{a'b'} - R_{ab'} + R_{a'b'}. \tag{B.9}$$

Since $a' \in V(a)$, we have $COD(a, a') \leqslant 2|R|_\infty$ so

$$\Delta(C) \leqslant C_{k(a)k(a)} + C_{k(a')k(a')} - 2C_{k(a)k(a')} \tag{B.10}$$

$$= \Sigma_{ab} - \Sigma_{a'b} - \Sigma_{ab'} + \Sigma_{a'b'} + R_{a'b} - R_{ab} - R_{a'b'} + R_{ab'} \tag{B.11}$$

$$\leqslant 8|R|_\infty, \tag{B.12}$$

which is in contradiction with $\Delta(C) > 8|R|_\infty$. So it cannot hold that $k(a') \neq k(a)$, which means that any $a' \in V(a)$ belongs to $G_{k(a)} \setminus \{a\}$, i.e. $V(a) \subset G_{k(a)} \setminus \{a\}$. This conclude the proof of the equality $V(a) = G_{k(a)} \setminus \{a\}$. □

B.3 Examples of Σ with $\rho(\Sigma, K) = 8$

Consider the matrix $\Sigma = \begin{bmatrix} 3r & r & 0 & 0 \\ r & 3r & 0 & 0 \\ 0 & 0 & 3r & r \\ 0 & 0 & r & 3r \end{bmatrix} + I_p$ with $r > 0$ and the convention that each entry in the matrix corresponds to a block of size 2. Fixing $K = 3$, we have the following decompositions for Σ

$$\Sigma - I_p = \underbrace{\begin{bmatrix} 4r & 0 & -r & -r \\ 0 & 4r & -r & -r \\ -r & -r & 2r & 2r \\ -r & -r & 2r & 2r \end{bmatrix}}_{=A_1 C_1 A_1^t} + \underbrace{\begin{bmatrix} r & -r & r & r \\ -r & r & r & r \\ r & r & r & -r \\ 0 & 0 & -r & r \end{bmatrix}}_{=R_1 + D_1 - I_p} \tag{B.13}$$

$$= \underbrace{\begin{bmatrix} 2r & 2r & -r & -r \\ 2r & 2r & -r & -r \\ -r & -r & 4r & 0 \\ -r & -r & 0 & 4r \end{bmatrix}}_{=A_2 C_2 A_2^t} + \underbrace{\begin{bmatrix} r & -r & r & r \\ -r & r & r & r \\ r & r & -r & r \\ r & r & r & -r \end{bmatrix}}_{=R_2 + D_2 - I_p} \tag{B.14}$$

For both decompositions, we have $\Delta(C_1) = \Delta(C_2) = 8r = 8|R_1|_\infty = 8|R_2|_\infty$, so that two different partitions G_1 and G_2 lead to the same ratio $\Delta(C_i)/|R_i|_\infty = 8$. It remains to prove that

$\max_{C,R} \Delta(C)/|R|_\infty = 8$ to conclude. For any decomposition associated to size 3 partition G, we have $|R|_\infty \geqslant r$. By symmetry we may assume that one group of $G_1 \cap G_1^* \neq \emptyset$ and $G_3 \cap G_3^* \neq \emptyset$ and $G_3 \cap G_4^* \neq \emptyset$. As consequence, $C_{3,3} \leqslant r + |R|_\infty$, $C_{1,1} \leqslant 3r + |R|_\infty$ and $C_{1,3} \geqslant -r - |R|_\infty$ implying that $\Delta(C) \leqslant 4r + 4|R|_\infty$. Since $|R|_\infty \geqslant r$, we obtain $\rho_2(\Sigma, 3) \leqslant 8$. So we have prove that $\rho_2(\Sigma, 3) = 8$ and that the maximum is achieved for at least 2 different partitions.

C Proofs for Section 3 p. 33: minimax lower bounds

These proofs of Theorems 3.2 and 3.1 are based on a version of Fano's Lemma that is appropriate for our application. Specifically, we will employ Birgé's Lemma (Corollary 2.18 in Massart, 2007), which we state below, translated to our problem. We let $\mathcal{S}$ denote generically either $\mathcal{M}(\eta, m)$ or $\mathcal{D}(\tau, m)$. Then:

Lemma C.1. *For any partition estimator $\widehat{G}$, and for any collection of distinct covariance matrices $\Sigma^{(j)} \in \mathcal{S}$, $1 \leqslant j \leqslant M$, we have*

$$\sup_{\Sigma \in \mathcal{S}} \mathbb{P}_\Sigma(\widehat{G} \neq G^*) \geqslant \max_{j=1,\dots,M} \mathbb{P}_{\Sigma^{(j)}}(\widehat{G} \neq G^{(j)}) \tag{C.1}$$

$$\geqslant \frac{1}{2e+1} \wedge \left(1 - \max_{j \geqslant 2} \frac{\mathcal{K}(\Sigma^{(j)}, \Sigma^{(1)})}{\log(M)}\right), \tag{C.2}$$

$$\geqslant \frac{1}{2e+1} \wedge \left(1 - \max_{j \geqslant 2} \frac{n\|(\Sigma^{(1)})^{-1}(\Sigma^{(j)} - \Sigma^{(1)})\|^2}{2\log(M)}\right), \tag{C.3}$$

where $\mathcal{K}(\Sigma^{(j)}, \Sigma^{(1)})$ denotes the Kulback-Leibler divergence between two Gaussian likelihoods based on n observations.

Proof. Inequality (C.2) is Birgé's Lemma (Corollary 2.18 in Massart, 2007), translated to our problem. We prove inequality (C.3) below. We only need to check that

$$\mathcal{K}(\Sigma^{(j)}, \Sigma^{(1)}) \leqslant n\|(\Sigma^{(1)})^{-1}(\Sigma^{(j)} - \Sigma^{(1)})\|^2/2. \tag{C.4}$$

We have

$$\mathcal{K}(\Sigma^{(j)}, \Sigma^{(1)}) = \frac{n}{2}\left(Trace((\Sigma^{(1)})^{-1}\Sigma^{(j)} - I) - \log det\left((\Sigma^{(1)})^{-1}\Sigma^{(j)}\right)\right) \tag{C.5}$$

$$= \frac{n}{2}(F((\Sigma^{(1)})^{-1}\Sigma^{(j)}) - F(I)). \tag{C.6}$$

with $F(S) = Trace(S) - \log det(S)$. Notice that F is convex, and therefore

$$F(I + H) - F(I) \leqslant \langle I - (I + H)^{-1}, H\rangle, \tag{C.7}$$

since the gradient of our function F in $I + H$ is $I - (I + H)^{-1}$. Let $\sigma_1 \geqslant \sigma_2 \geqslant \dots$ be the singular values of H. Then

$$\langle I - (I + H)^{-1}, H\rangle = \sum_k \frac{\sigma_k^2}{1 + \sigma_k} \leqslant \sum_k \sigma_k^2 = \|H\|^2. \tag{C.8}$$

Consequently, (C.3) holds, and the proof of this Lemma is complete. □

C.1 Minimax lower bounds with respect to the MCOD metric: Proof of Theorem 3.1

We derive the lower bound by constructing an example that corresponds to partitions with $K = 3$ blocks of size $p/3$ each, and thus the size of the smallest cluster $m = p/3$. We would expect that in this case the clusters will be easier to separate, and that m will play a role in the minimax bound, however our result below shows that the perfect recovery separation rate is $O(\sqrt{\log p/n})$. To construct the class of covariance matrices to which we will apply Lemma C.1 we let $0 < \epsilon < 1$ and define

$$C(\epsilon) = \begin{bmatrix} \epsilon & \epsilon - \epsilon^2 & -\epsilon \\ \epsilon - \epsilon^2 & \epsilon & \epsilon \\ -\epsilon & \epsilon & 2 \end{bmatrix}, \tag{C.9}$$

which is positive semi-definite. Consider the following covariance matrix with $K = 3$ blocks of equal size $m = p/3$ given by $\Sigma := AC(\epsilon)A^t + I$, where A is a $p \times 3$ hard assignment matrix given by $A_{j1} = 1$, if $1 \leqslant j \leqslant m$, and zero otherwise, $A_{j2} = 1$, if $m + 1 \leqslant j \leqslant 2m$, and zero otherwise and $A_{j3} = 1$, if $2m + 1 \leqslant j \leqslant 3m$, and zero otherwise. We observe that $\text{MCOD}(\Sigma) = 2\epsilon$ and $|\Sigma|_\infty = 3$. For $a \in \{1, \ldots, m\}$ and $b \in \{m + 1, \ldots, 2m + 1\}$, we construct $\Sigma^{(a,b)}$ by permuting the indices a and b. Naturally, the candidate matrices will still correspond to the same $K = 3$ and $m = p/3$, but the groups in the corresponding partitions will have switched labels. We collect the $M = m^2 + 1$ candidate matrices in a set $\mathcal{F}$.

Let $G^{'}$ be the partition associated with one of the generic matrices $\Sigma' \in \mathcal{F}$. We show below that if

$$\epsilon \leqslant \sqrt{\frac{0.8 \log(p/3)}{n}} \quad \text{and} \quad \eta \leqslant 0.8\epsilon, \tag{C.10}$$

then

$$\inf_{\widehat{G}} \sup_{\Sigma \in \mathcal{M}(\eta)} \mathbb{P}_\Sigma(\widehat{G} \neq G) \geqslant \inf_{\widehat{G}} \max_{\Sigma' \in \mathcal{F}} \mathbb{P}_{\Sigma'}(\widehat{G} \neq G^{'}) \tag{C.11}$$

$$\geqslant \frac{1}{2e + 1}\,. \tag{C.12}$$

To apply Lemma C.1, we calculate below the KL-divergence $\mathcal{K}(\Sigma^{(a,b)}, \Sigma)$. By symmetry, we can assume that $a = 1$ and $b = m + 1$. We write henceforth Q for the permutation matrix associated to the transposition of 1 and $m + 1$ and $\Sigma' = Q\Sigma Q^t$. We start by writing out the eigenvalues and the corresponding eigenvectors of $C(\epsilon)$:

- $\lambda_1 = \epsilon(2 - \epsilon)$ with $u_1^t = \frac{1}{\sqrt{2}}[1 \quad 1 \quad 0]$,
- $\lambda_2 = 2 + \epsilon^2$ with $u_2^t = \frac{\epsilon}{2\sqrt{1+\epsilon^2/2}}[-1 \quad 1 \quad 2/\epsilon]$,
- $\lambda_3 = 0$ with $u_3^t = \frac{1}{\sqrt{2+\epsilon^2}}[1 \quad -1 \quad \epsilon]$.

Let $v_k^t = \frac{1}{\sqrt{m}}(Au_k)^t$, for $k = 1, 2$. Then, with I denoting the $p \times p$ identity matrix, we have

$$\Sigma = \sum_{k=1}^{2} m\lambda_k v_k v_k^t + I, \quad \Sigma^{-1} = -\sum_{k=1}^{2} \frac{m\lambda_k}{1+m\lambda_k} v_k v_k^t + I \tag{C.13}$$

$$\Sigma' - \Sigma = \sum_{k=1}^{2} m\lambda_k [Qv_k v_k^t Q^t - v_k v_k^t]. \tag{C.14}$$

We have $Qv_k = v_k + \Delta_k$ with $\Delta_k^t = \frac{1}{\sqrt{m}}[(u_k)_2 - (u_k)_1, 0...0, (u_k)_1 - (u_k)_2, 0...0]$. Then, $\Delta_1 = 0$, $\Delta_2^t = \frac{\epsilon}{\sqrt{1+\epsilon^2/2}\sqrt{m}}[1, 0...0, -1, 0...0]$, and

$$\Sigma' - \Sigma = m\lambda_2(v_2\Delta_2^t + \Delta_2 v_2^t + \Delta_2\Delta_2^t). \tag{C.15}$$

We notice that $v_1 v_1^t \Delta_2 = 0$ and $v_1 v_1^t v_2 = 0$, so putting pieces together:

$$\Sigma^{-1}\Sigma' = I + \Sigma^{-1}(\Sigma' - \Sigma) = I + m\lambda_2 F, \tag{C.16}$$

where

$$F := (I - \frac{m\lambda_2}{1+m\lambda_2} v_2 v_2^t)(v_2\Delta_2^t + \Delta_2 v_2^t + \Delta_2\Delta_2^t). \tag{C.17}$$

Writing $s := \Delta_2^t v_2$, $\gamma := \Delta_2^t\Delta_2$ and $\rho := m\lambda_2/(1+m\lambda_2)$, we have

$$F = [v_2\Delta_2^t + \Delta_2 v_2^t + \Delta_2\Delta_2^t - \rho(v_2\Delta_2^t + sv_2 v_2^t + sv_2\Delta_2^t)] \tag{C.18}$$

$$= v_2\Delta_2^t(1 - \rho(1+s)) + \Delta_2 v_2^t + \Delta_2\Delta_2^t - \rho s v_2 v_2^t. \tag{C.19}$$

Let us compute the eigenvalues of F. We observe that the range of F is spanned by v_2 and Δ_2, so we seek for eigenvectors $\omega = v_2 + \alpha\Delta_2$. Since

$$s = -\frac{2\epsilon^2}{m(2+\epsilon^2)} = -\frac{2\epsilon^2}{m\lambda_2}, \quad \text{and} \quad \gamma = \frac{4\epsilon^2}{m\lambda_2} = -2s, \tag{C.20}$$

computing $F\omega$, we obtain

$$F\omega = [(1-\rho(1+s))(s+\alpha\gamma) - \rho s(1+\alpha s)]v_2 + [1 + \alpha s + s + \alpha\gamma]\Delta_2 \tag{C.21}$$

$$= [(1-\rho(1+s))(1-2\alpha) - \rho(1+\alpha s)]sv_2 + [1 + (1-\alpha)s]\Delta_2, \tag{C.22}$$

so

$$F\omega = \mu\omega \iff \begin{cases} 1 + (1-\alpha)s = \alpha\mu \\ [(1-\rho(1+s))(1-2\alpha) - \rho(1+\alpha s)]s = \mu \end{cases} \tag{C.23}$$

$$\iff \begin{cases} \alpha = \frac{1+s}{\mu+s} \\ 0 = \mu^2 + \mu[\rho s(2+s)] + (1-\rho)(s+2)s. \end{cases} \tag{C.24}$$

Therefore, the two non-zero eigenvalues μ_1 and μ_2 of F fulfill

$$\begin{cases} \mu_1 + \mu_2 = -\rho s(2+s) \\ \mu_1 \mu_2 = (1-\rho)s(2+s). \end{cases} \tag{C.25}$$

Since $\Sigma^{-1}\Sigma' = I + m\lambda_2 F$, with F of rank 2, we can now compute $\mathcal{K}(\Sigma', \Sigma)$:

$$\begin{aligned} \frac{2}{n}\mathcal{K}(\Sigma', \Sigma) &= \operatorname{tr}(m\lambda_2 F) - \log\det(I + m\lambda_2 F) && \text{(C.26)} \\ &= m\lambda_2(\mu_1+\mu_2) - \log(1 + m\lambda_2(\mu_1+\mu_2) + (m\lambda_2)^2\mu_1\mu_2) && \text{(C.27)} \\ &= -m\lambda_2\rho s(2+s) - \log\left(1 - m\lambda_2\rho s(2+s) + (m\lambda_2)^2(1-\rho)s(2+s)\right). && \text{(C.28)} \end{aligned}$$

We observe that $m\lambda_2(1-\rho) = \rho$, so the terms in the log cancel and finally, plugging the values in (C.20) into the above display, we obtain

$$\mathcal{K}(\Sigma', \Sigma) = -\frac{n}{2} m\lambda_2 \rho s(2+s) = 2n\rho\epsilon^2 \left(1 - \frac{\epsilon^2}{m(2+\epsilon^2)}\right) \leqslant 2n\epsilon^2. \tag{C.29}$$

Since we have $M \geqslant (p/3)^2$ candidate matrices, Lemma C.1 ensures that when $\epsilon \leqslant \sqrt{\frac{2e\log(p/3)}{(2e+1)n}}$ the probability of mis-clustering is larger than $1/(2e+1) \geqslant 1/7$. This completes the proof. □

C.2 Minimax cluster lower bounds with respect to the $\Delta(C^*)$-metric: Proof of Theorem 3.2

It is sufficient to consider the model corresponding to $C^* = \tau I_K$ and $\Gamma = I_p$, so that, given a partition G, the covariance matrix decomposes as

$$\Sigma_G = A_G(\tau I_K)A_G^t + I_p\,, \tag{C.30}$$

where A_G is the assignment matrix associated to the partition G. Note that, in this case $\Delta(C^*) = 2\tau$ and $|\Gamma|_\infty = 1$. Define $\mathcal{G}$ the class of all partitions of $\{1, \ldots, p\}$ into K groups of identical size $m = p/K$. Recall that, as before, $\mathbb{P}_{\Sigma_G}$ refers to the normal distribution with covariance Σ_G. We will once again use Lemma C.1. To this end, we construct the candidate matrices below.

Construction of the covariance matrices $\Sigma^{(j)}$ Let $A^{(0)}$ be the assignment matrix such that the m first variables belong to the first group, the next m belong to the second group and so on. In other words,

$$A^{(0)} = \begin{bmatrix} 1 & 0 & 0 & \ldots & & 0 \\ \vdots & & & & & \\ 1 & 0 & 0 & \ldots & & 0 \\ 0 & 1 & 0 & \ldots & & 0 \\ & \vdots & \ddots & & & \\ & & & & 0 & 1 \\ & & & & & \vdots \\ & & & & & 1 \end{bmatrix} \tag{C.31}$$

so that

$$\Sigma^{(0)} = \begin{bmatrix} 1+\tau & \tau & \tau & 0 & \dots & & & \\ \tau & \ddots & \tau & & & & & \\ \tau & \tau & 1+\tau & & & & & \\ 0 & & & 1+\tau & \tau & \tau & 0 & \dots \\ \vdots & & & \tau & \ddots & \tau & & \\ & & & \tau & \tau & 1+\tau & & \\ & & & 0 & & & \ddots & \\ & & & \vdots & & & & \end{bmatrix}, \tag{C.32}$$

where $\Sigma^{(0)} = A^{(0)}\tau I_K {A^{(0)}}^t + I_p$. Note that the associated partition for $G^{(0)}$ is $\{\{1...m\}...\{p-m+1...p\}\}$. For any $a = m+1, \dots, p$, denote μ_a the transposition between 1 and a in $\{1...p\}$. Then, for any $a = m+1, \dots, p$, define the assignement matrix $A^{(a)}$ and $\Sigma^{(a)}$ by

$$A^{(a)}_{ij} = A^{(0)}_{\mu_a(i),j}\,, \quad \Sigma^{(a)}_{ij} = \Sigma^{(0)}_{\mu_a(i),\mu_a(j)}\,. \tag{C.33}$$

In other words, the corresponding partition $G^{(a)}$ is obtained from $G^{(0)}$ by exchanging the role of the first and the a-th node.

Define the set $T := \{0, m+1, m+2, \dots, p\}$. Then, according to Lemma C.1, we have

$$\inf_{\hat{G}} \max_{j \in T} \mathbb{P}_j\left(\hat{G} \neq G_j\right) \geqslant \frac{1}{2e+1} \bigwedge \left(1 - \frac{\sum_{j \in T\setminus\{0\}} \mathcal{K}(\Sigma^{(j)}, \Sigma^{(0)})}{(|T|-1)\log(|T|)}\right), \tag{C.34}$$

where, as in the previous proof, $\mathcal{K}(\Sigma^{(j)}, \Sigma^{(0)})$ refers to the KL-divergence between two Gaussian likelihoods based on n observations. By symmetry, all the Kullback divergences are equal. Since $2e/(2e+1) \geqslant 0.8$ and $1/(2e+1) \geqslant 1/7$, the RHS of the above inequality is further larger than $1/7$, provide that we show that

$$n\mathrm{KL}(\mathbb{P}_{\Sigma^{(m+1)}}, \mathbb{P}_{\Sigma^{(0)}}) \leqslant 0.8\log(p-m+1) \tag{C.35}$$

where, $\mathrm{KL}(\mathbb{P}_{\Sigma^{(m+1)}}, \mathbb{P}_{\Sigma^{(0)}})$ is the KL-divergence between two Gaussian distributions (n = 1), and is evaluated below.

Lemma C.2. *For any $\tau > 1$ and any integers p and m, we have*

$$\mathrm{KL}(\mathbb{P}_{\Sigma^{(m+1)}}, \mathbb{P}_{\Sigma^{(0)}}) = \frac{2(m-1)\tau^2}{1+m\tau} \tag{C.36}$$

As a consequence, the desired result

$$\inf_{\widehat{G}} \sup_{\Sigma \in \mathcal{D}(\tau)} \mathbb{P}_\Sigma\left[\widehat{G} \neq G^*\right] \geqslant \frac{1}{7}, \tag{C.37}$$

holds as soon as

$$\frac{2n(m-1)\tau^2}{1+m\tau} \leqslant 0.8\log(p-m+1)\,. \tag{C.38}$$

This last condition is satisfied as soon as

$$\tau \leqslant c\left[\sqrt{\frac{\log(p)}{n(m-1)}} \bigvee \frac{\log(p)}{n}\right] \tag{C.39}$$

for some numerical constant $c > 0$. It remains to prove Lemma C.2, and we do so below.

Proof of Lemma C.2. The Kullback-Leibler divergence between two centered normal distributions writes as

$$2\mathrm{KL}(\mathbb{P}_{\Sigma^{(m+1)}}, \mathbb{P}_{\Sigma^{(0)}}) = -\log\det\big((\Sigma^{(0)})^{-1}\Sigma^{(m+1)}\big) + \mathrm{trace}\big((\Sigma^{(0)})^{-1}\Sigma^{(m+1)} - I_p\big), \tag{C.40}$$

so that we only have to compute the determinant and the trace of matrix $A := (\Sigma^{(1)})^{-1}\Sigma^{(m+1)}$. We shall see that see that A is a rank 2 perturbation of the identity matrix, so that we will only need to compute its two eigenvalues different from zero.

Observe that for $i = 0, m+1$, the matrices $A^{(i)}{A^{(i)}}^t$ admit exactly K non-zero eigenvalues that are all equal to m. As a consequence, we can decompose $A^{(i)}{A^{(i)}}^t = m\sum_{k=1}^K u_k^{(i)}(u_k^{(i)})^t$ where $u_k^{(i)}$ is a unit vector whose non zero components are all equal to $1/m$ and correspond to the k-th group in $G^{(i)}$. Note that $u_k^{(0)} = u_k^{(m+1)}$ for $k = 3, \dots, K$ as $A^{(0)}$ and $A^{(m+1)}$ only differ by rows 1 and $m+1$. The orthogonal projector $P_i = \sum_{k=1}^K u_k^{(i)}(u_k^{(i)})^t$ satisfies

$$\Sigma^{(i)} = m\tau P_i + I_p = (1+m\tau)P_i + (I_p - P_i)\,. \tag{C.41}$$

Since P_i and $I_p - P_i$ are orthogonal,

$$(\Sigma^{(i)})^{-1} = (1+m\tau)^{-1}P_i + (I_p - P_i) = I_p - \frac{m\tau}{1+m\tau}P_i \tag{C.42}$$

As a consequence of the above observations, we have

$$A = I_p + (\Sigma^{(0)})^{-1}\Big[\Sigma^{(m+1)} - \Sigma^{(0)}\Big] \tag{C.43}$$

$$= I_p + m\tau\big(P_{m+1} - P_0\big) - \frac{m^2\tau^2}{1+m\tau}P_0(P_{m+1} - P_0) =: I_p + B \tag{C.44}$$

The matrices P_0 and P_{m+1} are $k-1$ block diagonal with a first block of size $2m \times 2m$. Besides, P_0 and P_{m+1} take the same values on all the $K-2$ remaining blocks. To compute the non-zero eigenvalues of B, we only to consider the restrictions $\overline{P}_0$ and $\overline{P}_{m+1}$ of P_0 and P_{m+1} to the first $2m \times 2m$ entries. Also observe that the matrices $\overline{P}_0$ and $\overline{P}_{m+1}$ are 4×4 block-constant, with block size

$$\begin{bmatrix} 1\times 1 & 1\times(m-1) & 1\times 1 & 1\times(m-1) \\ (m-1)\times 1 & (m-1)\times(m-1) & (m-1)\times 1 & (m-1)\times(m-1) \\ 1\times 1 & 1\times(m-1) & 1\times 1 & 1\times(m-1) \\ (m-1)\times 1 & (m-1)\times(m-1) & (m-1)\times 1 & (m-1)\times(m-1) \end{bmatrix} \tag{C.45}$$

and the entries are

$$m\overline{P}_0 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \text{ and } m\overline{P}_{m+1} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}. \tag{C.46}$$

As a consequence, the non zero eigenvalues of B are the same as those of

$$C := m\tau\big(\underline{P}_{m+1} - \underline{P}_0\big) - \frac{m^2\tau^2}{1+m\tau}\underline{P}_0(\underline{P}_{m+1} - \underline{P}_0) \tag{C.47}$$

where $\underline{P}_{m+1}$ and $\underline{P}_0$ are two 4×4 matrices

$$m\underline{P}_0 = \begin{bmatrix} 1 & (m-1) & 0 & 0 \\ 1 & (m-1) & 0 & 0 \\ 0 & 0 & 1 & (m-1) \\ 0 & 0 & 1 & (m-1) \end{bmatrix} \text{ and } m\underline{P}_{m+1} = \begin{bmatrix} 1 & 0 & 0 & (m-1) \\ 0 & (m-1) & 1 & 0 \\ 0 & (m-1) & 1 & 0 \\ 1 & 0 & 0 & (m-1) \end{bmatrix}. \tag{C.48}$$

Working out the product of matrices, we get

$$C = -\tau \begin{bmatrix} 0 & (m-1) & 0 & -(m-1) \\ 1 & 0 & -1 & 0 \\ 0 & -(m-1) & 0 & (m-1) \\ -1 & 0 & 1 & 0 \end{bmatrix} + \frac{(m-1)\tau^2}{1+m\tau} \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \tag{C.49}$$

We observe that these two matrices have their first (resp. second) and third (resp. fourth) lines and columns opposite to each other. As a consequence, the two non-zero eigenvalues of C^* are the same as those of

$$D := -2\tau \begin{bmatrix} 0 & (m-1) \\ 1 & 0 \end{bmatrix} + \frac{2(m-1)\tau^2}{1+m\tau} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{C.50}$$

$$= \frac{2\tau}{1+m\tau} \begin{bmatrix} (m-1)\tau & -(m-1)[1+(m-1)\tau] \\ -(1+\tau) & (m-1)\tau \end{bmatrix}. \tag{C.51}$$

Straightforward computations then lead to

$$tr(D) = \frac{4(m-1)\tau^2}{1+m\tau}, \quad \det(D) = -tr(D) \tag{C.52}$$

Coming back to (C.40), we have

$$2\mathrm{KL}(\mathbb{P}_{\Sigma^{(m+1)}}, \mathbb{P}_{\Sigma^{(0)}}) = -\log\det\big(A\big) + \mathrm{trace}\big(A - I_p\big)\big] \tag{C.53}$$

$$= -\log\det(I+D) + tr(D) \tag{C.54}$$

$$= tr(D) - \log\big[1 + tr(D) + det(D)\big] \tag{C.55}$$

$$= \frac{4(m-1)\tau^2}{1+m\tau}. \tag{C.56}$$

□

D Results for the COD estimator: Sections 4 and 6.2

We give below the proofs of Theorems 4.1, 6.1 and Proposition 4.1.

D.1 Proof of Theorems 4.1 and 6.1

In order to avoid cluttered notation, in all this subsection, we write G for G^* (resp. $G[K]$) and m for m^* (resp. $m[K]$).

The proof consists in the application of Lemma D.1 and Lemma D.2 proved below.

Lemma D.1. *Consider any partition G. Let us set $\tau = \max_{a,b,c=1,\dots,p} |\widehat{\mathbf{cor}}(X_a - X_b, X_c) - \mathbf{cor}(X_a - X_b, X_c)|$,*

$$\mu = \max_{a \overset{G}{\sim} b} \mathrm{sCOD}(a,b), \text{and} \eta = \min_{a \overset{G}{\not\sim} b} \mathrm{sCOD}(a,b), \tag{D.1}$$

where $\mathrm{sCOD}(a,b) := \max_{c \neq a,b} |\mathbf{cor}(X_a - X_b, X_c)|$. Then, under the condition $\mu + \tau \leq \alpha < \eta - \tau$, the COD algorithm exactly recovers the partition G.

The control on the size of τ is given below.

Lemma D.2. *Under Assumption 1, there exist two constants c_1, c_2 such that*

$$\tau := \max_{a,b,c=1,\dots,p} |\widehat{\mathbf{cor}}(X_a - X_b, X_c) - \mathbf{cor}(X_a - X_b, X_c)| \leq c_1 L^2 \sqrt{\frac{\log(p)}{n}}, \tag{D.2}$$

with probability at least $1 - c_2/p$.

To conclude Theorem 3, we apply the two lemmas with $G = G^*$, implying that $\mu = 0$ and $\eta \geq \mathrm{MCOD}(\Sigma)/(2|\Sigma|_\infty)$. To conclude Theorem 6, we remark that $\eta \geq \mathrm{MCOD}(\Sigma, G)/(2|\Sigma|_\infty)$, and $\mu \leq \sqrt{2}|R|_\infty/\lambda_{\min}(\Sigma)$.

Proof of Lemma D.1. First, we notice that,

$$\widehat{\mathrm{sCOD}}(a,b) - \tau \leq \mathrm{sCOD}(a,b) \leq \widehat{\mathrm{sCOD}}(a,b) + \tau. \tag{D.3}$$

We then observe that

$$a \overset{G}{\sim} b \implies \mathrm{sCOD}(a,b) \leq \mu \implies \widehat{\mathrm{sCOD}}(a,b) \leq \mu + \tau, \tag{D.4}$$

and

$$a \overset{G}{\not\sim} b \implies \mathrm{sCOD}(a,b) \geq \eta \implies \widehat{\mathrm{sCOD}}(a,b) \geq \eta - \tau. \tag{D.5}$$

In particular, under the condition $\mu + \tau \leq \alpha < \eta - \tau$, we have

$$a \overset{G}{\sim} b \iff \widehat{\mathrm{sCOD}}(a,b) \leq \alpha. \tag{D.6}$$

Let us prove Lemma D.1 by induction on l. We consider the algorithm at some step l and assume that the algorithm was consistent up to this step, i.e. $\widehat{G}_j = G_{k(a_j)}$ for $j = 1, \dots, l-1$.

If $|S| = 1$, then it directly follows that $\widehat{G} = G$. Assume now that $|S| > 1$.

(i) If $\widehat{\mathrm{sCOD}}(a_l, b_l) > \alpha$, then according to (D.6) no $b \in S$ is in the same group as a_l. Since the algorithm has been consistent up to step l, it means that a_l is a singleton and $\widehat{G}_l := \{a_l\} = G_{k(a_l)}$.

(ii) If $\widehat{\mathrm{sCOD}}(a_l, b_l) \leq \alpha$, then $a_l \overset{G}{\sim} b_l$ according to (D.6). The equivalence (D.6) furthermore ensures that $\widehat{G}_l = S \cap G_{k(a_l)}$. Since the algorithm has been consistent up to this step we have $G_{k(a_l)} \subset S$ and hence $\widehat{G}_l = G_{k(a_l)}$.

To conclude, the algorithm remains consistent at step l and the Lemma D.1 follows by induction. □

Proof of Lemma D.2. We start with the following lemma.

Lemma D.3. *Let W be a random vector in $\mathbb{R}^d$, then*

$$\max_{a,b=1,\dots,d} |\widehat{\mathbf{cor}}(W_a, W_b) - \mathbf{cor}(W_a, W_b)| \leq 2 \max_{a,b=1,\dots,d} \frac{|\widehat{\mathbf{cov}}(W_a, W_b) - \mathbf{cov}(W_a, W_b)|}{\sqrt{\mathbf{var}(W_a)\mathbf{var}(W_b)}}. \tag{D.7}$$

Proof of Lemma D.3. Since both sides of (D.7) are invariant by rescaling any of the W_a, we can always assume that $\mathbf{var}(W_a) = 1$ for all $a = 1 \dots, d$. We denote by R and $\widehat{R}$ the correlation matrix of W and its empirical counterpart, and by S and $\widehat{S}$ its covariance matrix and its empirical counterpart. Since $\mathbf{var}(W_a) = 1$ for all $a = 1 \dots, d$, we have $R = S$, and the triangular inequality gives

$$|\widehat{R}_{ab} - R_{ab}| = |\widehat{R}_{ab}\big(1 - (\widehat{S}_{aa}\widehat{S}_{bb})^{1/2}\big) + \widehat{S}_{ab} - S_{ab}| \tag{D.8}$$

$$\leq |\widehat{R}_{ab}|\,|1 - (\widehat{S}_{aa}\widehat{S}_{bb})^{1/2}| + |\widehat{S}_{ab} - S_{ab}|. \tag{D.9}$$

We notice that

$$|1 - (\widehat{S}_{aa}\widehat{S}_{bb})^{1/2}| \leq |1 - \widehat{S}_{aa}| \vee |1 - \widehat{S}_{bb}| = |S_{aa} - \widehat{S}_{aa}| \vee |S_{bb} - \widehat{S}_{bb}|. \tag{D.10}$$

Since $|\widehat{R}_{ab}| \leq 1$, we conclude that for any $a, b \in \{1, \dots, d\}$

$$|\widehat{R}_{ab} - R_{ab}| \leq |\widehat{R}_{ab}| \left(|S_{aa} - \widehat{S}_{aa}| \vee |S_{bb} - \widehat{S}_{bb}|\right) + |\widehat{S}_{ab} - S_{ab}| \tag{D.11}$$

$$\leq 2|\widehat{S} - S|_\infty. \tag{D.12}$$

The proof of Lemma D.3 is complete. □

We now apply previous lemma to the random vector W in dimension $d = p(p+1)/2$ gathering all the X_c for $c = 1, \dots, p$ and all the $X_a - X_b$ for $1 \leq a < b \leq p$. We complete the proof of Lemma D.2 by combining (D.7) with (G.3) and by noticing that we always have $\tau \leq 2$ so the term with the square-root is dominant. □

D.2 Proof of Proposition 4.1

For any $a < b$, we introduce the notation $\epsilon_{ab}^{(i)} = \widehat{\Delta}_{ab}^{(i)} - \Delta_{ab}$ for $i = 1, 2$. We recall the notation $ECV(G) = \mathbb{E}^{(2)}[CV(G)]$. For any partition G, we have

$$ECV(G) - ECV(G^*) = \sum_{a \overset{G^*}{\sim} b} \mathbf{1}_{a \overset{G}{\nsim} b} \mathbb{E}^{(2)} \left[|\epsilon_{ab}^{(1)} - \epsilon_{ab}^{(2)}|_\infty^2 - |\epsilon_{ab}^{(1)}|_\infty^2 \right] \tag{D.13}$$

$$+ \sum_{a \overset{G^*}{\nsim} b} \mathbf{1}_{a \overset{G}{\sim} b} \mathbb{E}^{(2)} \left[|\Delta_{ab} + \epsilon_{ab}^{(1)}|_\infty^2 - |\epsilon_{ab}^{(1)} - \epsilon_{ab}^{(2)}|_\infty^2 \right]. \tag{D.14}$$

Let us prove that both terms are positive both in expectation and in probability with respect to $\mathbb{P}^{(2)}$.

The property (P1) ensures that since $\mathbb{E}[\epsilon_{ab}^{(2)}] = 0$. So, for the first sum, Jensen inequality implies that

$$\mathbb{E}^{(2)} \left[|\epsilon_{ab}^{(1)} - \epsilon_{ab}^{(2)}|_\infty^2 \right] \geq \left| \epsilon_{ab}^{(1)} - \mathbb{E}^{(2)} \left[\epsilon_{ab}^{(2)} \right] \right|_\infty^2 = |\epsilon_{ab}^{(1)}|_\infty^2. \tag{D.15}$$

Hence the first sum is always non-negative.

Let us now consider the second sum where $a \overset{G^*}{\nsim} b$ but $a \overset{G}{\sim} b$. The inequality

$$|\Delta_{ab}|_\infty^2 \leq (|\Delta_{ab} + \epsilon_{ab}^{(1)}|_\infty + |\epsilon_{ab}^{(1)}|_\infty)^2 \leq 2|\Delta_{ab} + \epsilon_{ab}^{(1)}|_\infty^2 + 2|\epsilon_{ab}^{(1)}|_\infty^2, \tag{D.16}$$

ensures the lower bound

$$\mathbb{E}^{(2)} \left[|\Delta_{ab} + \epsilon_{ab}^{(1)}|_\infty^2 - |\epsilon_{ab}^{(1)} - \epsilon_{ab}^{(2)}|_\infty^2 \right] \geq \frac{1}{2} |\Delta_{ab}|_\infty^2 - 2|\epsilon_{ab}^{(1)}|_\infty^2 - \mathbb{E}^{(2)}[|\epsilon_{ab}^{(2)}|_\infty^2]. \tag{D.17}$$

The Lemma D.3 together with the Corollary G.1 of Hanson-Wright inequality ensures that

$$\max_{a,b,c=1,\ldots,p} |[\epsilon_{ab}^{(i)}]_c|^2 \leq \min \left(2, cL^2 \left(\sqrt{\frac{\log(p)}{n}} + \frac{\log(p)}{n} \right) \right)^2 \leq 4c^2 L^4 |\Sigma|_\infty^2 \frac{\log(p)}{n}, \tag{D.18}$$

both with probability larger than $1 - 4/p$ and in expectation with respect to $\mathbb{P}^{(i)}$. Since

$$4|\Delta_{ab}|_\infty^2 \geq MCOD(\Sigma)^2 / |\Sigma|_\infty^2 > c_1^2 L^4 \log(p)/n, \tag{D.19}$$

the proof of Proposition 2 is complete.

E Proofs regarding cluster recovery with Pecok: Theorem 6.2 of Section 6.3

E.1 Proofs of the Lemmas A.4, A.5, A.6 and A.7 used in the proofs of Theorems A.1 and A.2 stated in Section A.3

We recall that these theorems, and the Lemmas that lead to their conclusion, are used to prove Theorem 6.2.

Proof of Lemma A.4. Fix $j \neq k \in [p]$. For $i = 1, \dots, n$, the random variables $T_i = \underline{\mathbf{Z}}_{ij} - \underline{\mathbf{Z}}_{ik}$ are independent and identically distributed with

$$\mathbb{E}[T_i^2] = C^*_{jj} + C^*_{kk} - 2C^*_{jl} + \gamma_{kk} + \gamma_{jj} - 2\gamma_{jl} = \Delta_{jk}(C^*) + \Delta_{jk}(\gamma). \tag{E.1}$$

Let us first work under assumption 1. We have $\|T_i\|^2_{\psi_2} \leqslant \operatorname{Var} T_i L^2$. Applying the Hanson-Wright inequality (G.1) to $|\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2$, we obtain, for some constant $c > 0$,

$$\begin{aligned} |\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2 - n(\gamma_{jj} + \gamma_{kk} - 2\gamma_{jk}) \geqslant & n\Delta_{jk}(C^*) - cL^2(\Delta_{jk}(C^*) \\ & + \Delta_{jk}(\gamma))(\log(p) + \sqrt{n\log(p)}) \,, \end{aligned} \tag{E.2}$$

with probability higher than $1 - 1/(pK^2)$. We derive from $|\Delta_{jk}(\gamma)| \leqslant 4|R|_\infty + 2|D|_\infty/m$, Condition (42) of Theorem 8 and the condition $L^4 \log(p) \leqslant c_1 n$ that

$$|\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2 - n(\gamma_{jj} + \gamma_{kk} - 2\gamma_{jk}) \geqslant n\frac{\Delta_{jk}(C^*)}{2}, \tag{E.3}$$

with probability higher than $1 - 1/(pK^2)$. Taking an union bound over all $j \neq k$ leads to (53) of Lemma 6.

Let us now turn to Assumption 1-bis. Since $\underline{\mathbf{Z}}_{ij} = m_j^{-1} \sum_{a \in G_j} \mathbf{X}_{ia}$, it follows that $|T_i| \leqslant 2M$ almost surely. Then, applying Hanson-Wright inequality for Bernstein-type random variables (G.2), we derive that, for some constant $c > 0$

$$\begin{aligned} |\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2 - n(\gamma_{jj} + \gamma_{kk} - 2\gamma_{jk}) \geqslant & n\Delta_{jk}(C^*) - cM^2\log(p) \\ & - cM\sqrt{(\Delta_{jk}(C^*) + \Delta_{jk}(\gamma))n\log(p)} \,, \end{aligned} \tag{E.4}$$

with probability higher than $1 - 1/(pK^2)$. We derive from $|\Delta_{jk}(\gamma)| \leqslant 4|R|_\infty + 2|D|_\infty/m$ and Condition (43) of Theorem 8 that

$$|\underline{\mathbf{Z}}_{:j} - \underline{\mathbf{Z}}_{:k}|_2^2 - n(\gamma_{jj} + \gamma_{kk} - 2\gamma_{jk}) \geqslant n\frac{\Delta_{jk}(C^*)}{2}, \tag{E.5}$$

with probability higher than $1 - 1/(pK^2)$. Taking an union bound over all $j \neq k$ leads to (53) of Lemma 6. □

Proof of Lemma A.5. For all $a \sim b$, one has $(W_2)_{ab} = 0$. For all $a \nsim b$, it follows from the covariance structure of $\underline{Z}$ and $\underline{E}$ that the expectation of $\mathbb{E}[(W_2)_{ab}]$ is zero. We will control the deviation of W_2 from zero with the Corollary G.1 of Hanson-Wright inequality, with $u = B^*(e_a - e_b)$ and $v = (I - B^*)(e_a - e_b)$. We observe that $\operatorname{Var} \underline{E}_a - \underline{E}_b \leqslant 16|\Gamma|_\infty$ and

$$\operatorname{Var}(A\underline{Z})_{:a} - (A\underline{Z})_{:b} = \operatorname{Var} \underline{Z}_{k(a)} - \underline{Z}_{k(b)} \leqslant \Delta_{k(a)k(b)}(C^*) + 2\frac{|D|_\infty}{m} + 3|R|_\infty. \tag{E.6}$$

Let us control $(W_2)_{ab}$ under Assumption 1. From the variance bounds and the Corollary G.1 of Hanson-Wright inequality, we derive that

$$|(W_2)_{ab}| \leqslant cL^2\sqrt{\operatorname{Var} \underline{E}_a - \underline{E}_b \operatorname{Var}(A\underline{Z})_{:a} - (A\underline{Z})_{:b}} \left[\sqrt{n\log(p)} + \log(p)\right] \tag{E.7}$$

$$\begin{aligned} & \leqslant c' L^2 \left[\sqrt{\Delta_{k(a)k(b)}(C^*)|\Gamma|_\infty} + \frac{|D|_\infty}{\sqrt{m}} + |R|_\infty + |R|_\infty^{1/2}|D|_\infty^{1/2}\right] \\ & \quad \times \left[\sqrt{n\log(p)} + \log(p)\right], \end{aligned} \tag{E.8}$$

with probability higher than $1 - 4/p^3$. Taking an union bound over all (a, b) such that $a \nsim b$ leads to the desired result.

Turning to Assumption 1-bis, we use Lemma G.2 below, which gives

$$|(W_2)_{ab}| \leqslant c \left[M \sqrt{n \log(p) \left[\Delta_{k(a)k(b)}(C^*) + \frac{|D|_\infty}{m} + |R|_\infty \right]} + M^2 \log(p) \right] \tag{E.9}$$

with probability higher than $1 - 4/p^3$. The result follows again from a union bound. □

Proof of Lemma A.6. By definition of $\underline{\mathbf{E}}$ and $\widetilde{\Gamma}$, one has $\widetilde{\Gamma} = (I - B^*)\widetilde{\Gamma}(I - B^*)$ and therefore $B^*\widetilde{\Gamma} = 0$. Then, Lemma A.3 entails

$$|\langle \widetilde{\Gamma} - (I - B^*)\Gamma(I - B^*), B^* - B \rangle| \leqslant \|\widetilde{\Gamma} - (I - B^*)\Gamma(I - B^*)\|_{op} \frac{\sum_{j \neq k} |B_{G_j G_k}|_1}{m}. \tag{E.10}$$

Let us first control the random variable $\|\widetilde{\Gamma} - (I - B^*)\Gamma(I - B^*)\|_{op}$ under Assumption. We apply Lemma G.3 and obtain

$$\|\widetilde{\Gamma} - (I - B^*)\Gamma(I - B^*)\|_{op} \leqslant c_L \|\Gamma\|_{op} \left[\sqrt{\frac{p}{n}} + \frac{p}{n} \right] , \tag{E.11}$$

with probability higher than $1 - 1/p$.

Turning to Assumption 1-bis, we observe that the random variables $|E_a|$ are all bounded by $2M$. Applying the matrix Bernstein inequality (Lemma G.4 below), we obtain

$$\|\widetilde{\Gamma} - (I - B^*)\Gamma(I - B^*)\|_{op} \leqslant cM \sqrt{\frac{p\|\Gamma\|_{op} \log(p)}{n}} + c\frac{pM^2 \log(p)}{n} , \tag{E.12}$$

with probability higher than $1 - 1/p$. Together with (E.10) and the definition of W_3, these two deviation inequalities allows us to prove the desired results. □

Proof of Lemma A.7. We first observe that

$$|\langle \Gamma - \widehat{\Gamma}, B^* - B \rangle| = |\langle R + D - \widehat{D}, B^* - B \rangle| \tag{E.13}$$

$$\leqslant |\langle R, B^* - B \rangle| + |\langle D - \widehat{D}, B^* - B \rangle|. \tag{E.14}$$

According to Lemma A.3 :

$$|\langle R, B^* - B \rangle| \leqslant \left[\sum_{j \neq k} |B_{G_j G_k}|_1 \right] \left(\frac{\|R\|_{op}}{m} + 6|B^* R|_\infty \right) \tag{E.15}$$

hence

$$|\langle \Gamma - \widehat{\Gamma}, B^* - B \rangle| \leqslant \left(\frac{\|R\|_{op}}{m} + 6|R|_\infty \right) \sum_{j \neq k} |B_{G_j G_k}|_1 + |\langle D - \widehat{D}, B^* - B \rangle|. \tag{E.16}$$

For the second term, we follow the same approach as for $\widetilde{\Gamma} - \Gamma$. The additional ingredient is that $\langle D - \widehat{D}, B^* - B\rangle = \langle D - \widehat{D} - \alpha I_p, B^* - B\rangle$ for any $\alpha \in \mathbb{R}$, since $\mathrm{tr}(B^*) = K = \mathrm{tr}(B)$. According to Lemma A.3, for any $\alpha \in \mathbb{R}$, the following holds:

$$|\langle D - \widehat{D}, B^* - B\rangle| = |\langle D - \widehat{D} - \alpha I_p, B^* - B\rangle| \tag{E.17}$$

$$\leqslant 2 \sum_{j \neq k} |B_{G_j G_k}|_1 \Big(\frac{\|D - \widehat{D} - \alpha I_p\|_{op}}{2m} + 3|B^*[(D - \widehat{D}) - \alpha I_p]|_\infty \Big) \tag{E.18}$$

$$\leqslant \frac{7\|D - \widehat{D} - \alpha I_p\|_\infty}{m} \left[\sum_{j \neq k} |B_{G_j G_k}|_1 \right], \tag{E.19}$$

where we used in the last line that αI_p, D and $\widehat{D}$ are diagonal matrices.

We fix $\alpha = \Big(\max_k (D_{kk} - \widehat{D}_{kk}) + \min_k (D_{kk} - \widehat{D}_{kk}) \Big) / 2$. The above inequality simplifies to

$$|\langle D - \widehat{D}, B^* - B\rangle| \leqslant \frac{7\|D - \widehat{D}\|_V}{2m} \sum_{j \neq k} |B_{G_j G_k}|_1 \, . \tag{E.20}$$

The result follows. □

E.2 Proof of (ii) of Proposition A.1 of Section A.4 and Proof of Lemma A.8

We recall that Proposition A.1 is also an intermediate step for Theorem 6.2 of Section E. We begin with the Proof of (ii). We will follow the same lines of proof as for part (i), adjusting the probabilistic inequalities to the bounded setting. Under Assumption 1-bis, Lemma G.2 below ensures that, simultaneously for all triplet (a, b_1, b_2) of distinct indices,

$$\Big| \frac{1}{n} \langle X_a - X_{b_1}, X_a - X_{b_2}\rangle - \mathrm{Cov}\left(X_a - X_{b_1}, X_a - X_{b_2}\right) \Big| \\ \leqslant c' \Big(M \sqrt{\frac{\mathrm{Var}\, X_a - X_{b_1} \log(p)}{n}} + M^2 \frac{\log(p)}{n} \Big),$$

with probability at least $1 - p^{-4}$. Plugging the estimates (A.80) and (A.81), we get with probability $1 - 1/p^2$

$$\Big| \widehat{\Gamma}_{aa} - D_{aa} \Big| \leqslant c' \Big(|R|_\infty + (t_1 + t_2 + t_{12}) + M \sqrt{|\Gamma|_\infty + t_1} \sqrt{\frac{\log(p)}{n}} + M^2 \frac{\log(p)}{n} \Big) \tag{E.21}$$

$$\leqslant c'' \Big(|R|_\infty + (t_1 + t_2 + t_{12}) + M \sqrt{\frac{|\Gamma|_\infty \log(p)}{n}} + M^2 \frac{\log(p)}{n} \Big), \tag{E.22}$$

where we used for the last inequality $a\sqrt{x+y} \leqslant y + a\sqrt{2x} + a^2$ for any $a, x, y \geqslant 0$.

So again, as in the first part of the proof, we only need to upper bound the bias terms t_1, t_2, t_{12} by $c' \left(\sqrt{|\Gamma|_\infty |R|_\infty} + M \sqrt{|\Gamma|_\infty \log(p)/n} + M^2 \log(p)/n \right)$.

For any c and d such that $k(a) = k(c)$ and $k(ne_1(a)) = k(d)$, Hanson-Wright inequality together with the condition $M^2 \log(p) \leqslant c_1 v^2 n$ give

$$\frac{1}{n}|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2^2 \leqslant \operatorname{Var} X_c - X_d + cM\sqrt{\frac{\operatorname{Var} X_c - X_d \log(p)}{n}} + cM^2\frac{\log(p)}{n} \tag{E.23}$$

$$\leqslant c' \operatorname{Var} X_c - X_d \leqslant c''(|\Gamma|_\infty + t_1), \tag{E.24}$$

with probability $1 - 1/p^2$. So combined with Lemma A.8 proved below, we get with probability at least $1 - 2p^{-2}$

$$\left|\langle \mathbf{X}_{:a} - \mathbf{X}_{:ne_1(a)}, \mathbf{X}_{:c} - \mathbf{X}_{:d}\rangle\right| \leqslant c'\left(\sqrt{n|R|_\infty} + M\sqrt{\log(p)}\right)\sqrt{n[|\Gamma|_\infty + t_1]}. \tag{E.25}$$

Applying again the Lemma G.2 and $M^2 \log(p) \leqslant c_1 v^2 n$, we obtain

$$\left|\langle \mathbf{X}_{:a} - \mathbf{X}_{:ne_1(a)}, \mathbf{X}_{:c} - \mathbf{X}_{:d}\rangle\right| \geqslant nt_1 - 4n|R|_\infty - cM\sqrt{n \operatorname{Var} X_c - X_d \log(p)} - cM^2 \log(p) \tag{E.26}$$

$$\geqslant nt_1 - 4n|R|_\infty - c'M\sqrt{|\Gamma|_\infty + t_1}\sqrt{n\log(p)}. \tag{E.27}$$

with probability $1 - 1/p^2$. Hence, as in the first part of the proof, we conclude that

$$t_1 \leqslant c'\left[|R|_\infty^{1/2}|\Gamma|_\infty^{1/2} + M\sqrt{\frac{|\Gamma|_\infty \log(p)}{n}}\right] \tag{E.28}$$

simultaneously for all a, with probability $1 - c_3/p$. Together with (E.22), this concludes the proof of the second part of the proposition.

Proof of Lemma A.8. In this proof, c' denotes a numerical constant, whose value can vary from line to line.

Since $m \geqslant 3$, there exists two indices b_1 and b_2 other than a belonging to the group $G_{k(a)}$. Fix any c and d different from a and b_1.

Under Assumption 1, by Hanson-Wright inequality and since $\log(p) \leqslant c_1 L^{-4} n$, it holds that

$$|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2^2 \geqslant n \operatorname{Var} X_c - X_d\Big(1 - c'L^2\Big(\sqrt{\frac{\log(p)}{n}} + \frac{\log(p)}{n}\Big)\Big) \tag{E.29}$$

$$\geqslant n \operatorname{Var} X_c - X_d/2\,, \tag{E.30}$$

with probability $1 - 1/p^5$.

Similarly, under Assumption 1-bis, by Hanson-Wright inequality and since $M^2 \log(p) \leqslant c_1 v^2 n$, we get

$$|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2^2 \geqslant n \operatorname{Var} X_c - X_d - c'\Big(M\sqrt{\operatorname{Var} X_c - X_d n \log(p)} + M^2 \log(p)\Big) \tag{E.31}$$

$$\geqslant n \operatorname{Var} X_c - X_d/2\,, \tag{E.32}$$

with probability $1 - 1/p^5$.

The random variable $\langle \mathbf{X}_{:a} - \mathbf{X}_{:b_i}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle$ has expectation $n(R_{ac} - R_{b_i c} + R_{b_i d} - R_{ad})$. Under the diagonal dominance assumption (6.5) and the assumption $\Delta(C^*) \geqslant 0$, we have

$$|\,\mathbb{E}\left[\langle \mathbf{X}_{:a} - \mathbf{X}_{:b_i}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle\right]| \leqslant n\left(|R_{ac}| + |R_{b_i c}| + |R_{b_i d}| + |R_{ad}|\right) \tag{E.33}$$

$$\leqslant 2n|R|_\infty^{1/2}\sqrt{D_{cc} + D_{dd} - 2R_{cd}} \tag{E.34}$$

$$\leqslant 4\sqrt{n|R|_\infty}|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2. \tag{E.35}$$

We observe that the inequality $|\,\mathbb{E}\left[\langle \mathbf{X}_{:a} - \mathbf{X}_{:b_i}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle\right]| \leqslant 4\sqrt{n|R|_\infty}|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2$ still holds when Γ is diagonal (i.e. $R = 0$), since in this case the left-hand side is equal to zero.

Under Assumption 1, by Hanson-Wright inequality, we obtain with probability $1 - 1/p^5$

$$\begin{aligned}|\langle \mathbf{X}_{:a} - \mathbf{X}_{:b_i}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle| &\leqslant |\,\mathbb{E}\left[\langle \mathbf{X}_{:a} - \mathbf{X}_{:b_i}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle\right]| \\ &\quad + cL^2\sqrt{\operatorname{Var} X_c - X_d|\Gamma|_\infty}\sqrt{n\log(p)} \qquad \text{(E.36)} \\ &\leqslant \left(4\sqrt{n|R|_\infty} + cL^2\sqrt{2|\Gamma|_\infty \log(p)}\right)|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2. \qquad \text{(E.37)}\end{aligned}$$

Similarly, under Assumption 1-bis, by Hanson-Wright inequality and $M^2\log(p) \leqslant c_1 v^2 n$, we obtain with probability $1 - 1/p^5$

$$\begin{aligned}|\langle \mathbf{X}_{:a} - \mathbf{X}_{:b_i}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle| &\leqslant |\,\mathbb{E}\left[\langle \mathbf{X}_{:a} - \mathbf{X}_{:b_i}, \mathbf{X}_{:c} - \mathbf{X}_{:d} \rangle\right]| \\ &\quad + cM\sqrt{\operatorname{Var} X_c - X_d}\sqrt{n\log(p)} + cM^2\log(p) \\ &\leqslant \left(4\sqrt{n|R|_\infty}\right) + c'M\sqrt{\log(p)}|\mathbf{X}_{:c} - \mathbf{X}_{:d}|_2. \qquad \text{(E.38)}\end{aligned}$$

By the definition of $ne_1(a)$ and $ne_2(a)$ and a union bound, we obtain the desired result. □

F Analysis of corrected spectral clustering: Section 5.4

In this section, we prove Proposition 5.3. We recall that we work here with the decomposition $\Sigma = AC^*A^t + \Gamma$ where Γ is diagonal. As in the previous section, we write m for m^* and G for G^*.

An alternative formulation of an approximate K-means is the following. Let $\eta > 1$ be a given positive number. Denote $\mathcal{A}_{p,K}$ the collection of membership matrices, that is $p \times K$ binary matrices whose rows contain exactly one non-zero entry. Note that a membership matrix $A \in \mathcal{A}_{p,K}$ defines a partition G. Given a $p \times K$ matrix $\widehat{U}$, the membership matrix $\widehat{A}$ is said to be an η-approximation K-means problem on $\widehat{U}$ if there exists a $K \times K$ matrix $\widehat{Q}$ such that

$$\|\widehat{U} - \widehat{A}\widehat{Q}\|_F^2 \leqslant \eta \min_{A \in \mathcal{A}p,k} \min_{Q} \|\widehat{U} - AQ\|_F^2\,. \tag{F.1}$$

The partition $\widehat{G}$ is then the partition corresponding to $\widehat{A}$.

The proof is based on the following Lemma by Lei and Rinaldo Lei and Rinaldo, 2015.

Lemma F.1. *Let M be any matrix of the form $M = AQ$ where $A \in \mathcal{A}_{p,K}$ is a membership matrix associated to G and $Q \in \mathbb{R}^{K \times q}$, and denote by δ the minimal distance between two rows of Q. Then, there exists a*

constant c_η, such that, for any matrix M' fulfilling $\|M - M'\|_F^2 < m\delta^2/c_\eta$, the classification of the rows of M' by an η-approximate K-means provides a clustering $\widehat{G}$ fulfilling

$$\bar{L}(\hat{G}, G) \leqslant c_\eta \frac{\|M - M'\|_F^2}{m\delta^2}. \tag{F.2}$$

We also need the following bounds

Lemma F.2. *We have $\|\Sigma\|_{op} \geqslant m\|C^*\|_{op}$ and $|\Gamma|_\infty \leqslant \|\Sigma\|_{op}$.*

In view of this Lemma, we obtain $\|\Sigma\|_{op} \geqslant m\|C^*\|_{op} \geqslant m\lambda_K(C^*)$. Then, condition (5.20) of Proposition 5.3 enforces that

$$\frac{Re(\Sigma) \vee \log(p)}{n} \leqslant 1/c_\eta^2 \,. \tag{F.3}$$

Let U be a $K \times p$ matrix which gathers the eigenvectors of AC^*A^t associated to the K leading eigenvalues. The associated eigenvectors are block constant. Therefore $U_0 = AQ_0$, and since $A^tA = mI$, the matrix $\sqrt{m}Q_0$ is orthogonal.

We apply Lemma F.1 with $M' = \widehat{U}$ and $M = U_0\widehat{O}$, where $\widehat{O}$ is a $K \times K$ orthogonal matrix to be chosen. We have $M = AQ$ with $\sqrt{m}Q = \sqrt{m}Q_0\widehat{O}$ orthogonal. In particular, the minimal distance between two rows of Q is $\delta = \sqrt{2/m}$. Lemma F.1 ensures that

$$\bar{L}(\hat{G}_S, G) \leqslant c_\eta \frac{\|\widehat{U} - U_0\widehat{O}\|_F^2}{2}, \tag{F.4}$$

whenever the right-hand side is smaller than 1. By Davis-Kahan inequality (e.g. Lei and Rinaldo, 2015), there exists an orthogonal matrix $\widehat{O}$ such that

$$\|\widehat{U} - U_0\widehat{O}\|_F^2 \leqslant \frac{8K\|\widetilde{\Sigma} - AC^*A^t\|_{op}^2}{m^2\lambda_K^2(C^*)} \,. \tag{F.5}$$

We can upper-bound the operator norm of $\widetilde{\Sigma} - AC^*A^t$ by

$$\|\widetilde{\Sigma} - AC^*A^t\|_{op} \leqslant \|\widehat{\Sigma} - \Sigma\|_{op} + \|\widehat{\Gamma} - \Gamma\|_{op} \,. \tag{F.6}$$

According to Theorem 9 in Koltchinskii and Lounici, 2017, there exists a constant $c > 0$ such that, with probability at least $1 - 1/p$

$$\|\widehat{\Sigma} - \Sigma\|_{op} \leqslant c_L\|\Sigma\|_{op}\left(\sqrt{\frac{Re(\Sigma)}{n}} \bigvee \frac{Re(\Sigma)}{n} \bigvee \sqrt{\frac{\log(p)}{n}} \bigvee \frac{\log(p)}{n}\right) \tag{F.7}$$

$$\leqslant c_L\|\Sigma\|_{op}\left(\sqrt{\frac{Re(\Sigma)}{n}} \bigvee \sqrt{\frac{\log(p)}{n}}\right), \tag{F.8}$$

where we used (F.3) in the second line.

Then, using that $\|\widehat{\Gamma}-\Gamma\|_{op} = |\widehat{\Gamma}-\Gamma|_\infty$ and Proposition A.1 together with $|\Gamma|_\infty \leqslant \|\Sigma\|_{op}$ (Lemma F.2 above), we obtain the inequality

$$\|\widetilde{\Sigma} - AC^*A^t\|_{op} \leqslant c_L\|\Sigma\|_{op}\left(\sqrt{\frac{Re(\Sigma)}{n}} \bigvee \sqrt{\frac{\log(p)}{n}}\right), \tag{F.9}$$

with probability at least $1-c/p$. So combining (F.4), with (F.5) and (F.9) we obtain the existence of $c'_\eta > 0$ such that we have

$$\bar{L}(\hat{G}_S, G) \leqslant \frac{c'_{\eta,L}K\|\Sigma\|^2_{op}}{m^2\lambda_K(C^*)^2}\left(\sqrt{\frac{Re(\Sigma)}{n}} \bigvee \sqrt{\frac{\log(p)}{n}}\right)^2, \tag{F.10}$$

with probability at least $1-c/p$, whenever the right-hand side is smaller than 1. The proof of Proposition 5 follows.

Proof of Lemma 5.1. We recall that $\widehat{V}$ is the $p \times K$ matrix stacking the K leading eigenvectors of $\widehat{\Sigma}$. We first prove that the matrix $\widehat{V}\widehat{V}^t$ is solution of the SDP in Lemma 2.

Let us write $\widehat{\Sigma} = \widetilde{V}\widetilde{D}\widetilde{V}^t$ for a diagonalisation of $\widehat{\Sigma}$ with $\widetilde{V}$ orthogonal and $\widetilde{D}_{11} \geqslant \ldots \geqslant \widetilde{D}_{pp} \geqslant 0$. We observe that $\langle\widehat{\Sigma}, B\rangle = \langle\widetilde{D}, \widetilde{V}^tB\widetilde{V}\rangle$, and that $B \in \overline{\mathcal{C}}$ iff $\widetilde{V}^tB\widetilde{V} \in \overline{\mathcal{C}}$ since the matrix $\widetilde{B} = \widetilde{V}^tB\widetilde{V}$ has the same eigenvalues as B. We observe also that $\widehat{V}\widehat{V}^t = \widetilde{V}\Pi_K\widetilde{V}^t$, where Π_K is the diagonal matrix, with 1 on the first K diagonal elements and 0 on the $p-K$ remaining ones. So proving that $\overline{B} = \widehat{V}\widehat{V}^t$ is solution of the SDP in Lemma 2 is equivalent to proving that

$$\Pi_K = \operatorname*{argmax}_{\widetilde{B}\in\overline{\mathcal{C}}}\langle\widetilde{D}, \widetilde{B}\rangle. \tag{F.11}$$

Let us prove this result.

To start with, we notice that

$$\sum_{k=1}^K \widetilde{D}_{kk} = \max_{0\leqslant\widetilde{B}_{kk}\leqslant 1;\ \sum_k \widetilde{B}_{kk}=K}\langle\widetilde{D}, \widetilde{B}\rangle. \tag{F.12}$$

Since the condition $I \succcurlyeq \widetilde{B} \succcurlyeq 0$ enforces $0 \leqslant \widetilde{B}_{kk} \leqslant 1$, we have $\overline{\mathcal{C}} \subset \{B : 0 \leqslant \widetilde{B}_{kk} \leqslant 1;\ \sum_k \widetilde{B}_{kk} = K\}$ and then

$$\max_{\widetilde{B}\in\overline{\mathcal{C}}}\langle\widetilde{D}, \widetilde{B}\rangle \leqslant \sum_{k=1}^K \widetilde{D}_{kk} = \langle\widetilde{D}, \Pi_K\rangle. \tag{F.13}$$

Hence Π_K is solution to the above maximisation problem and $\overline{B} = \widetilde{V}\Pi_K\widetilde{V}^t = \widehat{V}\widehat{V}^t$.

To conclude the proof, we notice that $\widehat{V}_{a:}\widehat{V}^t$ is an orthogonal transformation of $\widehat{V}_{a:}$, so we obtain the same results when applying a rotationally invariant clustering algorithm to the rows of $\widehat{V}$ and to the rows of $\widehat{V}\widehat{V}^t$. □

Proof of Lemma F.2. We claim that, for each group G_k, there is at most one negative value Γ_{aa} with $a \in G_k$. When it exists, this negative value also satisfies $|\Gamma_{aa}| \leqslant \min_{b\in G_k\setminus\{a\}}\Gamma_{bb}$.

Indeed, take any $a, b \in G_k$ such that $\Gamma_{aa} < 0$ (if it exists). Then, the corresponding 2×2 submatrix of Σ

$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} = \begin{pmatrix} C_{kk} - |\Gamma_{aa}| & C_{kk} \\ C_{kk} & C_{kk} + \Gamma_{bb} \end{pmatrix} \tag{F.14}$$

is positive semi-definite, which enforces $\Gamma_{bb} \geqslant |\Gamma_{aa}|$.

Let us now prove the first result of the Lemma. Since $A^t A = mI$, we have $\|\Sigma\|_{op} \geqslant \|A^t \Sigma A\|_{op}/m = \|mC + A^t \Gamma A/m\|_{op}$. Since $A^t \Gamma A/m$ is a diagonal matrix with entries $\sum_{a \in G_k} \Gamma_{aa}/|G_k|$, for $k = 1, \ldots, K$, we finally have

$$\|\Sigma\|_{op} \geqslant m\|C^*\|_{op} + \min_k \sum_{a \in G_k} \frac{1}{|G_k|} \Gamma_{aa} \geqslant m\|C^*\|_{op} \,, \tag{F.15}$$

where the second inequality is a consequence of the above claim. As for the second result, we observe that $|\Gamma|_\infty$ is achieved by some $\Gamma_{aa} > 0$. Then, $\|\Sigma\|_{op} \geqslant C^*_{k(a)k(a)} + \Gamma_{aa} \geqslant |\Gamma|_\infty$ since condition (5.20) of Proposition 5.3 enforces that C is positive. □

G Deviation inequalities

The following Lemma provides deviation inequalities for quadratic form of Sub-Gaussian and Bernstein-type random variables that we used in the proofs presented above. For the first inequality we use Rudelson and Vershynin, 2013, whereas the second one is proved in Bellec, 2014.

Lemma G.1 (Hanson-Wright Inequalities)**.** *There exists two positive constants c and c' such that the following holds for all $n \times n$ matrices A. Let $\xi = (\xi_1, \xi_2, \ldots, \xi_n)^t$ denote a vector of independent zero-mean random variables with respectives variances $\sigma_1^2, \ldots, \sigma_n^2$.*

(i) Assume that all variables ξ_i follow sub-Gaussian distributions, that for some $L > 0$, $\max_{i=1,\ldots,n} \|\xi_i\|_{\psi_2} \leqslant L$. Then, for all $t > 0$

$$\mathbb{P}\left[\xi^t A\xi - \mathbb{E}[\xi^t A\xi] > cL^2\big(\|A\|_2\sqrt{t} + c\|A\|_{op}t\big)\right] \leqslant e^{-t} \,. \tag{G.1}$$

(ii) Assume that, for some $M > 0$, ξ satisfies $\max_{i=1,\ldots,n} |\xi_i| \leqslant M$ almost surely. Then, for all $t > 0$

$$\mathbb{P}\left[\xi^t A\xi - \mathbb{E}[\xi^t A\xi] > c'\big(M\|AD_\sigma\|_2\sqrt{t} + cM^2\|A\|_{op}t\big)\right] \leqslant e^{-t} \,, \tag{G.2}$$

where $D_\sigma = \mathrm{Diag}(\sigma_1, \ldots, \sigma_n)$.

The following corollary will be useful for controlling cross-products of random variables.

Corollary G.1. *Let $\mathbf{X}$ denotes the observation matrix and let u, v be two p-dimensional vectors. Under Assumption 1, for any $t > 0$, with probability at least $1 - 4e^{-t}$, we have*

$$\Big|\frac{1}{n}\langle \mathbf{X}u, \mathbf{X}v\rangle - u^t \Sigma v\Big| \leqslant cL^2 \sqrt{u^t\Sigma u \, v^t \Sigma v}\left(\sqrt{\frac{t}{n}} + \frac{t}{n}\right). \tag{G.3}$$

Proof of Corollary G.1. Let us set $\alpha = (v^t\Sigma v/u^t\Sigma u)^{1/4}$. We first observe that

$$\langle \mathbf{X}u, \mathbf{X}v\rangle = \frac{1}{4}\left(\|\alpha\mathbf{X}u + \alpha^{-1}\mathbf{X}v\|^2 - \|\alpha\mathbf{X}u - \alpha^{-1}\mathbf{X}v\|^2\right). \tag{G.4}$$

Since for any $w \in \mathbb{R}^p$, we have $\|X^t w\|_{\psi_2} \leqslant L(w^t\Sigma w)^{1/2}$, Hanson-Wright inequality gives that with probability at least $1 - 4e^{-t}$, we have both

$$\begin{aligned} |\|\alpha\mathbf{X}u + \alpha^{-1}\mathbf{X}v\|^2 - \mathbb{E}\left[\|\alpha\mathbf{X}u + \alpha^{-1}\mathbf{X}v\|^2\right]| \leqslant \\ cL^2(\alpha u + \alpha^{-1}v)^t\Sigma(\alpha u + \alpha^{-1}v)(\sqrt{nt} + t) \end{aligned} \tag{G.5}$$

$$\begin{aligned} |\|\alpha\mathbf{X}u - \alpha^{-1}\mathbf{X}v\|^2 - \mathbb{E}\left[\|\alpha\mathbf{X}u - \alpha^{-1}\mathbf{X}v\|^2\right]| \leqslant \\ cL^2(\alpha u - \alpha^{-1}v)^t\Sigma(\alpha u - \alpha^{-1}v)(\sqrt{nt} + t). \end{aligned} \tag{G.6}$$

Combining the two bounds, we get

$$\left|\frac{1}{n}\langle \mathbf{X}u, \mathbf{X}v\rangle - u^t\Sigma v\right| \leqslant \frac{1}{2}cL^2\left(\alpha^2 u^t\Sigma u + \alpha^{-2}v^t\Sigma v\right)(\sqrt{t/n} + t/n). \tag{G.7}$$

Replacing α^2 by its value $\alpha^2 = (v^t\Sigma v/u^t\Sigma u)^{1/2}$, we get the desired result. □

We have a similar result under Assumption 1-bis.

Lemma G.2. *Let $\mathbf{X}$ denotes the observation matrix and let u, v be two p-dimensional vectors. Under Assumption 1-bis, for any $t > 0$, with probability at least $1 - 2e^{-t}$, we have*

$$\left|\frac{1}{n}\langle \mathbf{X}u, \mathbf{X}v\rangle - u^t\Sigma v\right| \leqslant 2M\left(\left(|v|_1\sqrt{u^t\Sigma u}\right) \wedge \left(|u|_1\sqrt{v^t\Sigma v}\right)\right)\sqrt{\frac{t}{n}} + 3M^2|u|_1|v|_1\frac{t}{n}. \tag{G.8}$$

Proof of Lemma G.2. The variable $Z_i = (X_i^t u)(X_i^t v) - u^t\Sigma v$ fulfills $|Z_i| \leqslant B = 2M^2|v|_1|u|_1$ and

$$\mathbb{E}\left[Z_i^2\right] \leqslant \mathbb{E}\left[\left(X_i^t u\, X_i^t v\right)^2\right] \leqslant M^2\min(|u|_1^2 v^t\Sigma v, |v|_1^2 u^t\Sigma u). \tag{G.9}$$

The inequality (G.8) then simply follows from Bernstein inequality. □

The two following Lemmas control the behavior of empirical covariance matrices both under sub-Gaussian and Bernstein-type distributional assumptions. The first one follows from Koltchinskii and Lounici, 2017; Vershynin, 2010, whereas the second one can be found in the expository paper Tropp, 2015.

Lemma G.3. *Consider a zero mean random vector Y of size p and invertible covariance matrix Σ. Assume that, for some $L > 0$, $\|\Sigma^{-1/2}Y\|_{\psi_2} \leqslant L$. Then, for some constant $c_L > 0$, the following holds. Given a n-sample $(Y_1, \ldots, Y_n)$, the empirical covariance matrix $\widehat{\Sigma} = \sum_{i=1}^n Y_i Y_i^t/n$ satisfies the following deviation inequality*

$$\mathbb{P}\left[\|\widehat{\Sigma} - \Sigma\|_{op} > c_L\|\Sigma\|_{op}\left[\sqrt{\frac{p}{n}} + \frac{p}{n} + \sqrt{\frac{t}{n}} + \frac{t}{n}\right]\right] \leqslant e^{-t}\,, \tag{G.10}$$

for any $t > 0$.

Lemma G.4 (Matrix Bernstein Inequality)**.** *Consider a zero mean random vector Y of size p and covariance matrix Σ. Assume that, for some $M > 0$, $\max_{i=1,\ldots,p} |Y_i| \leqslant M$ almost surely. Given a n-sample $(Y_1, \ldots, Y_n)$, the empirical covariance matrix $\widehat{\Sigma} = \sum_{i=1}^n Y_i Y_i^t / n$ satisfies the following deviation inequality*

$$\mathbb{P}\left[\|\widehat{\Sigma} - \Sigma\|_{op} > M\sqrt{2\frac{p\|\Sigma\|_{op}(t + \log(2p))}{n}} + \frac{4pM^2(t + \log(2p))}{3n}\right] \leqslant e^{-t}\,, \tag{G.11}$$

for any $t > 0$.

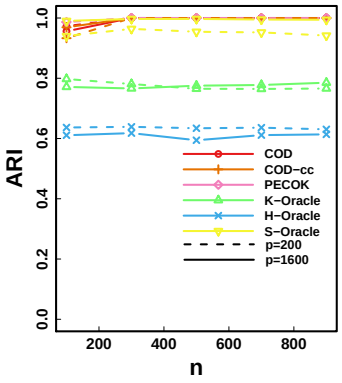
H Additional Simulation Results

Our work is devoted to exact cluster recovery. For completeness, we also assess the performance of our methods in terms of partial cluster recovery, and compare them with existing procedures. We employ the commonly used Adjusted Rand Index (ARI) to compare different methods. The largest ARI value is 1 when the recovered clusters match the truth exactly. The simulated data are the same as those illustrated in Figure 2.1 and 2.2. Figure 2.4 and 2.5 compares the average ARI values under different simulation settings. Our methods, COD, COD-CC, and PECOK, have their ARI measures close to 1 across all simulation settings, outperforming all other methods. Their performance also increases with n as predicted by our theory, while all other competing methods have flat ARI measures even if n increases. Under varying m, PECOK yields close to 1 ARI values for small $n = 60, 80, 100, 150$.

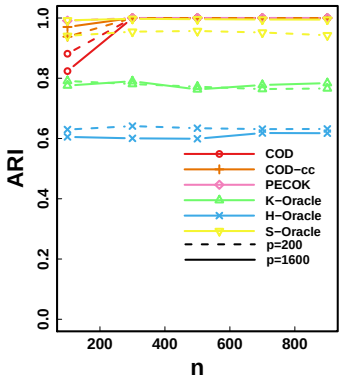
I Supplemental Materials for the fMRI Example

Preprocessing. We applied the preprocessing steps suggested by Xue, Aron, and Poldrack, 2008, which includes slice timing correction, alignment, registration, normalization to the average 152 T1 MNI template, smoothing with a 5mm full-width-half-maximum Gaussian kernel, denoising using the FSL MELODIC procedure, and a high pass filter with a 66s cut-off. The event-related activation and temporal correlation were removed using general linear models (GLM) for each voxel Friston et al., 1994. Following Power et al., 2011, we extract 180 mean activities within a 10mm spheres centered around each of 264 putative functional areas (see Table S2 from Power et al., 2011).

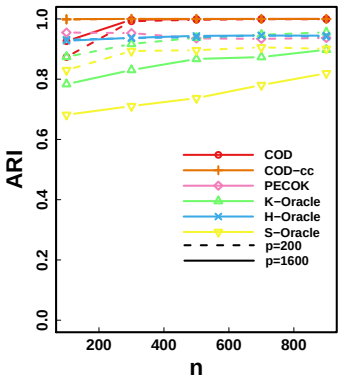
Figure 2.4 – Average ARI values by K-means (K-Oracle, medium green lines, triangle points), HC (H-Oracle, light blue lines, cross points), spectral clustering (S-Oracle, light yellow lines, upside-down triangle points), COD (dark red lines, circle points), COD-CC (light orange lines, plus points), PECOK (light pink lines, diamond points).



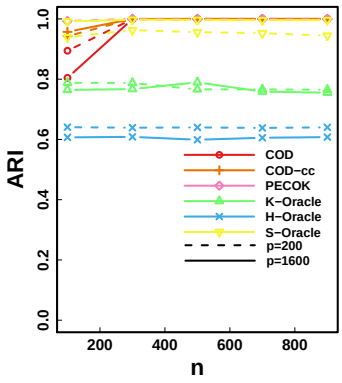
(a) M1



(b) M2

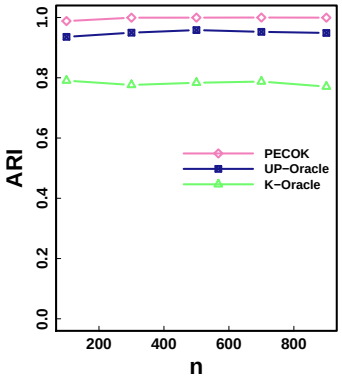


(c) M1S

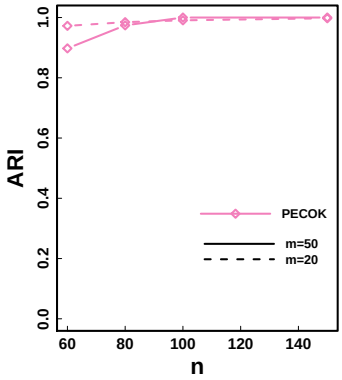


(d) M1P

Figure 2.5 – (a) The parameter K in K-means and Γ-uncorrected PECOK (UP-Oracle, navy blue lines, square points) are set to the true K while PECOK selects K based on our CV criterion. (b) Average ARI values of PECOK are shown under varying m.



(a) With/without correcting for Γ



(b) Varying m

Chapter 3

Adaptive Clustering through Semidefinite Programming

Contents

Ce chapitre a fait l'objet d'une publication Royer, 2017 dans *Advances in Neural Information Processing Systems 30* (NIPS 2017).

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 near-optimal in recovering the partition. Its performances are shown to be adaptive to the problem's effective dimension.

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 Azizyan, Singh, and Wasserman, 2013; Verzelen and Arias-Castro, 2014).

A famous early approach to clustering is to solve for the geometrical estimator K-means Steinhaus, 1956; Lloyd, 1982; MacQueen, 1967. 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 and Raftery, 2002 and Dasgupta and Schulman, 2007. 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 and Wei, 2007 show that the K-means objective can be relaxed into a convex, semidefinite program, leading Mixon, Villar, and Ward, 2016 to use this relaxation under a subgaussian mixture model to estimate the cluster centers. Yan and Sarkar, 2016 use a similar semidefinite program in the context of covariate clustering, when the network has nodes and covariates. Chrétien, Dombry, and Faivre, 2016 use a slightly different form of a semidefinite program to recover the adjacency matrix of the cluster graph with high probability. Lastly in the different context of variable clustering, Bunea et al., 2016 present a semidefinite program with a correction step to produce non-asymptotic exact recovery results.

In this work, we build upon the work and context of Bunea et al., 2016, and transpose and adapt their ideas for point clustering: we introduce a semidefinite estimator for point clustering inspired by the findings of Peng and Wei, 2007 with a correction component originally presented in Bunea et al., 2016. We show that it produces a very strong contender for clustering recovery in terms of speed, adaptivity and robustness to model perturbations. In order to do so we produce a flexible probabilistic model inducing an optimal partition of the data that we aim to recover. Using the same structure of proof in a different context, we establish elements of stochastic control (see for instance Lemma A.1 on the concentration of random subgaussian Gram matrices in the supplementary material) to derive conditions of exact clustering recovery with high probability and show optimal performances – including in high dimensions, improving on Mixon, Villar, and Ward, 2016, as well as adaptivity to the effective dimension of the problem. We also show that our results continue to hold without knowledge of the number of structures given one single positive tuning parameter.

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 $M \in \mathbb{R}^{d \times d'}$, $|M|_q$, $|M|_F$ and $|M|_{op}$ are respectively the entry-wise l_q-norm, the Frobenius norm associated with scalar

product $\langle .,. \rangle$ 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 2.1. *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 2.1. *Let $(\mathcal{G}_K^*, \boldsymbol{\mu}^*, \delta^*) \in \operatorname{argmax} \rho(\mathcal{G}, \boldsymbol{\mu}, \delta)$ over all $(\mathcal{G}, \boldsymbol{\mu}, \delta)$ such that $X_1, \ldots, X_n$ 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 2.1, $\mathcal{G}$ 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 Vershynin, 2012), we use a more flexible definition that can account for anisotropy.

Definition 2.2. *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 2.1. *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 2.1 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 and Wei, 2007 shows that the K-means estimator $\bar{B}$ can be expressed as

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

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:

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

Notice that $\mathbb{E}\,\widehat{\Lambda} = \Lambda + \Gamma$ for $\Lambda := (\langle \nu_a, \nu_b \rangle)_{(a,b) \in [n]^2} \in \mathbb{R}^{n \times n}$, where we note $\Gamma := \mathbb{E}\left[\langle E_a, E_b \rangle\right]_{(a,b) \in [n]^2} = \text{diag}\left(\text{tr}(\text{Var}(E_a))\right)_{1 \leqslant a \leqslant n} \in \mathbb{R}^{n \times n}$. 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 3.1. *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*

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

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 2.1. 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 3.2. *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*

$$B^* \notin \operatorname*{argmax}_{B \in \mathcal{C}_K^{\{0,1\}}} \langle \Lambda + \Gamma, B \rangle \quad \textit{and} \quad B^* \notin \operatorname*{argmax}_{B \in \mathcal{C}_K} \langle \Lambda + \Gamma, B \rangle. \tag{3.7}$$

So Proposition 3.2 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 $\Gamma = \operatorname{diag}(\operatorname{tr}(\operatorname{Var}(E_a)))_{1 \leqslant a \leqslant n}$.

Suppose we produce such an estimator $\widehat{\Gamma}^{corr}$. Then substracting $\widehat{\Gamma}^{corr}$ from $\widehat{\Lambda}$ can be interpreted as a correcting term, i.e. a way to de-bias $\widehat{\Lambda}$ as an estimator of Λ. Hence the previous results demonstrate the interest of studying the following semi-definite estimator of the projection matrix B^*, let

$$\widehat{B}^{corr} := \operatorname*{argmax}_{B \in \mathcal{C}_K} \langle \widehat{\Lambda} - \widehat{\Gamma}^{corr}, B \rangle. \tag{3.8}$$

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 2.1 there exist $\Sigma_1, ..., \Sigma_n$ such that $\forall a \in [n]$, $X_a \sim \text{subg}(\Sigma_a)$. Let

$$\sigma^2 := \max_{a \in [n]} |\Sigma_a|_{op}, \quad \mathcal{V}^2 := \max_{a \in [n]} |\Sigma_a|_F, \quad \gamma^2 := \max_{a \in [n]} \operatorname{tr}(\Sigma_a) \tag{3.9}$$

We now produce $\widehat{\Gamma}^{corr}$. Since there is no relation between the variances of the points in our model, there is very little hope of estimating $\operatorname{Var}(E_a)$. As for our quantity of interest $\operatorname{tr}(\operatorname{Var}(E_a))$, a form of volume, a rough estimation is challenging but possible. The estimator from Bunea et al., 2016 can be adapted to our context. For $(a, b) \in [n]^2$ let $V(a, b) := \max_{(c,d) \in ([n] \setminus \{a,b\})^2} \left| \langle X_a - X_b, \frac{X_c - X_d}{|X_c - X_d|_2} \rangle \right|$, $\widehat{b}_4 := \operatorname{argmin}_{b \in [n] \setminus \{a\}} V(a, b)$ and $\widehat{b}_{4'} := \operatorname{argmin}_{b \in [n] \setminus \{a, \widehat{b}_4\}} V(a, b)$. Then for $a \in [n]$, let

$$\widehat{\Gamma}^{corr} := \operatorname{diag}\left(\langle X_a - X_{\widehat{b}_4}, X_a - X_{\widehat{b}_{4'}} \rangle_{a \in [n]} \right). \tag{3.10}$$

Proposition 3.3. *Assume that $m > 2$. For $c_6, c_7 > 0$ absolute constants, with probability larger than $1 - c_6/n$ we have*

$$|\widehat{\Gamma}^{corr} - \Gamma|_\infty \leqslant c_7 \Big(\sigma^2 \log n + (\delta + \sigma\sqrt{\log n})\gamma + \delta^2 \Big). \tag{3.11}$$

So apart from the radius δ terms, that come from generous model assumptions, a proxy for Γ is produced at a $\sigma^2 \log n$ rate that we could not expect to improve on. Nevertheless, this control on Γ is key to attain the optimal rates below. It is general and completely independent of the structure of $\mathcal{G}$, as there is no relation between $\mathcal{G}$ and Γ.

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

$$m\Delta^2(\boldsymbol{\mu}) \geqslant c_2\big(\sigma^2(n + m\log n) + \mathcal{V}^2(\sqrt{n + m\log n}) + \gamma(\sigma\sqrt{\log n} + \delta) + \delta^2(\sqrt{n} + m)\big), \tag{3.12}$$

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

We call the right hand side of (3.12) 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 (3.12) 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 3.1? 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 (3.12) becomes

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

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:

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

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 Mixon, Villar, and Ward, 2016, 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 Chen and Xu, 2016).
In the high-dimensional setup $n \vee m\log n \lesssim p$, condition (3.13) becomes:

$$\frac{\Delta^2(\boldsymbol{\mu})}{\sigma^2} \gtrsim \sqrt{(K + \log n)\frac{pK}{n}}. \tag{3.15}$$

There are few information-theoretic bounds for high-dimension clustering. Recently, Banks et al., 2018 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 (3.12) by introducing an effective dimension r_*, measuring the largest volume repartition for our variance-bounding matrices $\Sigma_1, ..., \Sigma_n$. We will show that our estimator adapts to this effective dimension. Let

$$r_* := \frac{\gamma^2}{\sigma^2} = \frac{\max_{a\in[n]} \operatorname{tr}(\Sigma_a)}{\max_{a\in[n]} |\Sigma_a|_{op}}, \tag{3.16}$$

r_* can also be interpreted as a form of global effective rank of matrices Σ_a. Indeed, define $Re(\Sigma) := \operatorname{tr}(\Sigma)/|\Sigma|_{op}$, 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 (3.12) can be written as

$$\frac{\Delta^2(\boldsymbol{\mu})}{\sigma^2} \gtrsim \big(K + \log n\big) + \sqrt{(K + \log n)\frac{r_* K}{n}}. \tag{3.17}$$

By comparing this equation to (3.13), notice that r_* is in place of p, indeed playing the role of an effective dimension for the problem. This shows that our estimator adapts to this effective dimension, without the use of any dimension reduction step. In consequence, equation (3.17) 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 (3.14).
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, (3.17) 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:

$$\widetilde{B}^{corr} := \operatorname*{argmax}_{B\in\mathcal{C}} \langle \widehat{\Lambda} - \widehat{\Gamma}^{corr}, B\rangle - \widehat{\kappa} \operatorname{tr}(B). \tag{4.1}$$

Theorem 4.1. *Suppose that $m > 2$ and* (3.12) *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 (3.12). 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 Conclusion

In this paper we analyzed a new semidefinite positive algorithm for point 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 also demonstrated that our estimator can be optimally adapted to a data-driven choice of K, with a single tuning parameter. Lastly we illustrated on high-dimensional experiments that our approach is empirically stronger than other classical clustering methods. The $\widehat{\Gamma}^{corr}$ correction step of the algorithm, it can be interpreted as an independent, denoising step for the Gram matrix, and we recommend using such a procedure where the probabilistic framework we developed seems appropriate.

In practice, it is generally more realistic to look at approximate clustering results, but in this work we chose the point of view of exact clustering for investigating theoretical properties of our estimator. Our experimental results provide evidence that this choice is not restrictive, i.e. that our findings translate very well to approximate recovery. We expect our results to hold with similar speeds for approximate clustering, up to some logarithmic terms. One could think of adapting works on community detection by Guédon and Vershynin, 2016 based on Grothendieck's inequality, or work by (Fei and Chen, 2017) from the stochastic-block-model community on similar semidefinite programs. In fact, referring to a detection bound by Banks et al., 2018, our only margin for improvement on the separation speed is to transform the logarithmic factor $\sqrt{\log n}$ into $\sqrt{\log K}$ when the number of clusters K is of order $O(\log n)$ – otherwise the problem is rather open.

As for the robustness of this procedure, a few aspects are to be considered: the algorithm we studied solves for a convexified objective, therefore its performances are empirically more stable than that of an objective that would prove non-convex, especially in the high-dimensional context. In this work we also benefit from a permissive probabilistic framework that allows for multiple deviations from the classical gaussian cluster model, and come at no price in terms of the performance of our estimator. Points from a same cluster are allowed to have significantly different means or fluctuations, and the results for exact recovery with high probability are unchanged, near-minimax and adaptive.

Appendix

A Intermediate results

Generic controls for exact recovery

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

Theorem A.1. *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 $\widetilde{B} := \operatorname{argmax}_{B \in \mathcal{C}} \langle \widehat{\Lambda} - \widehat{\Gamma}, B \rangle - \widehat{\kappa}\,\mathrm{tr}(B)$ for $\widehat{\kappa} \in \mathbb{R}$.

Theorem A.2. *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 $1 - c_3/n$

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 A.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 $\forall t \geqslant 0$*

$$\mathbb{P}\left[|\mathbf{E}\mathbf{E}^T - \mathbb{E}\left[\mathbf{E}\mathbf{E}^T\right]|_{op} \geqslant 2\max_{a\in[n]}|\Sigma_a|_F\sqrt{t} + 2\max_{a\in[n]}|\Sigma_a|_{op}t\right] \leqslant 9^n 2e^{-c_* t}. \tag{A.3}$$

B Main proofs

B.1 Proof of Proposition 2.1: 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 3.1 (respectively Theorem 4.1) is a composition of Theorem A.1 (respectively Theorem A.2) and Proposition 3.3.
In this section, under Hypothesis 2.1, we have $\forall k \in [K], \forall a \in G_k : X_a \sim \text{subg}(\Sigma_a)$. For $k \in [K]$, we define $\sigma_k^2 := \max_{a \in G_k} |\Sigma_a|_{op} \leqslant \sigma^2$, as well as $\mathcal{V}_k^2 := \max_{a \in G_k} |\Sigma_a|_F \leqslant \mathcal{V}^2$, $\gamma_k^2 := \max_{a \in G_k} \text{tr}(\Sigma_a) \leqslant \gamma^2$.
A number of proofs in this section are adapted from the proof ensemble of Bunea et al., 2016. 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.

Proof of Theorem A.1

In this theorem we only need to consider $B \in \mathcal{C}_K$, but the proof of Theorem A.2 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 B.1. *Proving* (B.1) *reduces to proving*

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

The proof for Lemma B.1 is found in section B.2. 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:

Lemma B.2. *For $c_2' > 0$ absolute constant, with probability greater than $1 - 1/n$:*

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

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

Lemma B.3. *For any symmetric matrix $W \in \mathbb{R}^{n\times n}$ we have:*

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

The proof for Lemma B.3 will be found in Bunea et al., 2016, 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 $1 - 1/n$, $\forall B \in \mathcal{C}$:

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

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

For the term W_4 we introduce the following Lemma, proved in section B.2:

Lemma B.4. *For $c_4', c_4'' > 0$ absolute constants, with probability larger than $1 - 2/n$:*

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

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:

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

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 $1 - c_1'/n$: $\forall B \in \mathcal{C}$,

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

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 B.3 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:

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

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 Bunea et al., 2016 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 $1 - c_1/n$:

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

This concludes the proof for Theorem A.1. □

Proof of Theorem A.2: 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 A.2 we can follow the same proof as in Theorem A.1 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 $1 - c_1'/n$: $\forall B \in \mathcal{C}$,

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

Using the assumption (A.1) of Theorem A.2 there exist $c_2' > 0$ such that with probability greater than $1 - c_1'/n$: $\forall B \in \mathcal{C}$,

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

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 Bunea et al., 2016 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 A.2. □

Proof of Lemma B.1

$$(\widehat{\Lambda} - \widehat{\Gamma})_{ab} = \langle X_a, X_b\rangle - \widehat{\Gamma}_{ab} = \langle \nu_a, \nu_b\rangle + \langle \nu_a, E_b\rangle + \langle \nu_b, E_a\rangle + \langle E_a, E_b\rangle - \widehat{\Gamma}_{ab} \tag{B.16}$$

$$= \langle \nu_a, \nu_b\rangle + \langle \nu_a - \nu_b, E_b - E_a\rangle + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} \tag{B.17}$$

$$= \langle \nu_a, \nu_b\rangle + \langle \mu_k - \mu_l, E_b - E_a\rangle + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} \tag{B.18}$$

$$= -\langle \mu_k, \mu_l\rangle + \langle \nu_a - \mu_k, \nu_b - \mu_l\rangle + \langle \nu_a, \mu_l\rangle + \langle \mu_k, \nu_b\rangle + \langle \mu_k - \mu_l, E_b - E_a\rangle + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} \tag{B.19}$$

$$= -(S_1)_{ab} - \frac{1}{2}(|\mu_k|_2^2 + |\mu_l|_2^2) + (W_1)_{ab} + \langle \nu_a, \mu_l\rangle + \langle \mu_k, \nu_b\rangle + \langle \mu_k - \mu_l, E_b - E_a\rangle + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} \tag{B.20}$$

$$= -(S_1)_{ab} - \frac{1}{2}(|\mu_k|_2^2 + |\mu_l|_2^2) + (W_1)_{ab} + \langle \nu_a, \mu_k\rangle + \langle \mu_l, \nu_b\rangle + \langle \mu_k - \mu_l, \nu_b - \nu_a + E_b - E_a\rangle + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} \tag{B.21}$$

$$= -(S_1)_{ab} - \frac{1}{2}(|\mu_k|_2^2 + |\mu_l|_2^2) + (W_1)_{ab} + \langle \nu_a, \mu_k\rangle + \langle \mu_l, \nu_b\rangle + 2(S_1)_{ab} + (W_2)_{ab} + (W_3)_{ab} + \langle \nu_a, E_a\rangle + \langle \nu_b, E_b\rangle + (W_4 + W_5)_{ab} \tag{B.22}$$

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

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

□

Proof of Lemma B.2: 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:

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

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 C.1 $\langle E_a - E_b, z\rangle$ 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 Vershynin, 2012), there exist $c_* > 0$ absolute constant such that $\forall t \geqslant 0$:

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

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 $t = \sqrt{(1+3\log n)/c_*}$. □

Proof of Lemma B.4: control of $|\langle W_4, B^* - B\rangle|$

Recall $(W_4)_{ab} = \langle E_a, E_b\rangle - \Gamma_{ab}$. We will prove Lemma B.4 by using the derivation of (B.6) combined with Lemma A.1 for control of the operator norm and the following Lemma for the remaining part.

Lemma B.5. *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:

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

The bound for u_b uses Lemma C.3: $\forall t \geqslant 0\ \mathbb{P}\left[||E_b|_2^2 - \mathbb{E}\,|E_b|_2^2| \geqslant \mathcal{V}_l^2\sqrt{t} + \sigma_l^2 t\right] \leqslant 2e^{-c_* t}$ so only the scalar product remains to be controlled. Notice that by Lemma C.1, $\sqrt{|G_k|}\widetilde{E}_k$ 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 C.3 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 A.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 B.3 to W_4:

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

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

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

This concludes the proof for Lemma B.4. □

B.3 Proof of Proposition 3.3, control of $\widehat{\Gamma}^{corr}$

Let $a \in G_k$, $b_1 \in G_{l_1}, b_2 \in G_{l_2}$, using decomposition (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 C.3 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 $Y = \Sigma_b^{-1/2} E_b \sim$ subg(I_p) and $A = \Sigma_b^{1/2^T}(zz^T)\Sigma_b^{1/2}$, so that we have: $\langle z, E_b\rangle^2 = E_b^T zz^T E_b = Y^T AY$. Because $|z|_2 = 1$ and zz^T is symmetric of rank 1 we have $|A|_F = |A|_{op} = \mathrm{tr}(A) \leqslant \sigma^2$ therefore we use Lemma C.2 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 C.3 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 $(a_1, a_2) \in G_k\backslash\{a\}$, $(c, d) \in ([n]\backslash\{a, a_1, a_2\})^2$, 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 C.1, $\langle E_a - E_{a_u}, Z\rangle$ 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:

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

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:

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

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 C.1, $E_c - E_d$ is subg$(\Sigma_c + \Sigma_d)$. We have $\mathbb{E}\left[|E_c - E_d|_2^2\right] \leqslant \mathrm{tr}(\Sigma_c + \Sigma_d) \leqslant 2\gamma^2$, $|\Sigma_c + \Sigma_d|_F \leqslant 2\mathcal{V}^2 \leqslant 2\sigma\gamma$ and $|\Sigma_c + \Sigma_d|_{op} \leqslant 2\sigma^2$. Therefore by the first inequality of Lemma C.3 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 $1 - 1/n^2$:

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

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

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

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 $1 - 4/n^2$:

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

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 3.3 □

B.4 Proof of Proposition 3.1

For this proof we rely heavily on the proof of Theorem A.1: 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 A.1, that is using (B.4), (B.7) and (B.12), we have $\forall B \in \mathcal{C}_K$:

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

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

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

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

Thus we have:

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

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 $\operatorname{argmax}_{B \in \mathcal{C}_K} \langle \Lambda + \Gamma, B \rangle = B^*$. Lastly as B^* is in $\mathcal{C}_K^{\{0,1\}} \subset \mathcal{C}_K$, this concludes the proof. □

B.5 Proof of Proposition 3.2

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

$$B^* := \begin{bmatrix} \boxed{\frac{1}{m}} & & \\ & \boxed{\frac{1}{m}} & \\ & & \boxed{\frac{1}{m}} \end{bmatrix}, \quad B_1 := \begin{bmatrix} \boxed{2/m} & & \\ & \boxed{2/m} & \\ & & \boxed{\frac{1}{2m}} \end{bmatrix}$$

  matrices 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 C.1. *$\forall a \in [n]$ let $Y_a \sim subg(\Sigma_a)$, independent, $\Sigma_a \in \mathbb{R}^{d \times d}$ then*

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

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

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

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

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

□

Lemma C.2. *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 $t \geqslant 0$

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

Proof. A variation of the original Hanson-Wright inequality (Theorem 1.1 from Rudelson and Vershynin, 2013), it holds as $\sigma = 1$ bounds the subgaussian norm $|Y|_{\Psi_2} := \sup_{x \in \mathcal{S}_{d-1}} \sup_{p \geqslant 1} p^{-1/2} (\mathbb{E}\,|x^T Y|^p)^{1/p}$, a consequence of Lemma 5.5 from Vershynin, 2012. □

Lemma C.3. *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 C.2 with $Y = \Sigma^{-1/2}E$ and $A = \Sigma$. As for the second inequality, by Lemma C.1 we have $Y = (E^T\Sigma^{-1/2}, E'^T\Sigma'^{-1/2T})^T \sim \text{subg}(I_{2d})$. Then let us use Lemma C.2 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 A.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 Rigollet, 2015, let $\mathcal{N}$ be a 1/4-net for $\mathcal{S}_{n-1}$ such that $|\mathcal{N}| \leqslant 9^n$ (see Lemma 5.2 Vershynin, 2012), 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

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

Notice that this rewrites

$$x^TWy = \sum_{a=1}^n\sum_{b=1}^n x_a(E_a^TE_b - \Gamma_{ab})y_b \tag{C.7}$$

$$= (\sum_{a=1}^n E_a^Tx_a)(\sum_{b=1}^n E_b^Ty_b)^T - \mathbb{E}(\sum_{a=1}^n E_a^Tx_a)(\sum_{b=1}^n E_b^Ty_b)^T. \tag{C.8}$$

For $x, y \in \mathcal{N}^2$, let us define $x\otimes\Sigma^{1/2} := (x_1\Sigma_1^{1/2}, ..., x_n\Sigma_n^{1/2})^T \in \mathbb{R}^{np\times p}$ 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 C.1 we have $Y \sim \text{subg}(I_{np})$). We have

$$x^TWy = Y^T(x\otimes\Sigma^{1/2})(y\otimes\Sigma^{1/2})^TY - \mathbb{E}[Y^T(x\otimes\Sigma^{1/2})(y\otimes\Sigma^{1/2})^TY]. \tag{C.9}$$

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: $|(x\otimes\Sigma^{1/2})(y\otimes\Sigma^{1/2})^T|_F^2 = \sum_{a=1}^n\sum_{b=1}^n x_a^2y_b^2|\Sigma_a^{1/2}\Sigma_b^{1/2}|_F^2 \leqslant \max_{a\in[n]}|\Sigma_a|_F^2$. Therefore using Lemma C.2 on Y we have

$$\forall t \geqslant 0 : \mathbb{P}\left[|Y^TAY - \mathbb{E}\left[Y^TAY\right]| \geqslant \max_{a\in[n]}|\Sigma_a|_F\sqrt{t} + \max_{a\in[n]}|\Sigma_a|_{op}t\right] \leqslant 2e^{-ct} \tag{C.10}$$

Hence in conjunction with (C.6) we conclude the proof. □

Chapitre 4

Partitionnement et optimisation

Contents

Dans cette partie, on présente une discussion introductive aux approches état-de-l'art pour les solutions des programmes semi-défini positifs (SDPs), puis on propose des algorithmes approchés présentant une moindre complexité. Enfin on évalue numériquement ces différentes approches.

Notre discussion est orientée par la question de la résolution du programme :

Définition 0.1 (K-MEANS SDP)**.** *Soit $A \in \mathbb{R}^{N\times N}$ une matrice des données, résoudre*

$$\underset{B \text{ sym.}}{\operatorname{argmin}} \langle A, B\rangle \text{ s.t. } \begin{array}{l} \bullet\ B.1 = 1, \operatorname{tr}(B) = K \\ \bullet\ B \geqslant 0 \\ \bullet\ B \succcurlyeq 0 \end{array} . \tag{0.1}$$

On a vu que ce programme était le produit d'une relaxation convexe des K-moyennes (2.1) (voir page 5). C'est un programme SDP qui possède la particularité d'être lourdement contraint, du fait principal des contraintes de positivité $B \geqslant 0$. On parlera de **SDP à contraintes positives**.

1 Partitionnement et optimisation SDP

L'optimisation semi-défini positive concerne les problèmes d'optimisation d'une fonction convexe sur le cône semi-défini positif et sous contraintes linéaires, confère Vandenberghe et Boyd, 1996. Elle est souvent employée pour relâcher des problèmes *durs*, à l'instar de problèmes d'optimisation linéaire en nombres entiers, en problèmes convexes qu'on peut donc résoudre en temps polynomiaux. On s'intéressera plus particulièrement à ce type de programmes :

Définition 1.1 (Programme semi-défini positif à m contraintes d'égalité)**.** *Pour A matrice réelle symétrique, $b \in \mathbb{R}^m$, $\mathcal{A} : \mathbb{R}^{N\times N} \to \mathbb{R}^m$ un opérateur linéaire (capturant m contraintes d'égalité), résoudre :*

$$\min_B \langle A, B\rangle \quad t.q. \quad B \succcurlyeq 0, \mathcal{A}(B) = b \tag{1.1}$$

1.1 Solveurs pour l'optimisation SDP

Les problèmes faisant intervenir des programmes comme (0.1), (1.1) ne manquent pas : on peut voir remonter les premières applications à l'optimisation de graphes avec Lovasz, 1979, et plus concrètement au problème de coupe maximum avec Goemans et Williamson, 1995. Pour le problème de la coupe normalisée, Xing et Jordan, 2003 présentent des résultats de contrôle d'une relaxation SDP qui améliorent ceux d'algorithmes plus relâchés de partitionnement spectraux (mais les auteurs ne constataient pas en pratique d'amélioration effective pour le partitionnement). Plus proche de notre problématique encore, Alon et Naor, 2006 analysent les relaxations SDPs en utilisant l'inégalité de Grothendieck, grâce à laquelle Guédon et Vershynin, 2016 montrent des contrôles théoriques en détection de communauté, qui seront suivis par les travaux de Chrétien, Dombry et Faivre, 2016 ; Mixon, Villar et Ward, 2016 pour le partitionnement de points dans les modèles de mélanges. Dans Chrétien, Dombry et Faivre, 2016, les auteurs étudient un programme proche de (0.1) à trace contrainte, problème dont Helmberg et Rendl, 2000 ont donné une reformulation en problème d'optimisation de valeurs propres et une méthode du premier ordre pour résoudre celui-ci.

Résoudre en généralité l'optimisation (1.1) n'a rien d'évident. Une méthode du premier ordre est la méthode des multiplicateurs à directions alternées Boyd et al., 2011, ou ADMM, déjà utilisée en détection de communauté par Cai et Li, 2015 ; Amini et Levina, 2018. Essentiellement, il s'agit d'encoder la contrainte conique avec une autre variable primale, soit de ré-écrire (1.1) comme

$$\min_{(B,C):B=C} \langle A, B\rangle + \delta_{C\succcurlyeq 0}{}^{1} \quad \text{t.q.} \quad \mathcal{A}(B) = b, \tag{1.2}$$

puis d'introduire le lagrangien augmenté qu'on optimisera itérativement et alternativement selon les directions de descente (ici B et C). L'intérêt d'utiliser ADMM est qu'on peut exploiter la séparabilité de l'objectif selon les différentes contraintes et ainsi résoudre séparément, donc à moindre coût, les minimisations dans les différentes directions. Autrement dit les calculs peuvent être distribués, ce qui aura un intérêt particulier lorsqu'on voudra faire de la factorisation (voir plus bas). Naturellement cette méthode ne permet pas de diminuer directement le principal coût par itération $\mathcal{O}(N^3)$ des calculs, qui provient de la contrainte de semi-défini positivité. Pour paraphraser ses concepteurs, ADMM est une méthode standard d'optimisation, modulaire, située

[0] δ indique l'ensemble avec la convention $\delta_{C\succcurlyeq 0} = 0$ si $C \succcurlyeq 0$, et $+\infty$ sinon

à un niveau plutôt haut conceptuellement. Ici elle pourra servir d'étalon, c'est une méthode simple à implémenter, qui convergera globalement mais lentement, et dont on s'attend à ce que les performances génériques puissent être battues par une méthode plus spécialisée.

L'essor du domaine de l'optimisation convexe est le fruit du développement des méthodes de points intérieurs (voir Nesterov et Nemirovskii, 1994) dans les années 80 et 90, qui donnent le moyen d'analyser et de résoudre les programmes comme (0.1) et (1.1) en temps polynomial. En principe ces méthodes partent d'un point admissible du programme, et optimisent dans une direction obtenue grâce à une méthode du second ordre (qui nécessite d'évaluer le hessien), en s'assurant de rester admissible (donc de rester à *l'intérieur* du cône) par une pénalisation adéquate des bords du cône (méthode de barrière logarithmique). Bien que ces méthodes soient très efficaces en tant qu'algorithme d'optimisation appliqué à un problème convexe conique, elles sont aussi connues pour être très coûteuses en temps et en mémoire : pour les algorithmes de chemin central primal-dual, le coût par itération sera de l'ordre de $\mathcal{O}(N^3)$ pour une convergence en $\mathcal{O}(\sqrt{N})$ itérations. Dans le contexte du K-MEANS SDP (0.1), la matrice des données $A \in \mathbb{R}^{N\times N}$ (matrice de Gram ou matrice de covariance empirique) n'est pas creuse en général, et le programme est lourdement contraint avec $N(N+3)/2$ contraintes en plus de la contrainte conique ; dès lors les bons solveurs utilisant ces méthodes comme Andersen et al., 2011 ne parviennent à résoudre des problèmes où N excède quelques dizaines.

On peut lire ce passage éclairant les différences entre les deux approches dans Yang, Sun et Toh, 2014 page 334 :

> While there has been a recent focus on using first order methods such as those based on ADMM or accelerated proximal gradient methods to solve structured convex optimization problems arising from machine learning and statistics, the extensive numerical results we obtained here for matrix conic programming problems serve to demonstrate that second order methods with good local convergence property are essential, if used wisely, for mitigating the inherent slow local convergence of first order methods, especially on difficult problems.

Yang, Sun et Toh, 2014 présente la méthode SDPNAL+, qui réalise une optimisation approchée à l'aide d'une méthode de Newton mélangée à une méthode de gradient conjugué, employée après une étape d'initialisation par ADMM. Mixon, Villar et Ward, 2016 utilise cette méthode pour résoudre le K-MEANS SDP (0.1) dans des problèmes de taille $N = 10^3$.

Il semble que pour optimiser le K-MEANS SDP (0.1) en grande dimension, on doive choisir entre une convergence trop lente si on utilise une méthode du premier ordre, ou trop coûteuse avec une méthode du second ordre. C'est Charybde et Scylla.

1.2 Factorisation de faible rang

Des travaux ont cherché à mieux utiliser la structure intrinsèque du problème. Burer et Monteiro, 2003 exploite une caractérisation des matrices semi-défini positives : il introduit le changement de variable $B = VV^T$ pour V matrice réelle dans $\mathbb{R}^{N\times p}$ et montre que le problème (1.1) est équivalent à

$$\min_V \langle A, VV^T\rangle \quad \text{t.q.} \quad \mathcal{A}(VV^T) = b \tag{1.3}$$

pour p tel que $p(p+1)/2 \geqslant m$. Si p est suffisamment petit, cette factorisation peut engendrer une importante réduction de la dimension de l'espace de recherche : c'est la **factorisation de faible rang**.

Par cette méthode de factorisation, on s'affranchit de la pénible contrainte $B \succcurlyeq 0$ tout en introduisant de la non-linéarité, c'est-à-dire qu'on a échangé des bonnes propriétés théoriques pour une capacité à améliorer les performances. Néanmoins les auteurs donnent des conditions assurant que la convergence vers un point stationnaire V^* de l'objectif factorisé coïncide avec un optimum global $B^* = V^*V^{*T}$ du problème initial. Un résultat similaire apparaît dans Journée et al., 2010 et depuis, Boumal, Voroninski et Bandeira, 2018 montre sous des conditions plus faibles de régularité que si on a $p(p+1)/2 > \mathrm{rang}(\mathcal{A})$, l'optimisation de la reformulation non-linéaire (1.3) conduira presque tout le temps à trouver un optimum global du problème initial. De nombreux travaux dans des contextes similaires viennent valider cette approche (voir Ge, Lee et Ma, 2016 ; Ge, Jin et Zheng, 2017). Malheureusement à ce jour ces résultats n'ont pas été adaptés au problème des SDP à contraintes positives (0.1).

Les méthodes de résolution de ces problèmes factorisés ont des origines diverses : Burer et Monteiro, 2003 utilise une méthode BFGS à mémoire limitée pour optimiser le lagrangien augmenté. En rapport avec la méthode de cavité en physique statistique, Lesieur et al., 2016 fait du Bayes-Optimal approximate-message-passing (AMP) pour un mélange gaussien (Matsushita et Tanaka, 2013 a montré comment faire de la reconstruction de faible rang avec l'AMP). Une autre méthode provenant de la physique, méthode de descente par blocs de coordonnées, est employée par Ricci-Tersenghi, Javanmard et Montanari, 2016 pour résoudre la factorisation faible rang issue d'une relaxation SDP. Toutes ces méthodes sont de nature approchée.

Journée et al., 2010 souligne que l'invariance des solutions de (1.3) (pour des programmes dont l'objectif n'est plus simplement linéaire mais généralement convexe) à la multiplication à droite par une matrice orthogonale affaiblit grandement la vitesse de convergence des méthodes du second ordre. En optimisant plutôt sur la variété riemannienne des classes d'équivalence engendrées par cette invariance, les auteurs obtiennent une convergence quadratique grâce à un algorithme des régions de confiance.

Récemment Carson, Mixon et Villar, 2017 ont proposé une adaptation de la théorie de l'optimisation de variété pour le problème K-MEANS SDP (0.1). Ils montrent qu'une méthode de descente de gradient avec la règle d'Armijo sur la variété K-means permet une résolution pratique de l'objectif, et dont la convergence théorique est assurée par les travaux de Boumal, Absil et Cartis, 2018. Mais les seuls résultats présentés pour l'instant exhibent $N = 100$ points.

2 Nouveaux estimateurs computationnels

De nos recherches empiriques menées sur le partitionnement, on a extrait un ensemble d'estimateurs ou méthodes dignes d'intérêt qu'on introduit à présent.

2.1 Estimateurs pour le partitionnement

ADMM pour PECOK

En première partie on a expliqué que la méthode ADMM pouvait être employée pour résoudre le problème (0.1). On donne ici l'algorithme final adapté à la résolution de ce SDP pour le partition-

nement de points ou variables.

On suit fidèlement la construction de Amini et Levina, 2018 pour formuler le programme ADMM. Soit $\mathcal{A} : \mathbb{R}^{N\times N} \to \mathbb{R}^{N+1}$ l'opérateur linéaire capturant les $N+1$ contraintes affines de (0.1), avec $b \in \mathbb{R}^{N+1}, b_i = 1$ pour $i = 1, ..., N$, $b_{N+1} = K$ de sorte que $\mathcal{A}(X) = b$ encode $X1 = 1$ et $\mathrm{tr}(X) = K$. Le programme (0.1) est équivalent à :

$$\underset{X,Y,Z}{\operatorname{argmin}} \langle A, X\rangle + \delta_{\mathcal{A}(X)=b} + \delta_{Y\succcurlyeq 0} + \delta_{Z\geqslant 0} \quad \text{t.q.} \quad X = Y = Z \tag{2.1}$$

En introduisant le lagrangien augmenté $\mathcal{L}_\rho$, la variable duale $U = (U_X, U_Y, U_Z)$ et en notant $\Pi_{\mathcal{A}}$ l'opérateur de projection sur l'ensemble $\{X : \mathcal{A}(X) = b\}$, en notant encore $\bar{X} := (X+Y+Z)/3$ on obtient les étapes alternées ADMM suivantes :

ADMM pour PECOK

1. Calculer la matrice relationnelle corrigée des données $\widetilde{\Lambda} := \widehat{\Lambda} - \widehat{\Gamma}$
2. Réaliser la descente ADMM suivante :

$$X^{k+1} = \Pi_{\mathcal{A}}(\bar{X}^k - U_X^k + (1/\rho)\widetilde{\Lambda}) \tag{2.2}$$
$$Y^{k+1} = \Pi_{\succcurlyeq 0}(\bar{X}^k - U_Y^k) \tag{2.3}$$
$$Z^{k+1} = \Pi_{\geqslant 0}(\bar{X}^k - U_Z^k) \tag{2.4}$$
$$U^{k+1} = U^k + (X, Y, Z)^{k+1} - (\bar{X}, \bar{X}, \bar{X})^{k+1} \tag{2.5}$$

On utilisera l'algorithme ADMM avec des critères d'arrêt multiples inspirés des recommandations de Vandenberghe et Boyd, 1996, un paramètre de pénalisation ρ constant dans toutes les expériences d'un même contexte pour lequel il sera dimensionné (on a en effet $\rho \in \{1, 10\}$). Les performances de cette méthode sont illustrées figures 4.3, 4.2, 4.4, 4.1 et figures 4.7, 4.6, 4.8, 4.5. Pour les détails et calcul de $\Pi_{\mathcal{A}}$ on se réfèrera à la section A en annexe.

Une factorisation faible rang pour PECOK

On l'a dit précédemment, l'étape de projection conique (2.3) ci-dessus est la plus coûteuse, celle qui demande un temps de calcul de l'ordre de $\mathcal{O}(N^3)$ (coût d'une SVD complète). On trouve dans Cai et Li, 2015 remarque 2.2. page 13, le commentaire suivant :

> This step has dominating computational complexity in each iteration of ADMM. In fact, an exact implementation of this subproblem of optimization requires a full SVD of $Z-\Lambda-E/\rho$, whose computational complexity is $O(N^3)$. When N is as large as hundreds of thousands, the full SVD has scalability issue. An open question is how to facilitate the implementation, or whether there exists a surrogate that is computationally inexpensive. A possible remedy is applying the low-rank iterative method, which means in each iteration of ADMM, the full SVD is replaced by a partial SVD where only the leading eigenvalues and eigenvectors are computed.

Plus tôt on a présenté l'idée de Burer et Monteiro, 2003, qui suggérait de factoriser le problème. Comme ADMM est modulaire, on peut utiliser cette factorisation pour remplacer l'étape de

projection semi-définie positive et transformer (2.3) de cette façon :

$$V^{k+1} = \underset{V}{\operatorname{argmin}} \|VV^T - \bar{X}^k + U_Y^k\|^2 \tag{2.6}$$

$$Y^{k+1} = V^{k+1}(V^{k+1})^T \tag{2.7}$$

Aussi l'approche suggérée par Cai et Li, 2015 va nous intéresser et on propose une approximation qui va dans ce sens – on note un mode opératoire très similaire, et dans l'ambition, chez Ma et Ma, 2017 pour la détection de communauté. Remplaçons l'étape factorisée (2.6) de l'optimisation par une optimisation approchée sur l'ensemble des matrices $\mathbb{R}^{N\times K}$ dans la direction de descente.

On propose de transformer l'étape $V^{k+1} = \operatorname{argmin}_V \|VV^T - T^k\|^2 =: f_k(V)$ (où $T^k := \bar{X}^k + U_Y^k$) par quelques pas dans la direction du gradient. Soit $\kappa \in \mathbb{N}$, pour $i \in 1 \dots \kappa$ faire :

$$V^{k+1} \leftarrow V^{k+1} - \eta_i \partial_V f_k(V^{k+1}), \tag{2.8}$$

avec une longueur de pas η_i calculée avec les conditions d'Armijo pour assurer la convergence. On notera que le gradient a une forme explicite :

$$\partial_V f(V) = 4(VV^T - T^k)V \tag{2.9}$$

Pour résumer on propose l'algorithme suivant :

Factorisation ADMM pour PECOK (FACTPECOK)

1. Calculer la matrice relationnelle corrigée des données $\widetilde{\Lambda} := \widehat{\Lambda} - \widehat{\Gamma}$
2. Réaliser la descente ADMM suivante :

$$X^{k+1} = \Pi_{\mathcal{A}}(\bar{X}^k - U_X^k + (1/\rho)\widetilde{\Lambda}) \tag{2.10}$$

$$Y^{k+1} = V^{k+1}(V^{k+1})^T \qquad \text{où } V^{k+1} \text{ donné par (2.8)} \tag{2.11}$$

$$Z^{k+1} = \Pi_{\geqslant 0}(\bar{X}^k - U_Z^k) \tag{2.12}$$

$$U^{k+1} = U^k + (X, Y, Z)^{k+1} - (\bar{X}, \bar{X}, \bar{X})^{k+1} \tag{2.13}$$

Les performances de cette approche sont illustrées figures 4.11 et 4.10.

Approximations des K-moyennes corrigées

L'intuition et l'importance de la correction $\widehat{\Gamma}$ ont été présentées en introduction : il s'agissait de pallier le défaut de l'estimateur des K-moyennes, qui à tendance à scinder les groupes trop larges. On peut penser que si ce défaut a été suffisamment corrigé, on se trouve ramené au cas où ce dernier estimateur est optimal (le cas isovolumique).

Ainsi cette correction peut donner de bonnes raisons de croire qu'on a ainsi déplacé le problème à un cas où l'on sait (à la manière de Kumar et Kannan, 2010 ou Lu et Zhou, 2016) que l'algorithme de Lloyd, 1982 permet de retrouver la partition sous-jacente. On propose l'algorithme suivant :

KMEANZ

1. Calculer la matrice relationnelle corrigée des données $\widetilde{\Lambda} := \widehat{\Lambda} - \widehat{\Gamma}$ et une approximation

semi-définie positive :

$$\widetilde{\Lambda}' := \widetilde{\Lambda} - a'\widehat{\Gamma} \text{ où } a' := \max\{a \in [0,1] : \widehat{\Lambda} - a\widehat{\Gamma} \succcurlyeq 0\} \tag{2.14}$$

2. En extraire une racine, soit $\widetilde{Z} \in \mathbb{R}^{N\times N}$ telle que $\widetilde{\Lambda}' = \widetilde{Z}\widetilde{Z}^T$
3. Appliquer l'algorithme de Lloyd sur les lignes de $\widetilde{Z}$:

$$\widehat{\mathcal{G}}_{\text{Kmeanz}} := \text{Lloyd}(\widetilde{Z}) \tag{2.15}$$

Les performances de cette méthode sont illustrées figures 4.3, 4.2, 4.4, 4.1 et figures 4.11, 4.10.

Si cet estimateur semble naturel, il exige néanmoins de faire une SVD complète dont on pourrait s'affranchir en ré-employant les techniques spectrales. On propose l'algorithme dérivé suivant, qui reprend le même procédé mais procède à une approximation de faible rang de la matrice relationnelle avant d'en extraire une racine :

KMEANZ+

1. Calculer la matrice relationnelle corrigée des données $\widetilde{\Lambda} := \widehat{\Lambda} - \widehat{\Gamma}$ et une approximation semi-définie positive :

$$\widetilde{\Lambda}' := \widetilde{\Lambda} - a'\widehat{\Gamma} \text{ où } a' := \max\{a \in [0,1] : \widehat{\Lambda} - a\widehat{\Gamma} \succcurlyeq 0\} \tag{2.16}$$

2. En déduire une approximation de rang K : K-SVD($\widetilde{\Lambda}'$) et en extraire une racine, soit $\widetilde{Z} \in \mathbb{R}^{N\times K}$ telle que K-SVD($\widetilde{\Lambda}'$) $= \widetilde{Z}\widetilde{Z}^T$
3. Appliquer l'algorithme de Lloyd sur les lignes de $\widetilde{Z}$:

$$\widehat{\mathcal{G}}_{\text{Kmeanz}} := \text{Lloyd}(\widetilde{Z}) \tag{2.17}$$

Non seulement cette méthode n'exige plus qu'une SVD partielle, mais en plus l'algorithme de partitionnement s'appliquant a posteriori ne travaille plus que sur une matrice de plus faible dimension $\mathbb{R}^{N\times K}$. Les performances de cette méthode seront illustrées figures 4.11 et 4.10.

Comme on avait aussi constaté que les algorithmes cherchant à minimiser l'objectif des K-moyennes (2.1) page 5 n'observaient les données qu'à travers leurs produits scalaires, voir (2.5) page 7, on propose aussi d'appliquer l'algorithme de Lloyd à la matrice relationnelle corrigée des données :

KMEANX

1. Calculer la matrice relationnelle corrigée des données $\widetilde{\Lambda} := \widehat{\Lambda} - \widehat{\Gamma}$
2. Appliquer l'algorithme de Lloyd sur les lignes de cette approximation :

$$\widehat{\mathcal{G}}_{\text{Kmeanx}} := \text{Lloyd}(\widetilde{\Lambda}) \tag{2.18}$$

Les performances de cette méthode sont illustrées figures 4.11, 4.10. On note ici que comme précédemment, on pourrait souhaiter intercaler entre ces deux étapes une étape de projection de

faible rang de la matrice relationnelle corrigée. On serait alors dans le cas de l'algorithme spectral de faible rang McSherry, 2001, appliqué à la matrice relationnelle corrigée.

2.2 Des substituts pour estimer le biais Γ

Les chapitres 3 et 2 dérivent des contrôles quasi-optimaux en utilisant un estimateur de Γ en $\mathcal{O}(N^4)$. On présente ici des substituts moins coûteux :

Définition 2.1 (Estimateur cubique)**.** *Soit* $\widehat{\Gamma}_a^{(3)} := \langle X_a - X_{\widehat{b}_3(a)}, X_a \rangle$ *où*

$$\widehat{b}_3(a) := \underset{b:b\neq a}{\operatorname{argmin}} \max_{c:c\neq a,b} |\langle X_a - X_b, \frac{X_c}{|X_c|_2}\rangle| \tag{2.19}$$

On a déjà montré des façons d'obtenir des garanties théoriques pour cet estimateur (confère Lemme 3.1 page 10 en introduction et une première version du papier introduisant PECOK Bunea et al., 2016). C'est un estimateur qui est moins indépendant du signal que l'estimateur servant dans Bunea et al., 2018a ; Royer, 2017, mais il a le mérite d'être moins coûteux en temps de calcul, de l'ordre de $\mathcal{O}(N^3)$.

Enfin on propose aussi deux estimateurs quadratiques qui cherchent à imiter les propriétés des estimateurs précédents pour un coût moindre en $\mathcal{O}(N^2)$:

Définition 2.2 (Estimateur quadratique et quadratique simplifié)**.** *Soient*

$$\widehat{\Gamma}_a^{(1)} := \langle X_a - X_{\widehat{b}_1(a)}, X_a \rangle \tag{2.20}$$

$$\widehat{\Gamma}_a^{(2)} := \langle X_a - X_{\widehat{b}_1(a)}, X_a - X_{\widehat{b}_2(a)} \rangle \tag{2.21}$$

où on a utilisé

$$\widehat{b}_1(a) := \underset{c:c\neq a}{\operatorname{argmax}} \langle X_a, \frac{X_c}{|X_c|_2}\rangle \quad \textit{et} \quad \widehat{b}_2(a) := \underset{c:c\neq a,\widehat{b}_1(a)}{\operatorname{argmax}} \langle X_a, \frac{X_c}{|X_c|_2}\rangle. \tag{2.22}$$

Ainsi pour donner une façon de trouver un "voisin" d'un point a (voir discussion page 10), les estimateurs b_1, b_2 maximisent l'alignement dans la direction du point X_a recherché.

Les performances de ces estimateurs sont illustrées figure 4.9.

2.3 Un changement de perspective

Le point de départ de ces travaux était la minimisation de l'objectif suivant :

$$\underset{\mathcal{G}=\{G_k\}_{1\leq k\leq K}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{a\in G_k} \left\|X_a - \bar{X}_{G_k}\right\|^2 \tag{2.23}$$

où $\bar{X}_{G_k} = \frac{1}{|G_k|}\sum_{b\in G_k} X_b$. Dans sa version plus générale, l'objectif des K-moyennes s'écrit :

$$\underset{\mathcal{G}=\{G_k\}_{1\leq k\leq K}}{\operatorname{argmin}} \underset{\mu_1,\dots,\mu_K}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{a\in G_k} \left\|X_a - \mu_k\right\|^2. \tag{2.24}$$

Ce critère correspond au maximum de vraisemblance dans le cas homoscédastique et donne en espérance :

$$\mathbb{E}\Big[\sum_{k=1}^{K}\sum_{a\in G_k}\big\|X_a-\mu_k\big\|^2\Big]=\sum_{k=1}^{K}\sum_{a\in G_k}\big\|\,\mathbb{E}[X_a]-\mu_k\big\|^2+\operatorname{tr}\operatorname{Cov}(X_a). \tag{2.25}$$

On voit que dans un modèle simplifié comme (1.6), il sera minimal de valeur la somme des traces des covariances des points en la partition recherchée. Or lorsqu'on minimise sur l'espace $\mu_1,\dots,\mu_K$, on obtient que $\forall k\in[K],\mu_k=\bar{X}_{G_k}$, les moyennes empiriques sur la partition. Et l'espérance n'est plus la même, soient $\mu_1,\dots,\mu_K$, on a (voir calculs en introduction page 9) :

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=1}^{K}\sum_{a\in G_k}\big\|X_a-\bar{X}_{G_k}\big\|^2\Big]=\sum_{k=1}^{K}\sum_{a\in G_k}\big\|\,\mathbb{E}[X_a]-\mu_k\big\|^2+\operatorname{tr}\operatorname{Cov}(X_a)\times(1-\frac{1}{|G_k|})\\ -\big\|\,\mathbb{E}[\bar{X}_{G_k}]-\mu_k\big\|^2.\end{aligned} \tag{2.26}$$

Aussi on voit que l'optimisation (2.23) ne porte pas tout à fait sur la bonne quantité, qui fait en moyenne intervenir un compromis entre l'écart au centre et une quantité dépendant des variances et de la taille du groupe considéré. On a bien un biais par rapport à (2.25). Dans (2.23), il apparaît qu'on estime à la fois les moyennes et les traces des covariances des distributions (implicitement). Au lieu de ça on pourrait considérer la variante suivante :

Définition 2.3 (Variante des K-moyennes)**.** *Minimiser sur l'ensemble des partitions à K éléments l'un des critères suivant :*

$$\mathit{crit}(\mathcal{G})=\sum_{k=1}^{K}\frac{|G_k|}{|G_k|-1}\sum_{a\in G_k}\big\|X_a-\bar{X}_{G_k}\big\|^2 \tag{2.27}$$

Le critère proposé est par construction sans biais. Malheureusement, la simplicité de l'algorithme de Lloyd, qui permettait d'optimiser séparément sur l'affectation des points à centres fixés, n'opère plus ici et il faut par exemple recourir à un algorithme de type glouton. Des essais d'implémentation ont produit une amélioration par rapport à Lloyd, mais qui restait marginale cependant.

3 Numerical experiments

We now want to make a general all-round comparison between well-known methods and the ones that we have crafted in this manuscript. At first, we will present an extensive survey of the ADMM PECOK algorithm implementation that we have introduced above, in the different contexts of variable and point clustering. This showcases strength as well as limitations of the compared methods. This prompts us to then look at other efficient clustering algorithms for higher dimensional problems.

The PYTHON3 implementation of the methods used are found in open access here : github.com/martinroyer/pecok Royer, October, 2017

3.1 Comparing with recognized clustering algorithms

To compare performances with, we use the following well-known methods :

1. the hierarchical clustering method of Ward, 1963, applied on the columns of the data matrix for variable clustering, and on the lines of the data matrix for point clustering
2. a Lloyd approximate K-means algorithm (Lloyd, 1982) with several initializations of Arthur et Vassilvitskii, 2007 (with the same targets)
3. a spectral clustering algorithm on either the empirical covariance matrix or the gram matrix

For those algorithms we use the robust implementation from acclaimed software library SCIKIT-LEARN Pedregosa et al., 2011.

We present results in exact recovery, as well as approximate recovery. Throughout this section in order to measure estimator performance in terms of approximate recovery we use the variation of information (VI) metric. It is an entropy-based measure between partitions designed in Meilă, 2007 by analogy with the total variation. Let $\mathcal{H}(\mathcal{G})$ be the entropy of random variable $\mathcal{G}$ and $I(\mathcal{G}, \mathcal{G}')$ be the mutual information between $\mathcal{G}$ and $\mathcal{G}'$, the variation of information is then defined as :

$$\mathrm{VI}(\mathcal{G}, \mathcal{G}') := \mathcal{H}(\mathcal{G}) + \mathcal{H}(\mathcal{G}') - 2I(\mathcal{G}, \mathcal{G}') \tag{3.1}$$

We stress that the more two partition $\mathcal{G}$ and $\mathcal{G}'$ resemble one another, the more $\mathrm{VI}(\mathcal{G}, \mathcal{G}')$ goes towards zero.

Variable clustering

First we look at the context of variable clustering, and showcase our ADMM implementation of the PECOK program as well as an oracle version of the COD algorithm, that is : a hierarchical clustering algorithm based on the SCOD metric (4.2) (voir page 34) using Ward's linkage.

In order to evaluate comparative clustering performances, we look at various scenarii and for each scenario we will be working with the task of clustering in the context of the following default setup : we set to the task of clustering $p = 200$ variables split into $K = 10$ equally-sized groups with standard deviations ranging from 0.1 to 1, and where the covariance matrix C is block diagonal and each block is a 2×2 block matrix B, where $B_{11} = 0.6$, $B_{22} = 2$, $B_{12} = B_{21} = 0.8$. We display respective performances as the number of observations n is grown and all reported values are averaged over 100 identical experiments. Results for the 'default' setup are presented on the left panel, from which the middle and right panels depart in one aspect (respectively : increased SNR in Figure 4.1, difference in heteroscedasticity in Figure 4.2, varying number of clusters in Figure 4.3 and distorted balance in cluster size in Figure 4.4). Hence the default situation is identical throughout all experiments, i.e. for Figures 4.1, 4.2, 4.3 and 4.4, the left panel is always identical.

The experiment shown in Figure 4.1 shows that PECOK and COD are most relevant for tackling hard problems in the presence of heteroscedasticity. When heteroscedasticity is less important as shown in Figure 4.2, all procedures enjoy roughly the same behaviour. In the presence of a larger number of structures, Figure 4.3 shows the two procedures to be dominant again, and the right panel for $K = 2$ is a rare one where another procedure than PECOK (the hierarchical procedure) has the better hand. It seems that the two cluster setup really is a specific clustering problem. Lastly Figure 4.4 shows that when the clusters tend to be imbalanced, the SCOD metric can become most relevant.

Overall for variable clustering the ADMM implementation clearly has the cleanest of records, displaying a robust behaviour in most scenarii. When the amount of available information n is low, almost no method can correctly cluster all points, and on the contrary when n is high, the de-biasing procedure $\widehat{\Gamma}$ seems to give PECOK (ADMM) a strong advantage over its competitors. Naturally as discussed above this comes at a very high computing cost, but encouraging results for strong alternatives are presented further below.

Variable clustering : varying SNR

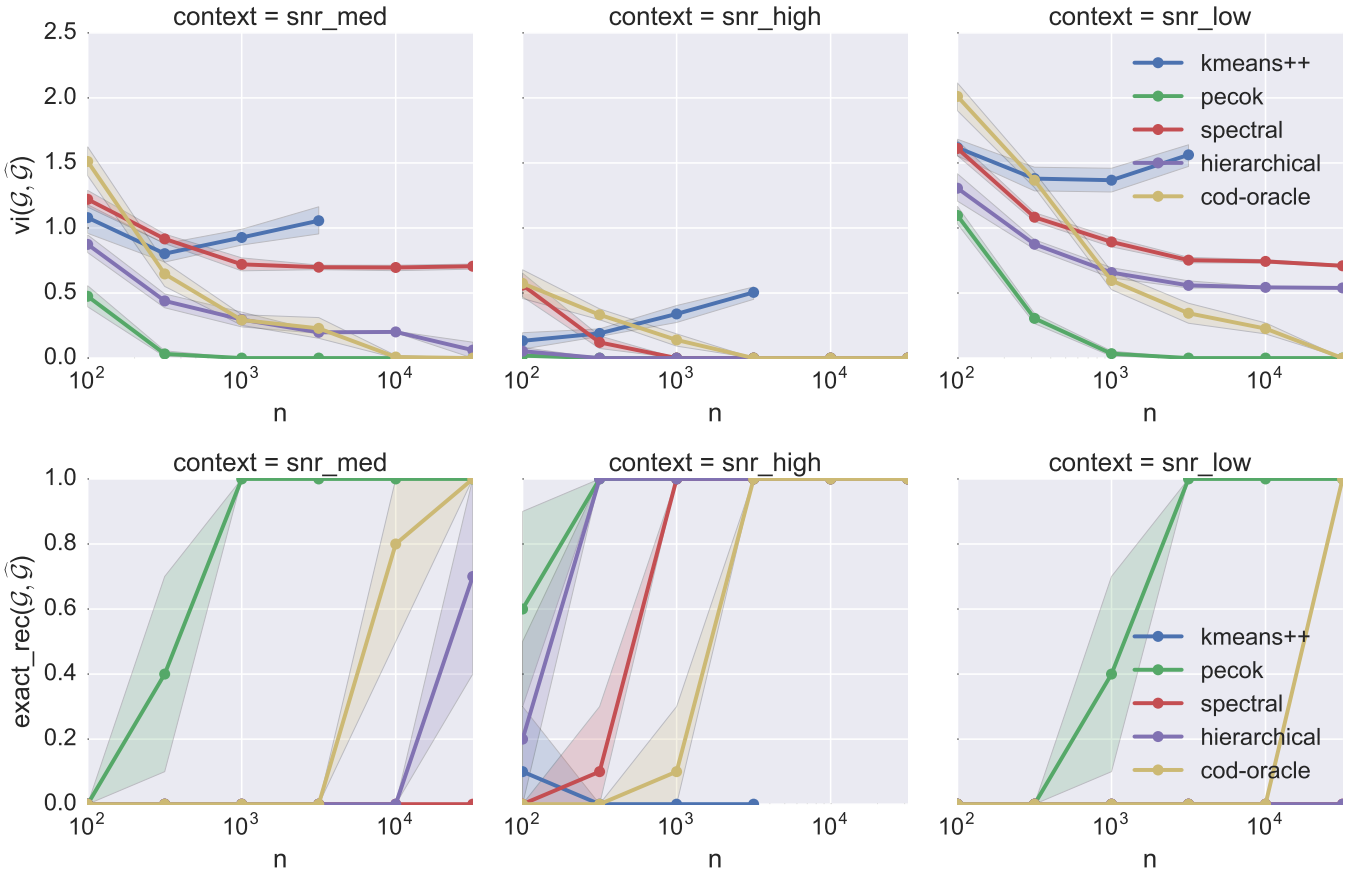


FIGURE 4.1 – Performance comparison with different amount of underlying SNR $(= \Delta(C)/\sqrt{|D|_\infty})$. (left, default) medium SNR = 0.1. (middle) high SNR = 0.3. (right) low SNR = 0.05. Top row : approximate results, bottom row : exact results

Variable clustering : varying heteroscedasticity

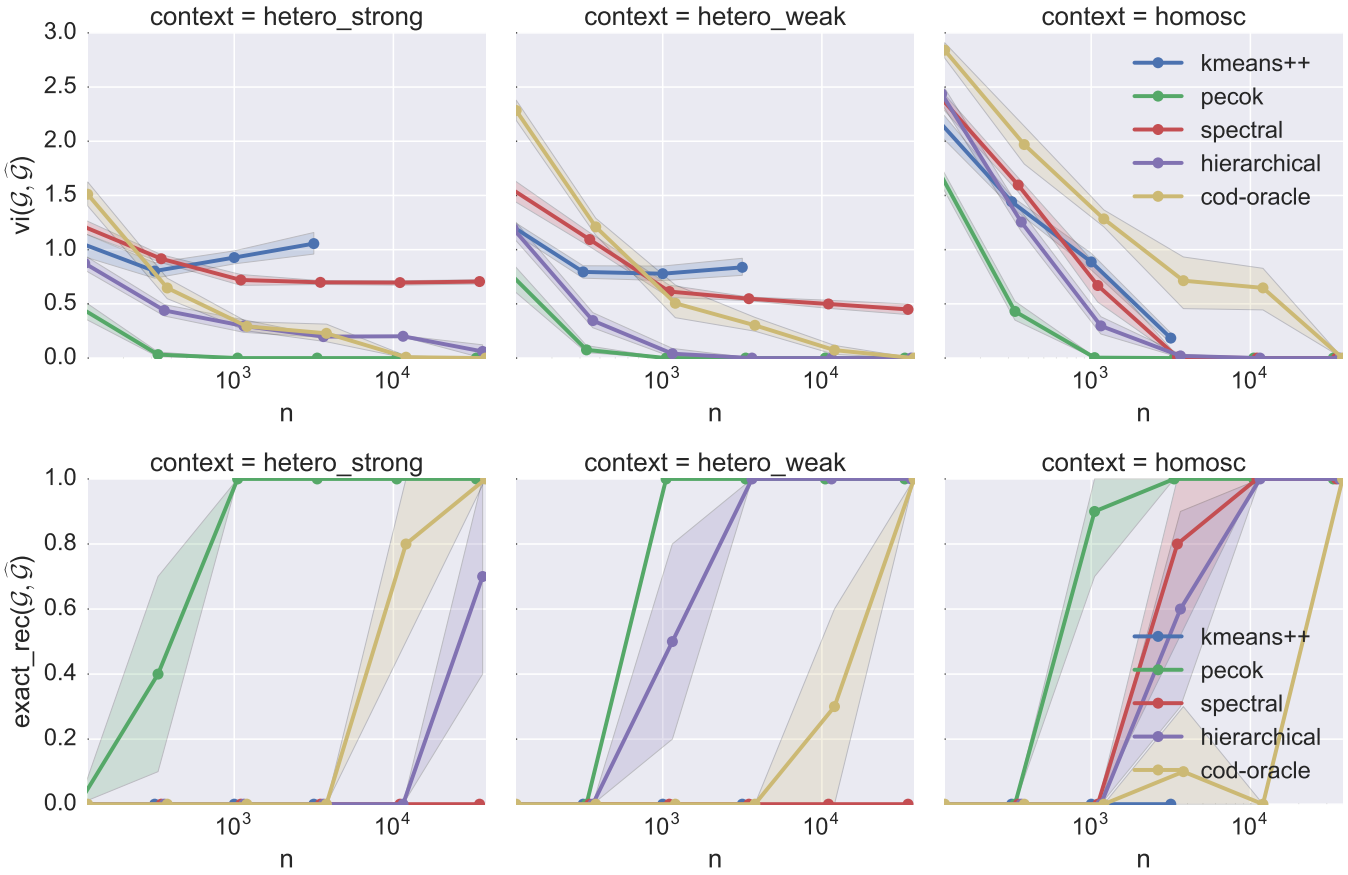


FIGURE 4.2 – Performance comparison with respect to heteroscedasticity. (left, default) standard deviations for groups : [0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.] (so for variable i in group 2, $\sqrt{\mathrm{Var}\, E_i} = 0.2$ and for variable j in group 10, $\sqrt{\mathrm{Var}\, E_j} = 1$). (middle) 10 deviations equally spaced between 0.5 and 1. (right) homoscedascetic variables. Top row : approximate results, bottom row : exact results

Variable clustering : varying number of structures

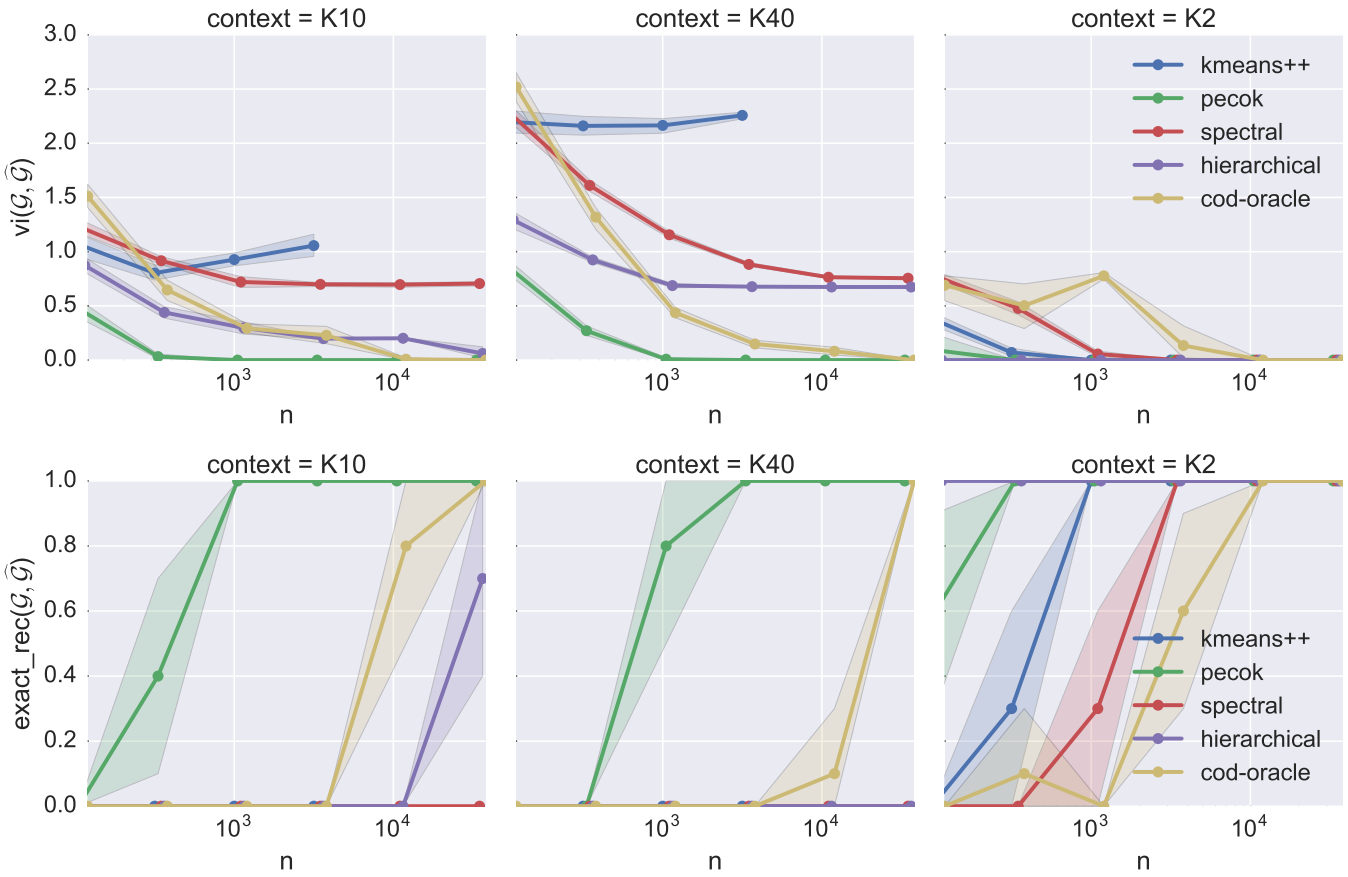


FIGURE 4.3 – Performance comparison for varying number of structures. (left, default) $K = 10$. (middle) K=40. (right) K=2. Top row : approximate results, bottom row : exact results

Variable clustering : varying balance in group size

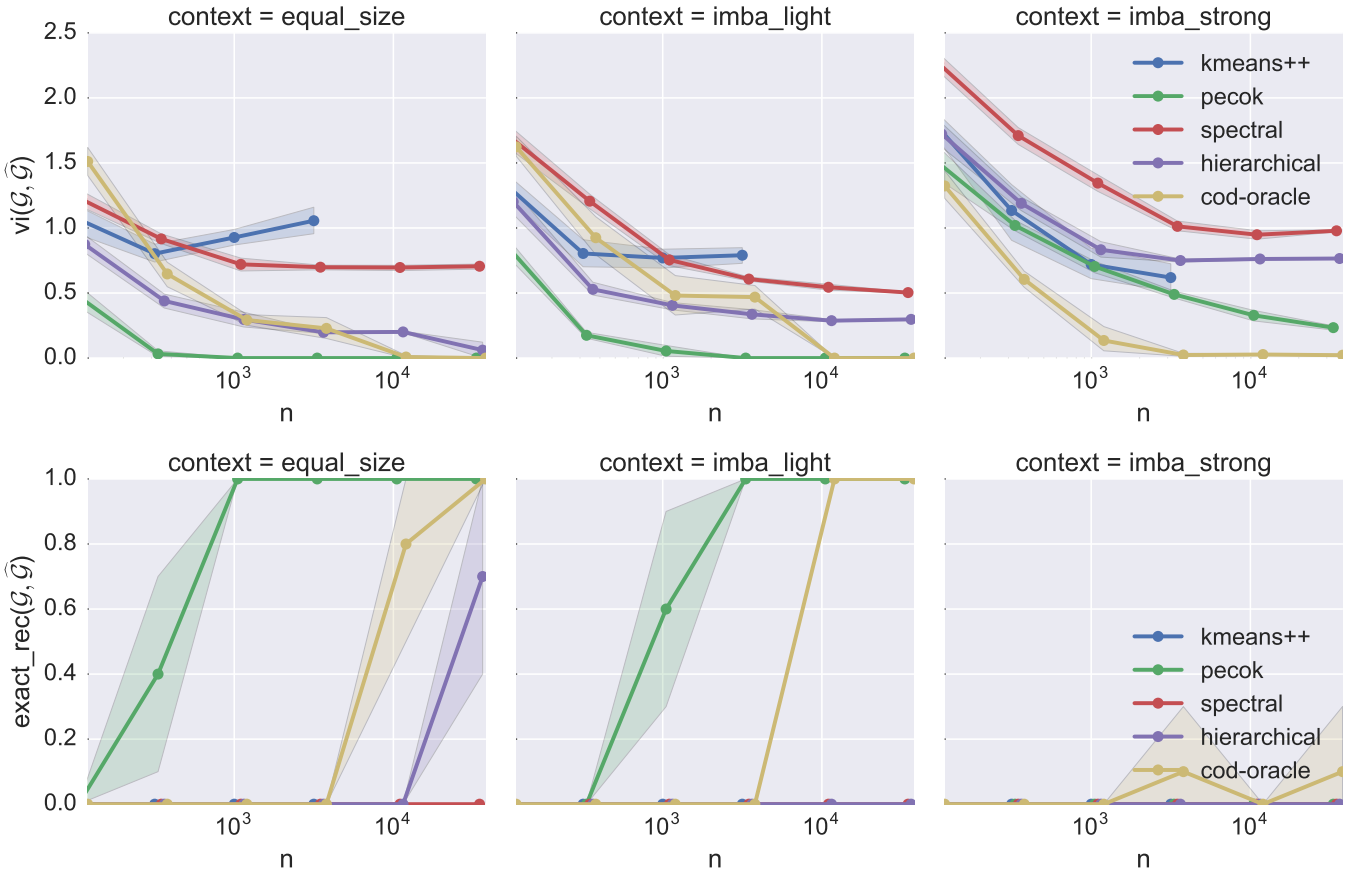


FIGURE 4.4 – Performance comparison with respect to balance in group size. (left, default) equally-sized groups ($K = 10$ groups of $m = 20$ variables each). (middle) group size distribution : [1, 5, 10, 14, 18, 22, 26, 30, 35, 39]. (right) group size distribution : [1, 1, 1, 2, 2, 2, 2, 55, 63, 71]. Top row : approximate results, bottom row : exact results.

Point clustering

In this second part, we look at the context of point clustering : the perspective is slightly shifted, as we "cluster" on the lines of the input data matrix, so there is a transposition of sort. Despite the closeness of the variable and point clustering contexts, this second batch of experiments exemplifies how they really differ : in the point clustering context, the increase of n still corresponds to an increase in the amount of information, perhaps more precisely on the understanding one can gather on estimating the various group distributions. But it also comes with an increase in clustering difficulty, as there are more and more points to split from one another, so simply increasing the number of points no longer mean that the problem to tackle has become easier. We showcase our ADMM implementation of the PECOK program as well what can be seen as a direct way of correcting the K-means estimator introduced in Section 2 page 116 : KMEANZ.

We will be working with the task of clustering in the context of what we will call a 'default' setup : $n = 200$ variables split into $K = 10$ equally-sized groups with standard deviations from 0.1 to 1, and where means are randomly drawn onto a sphere so that the minimum distance between two groups corresponds to a given SNR. All reported values are averaged over 100 identical experiments. We measure performances as the level of SNR is increased. All experiments showcase this setup and others closely linked to it, departing in one aspect (respectively : varying space dimension in Figure 4.5, difference in heteroscedasticity in Figure 4.6, varying number of clusters in Figure 4.7 and distorted balance in cluster size in Figure 4.8). Hence the default situation is identical throughout all experiments, i.e. for Figures 4.7, 4.6, 4.8 and 4.5, the left panel is always identical.

Figure 4.5 shows PECOK and KMEANZ to be dominant in the context of high dimensional recovery, whereas all procedures tend to perform equally well when the space dimension is moderate. Figure 4.6 also shows that the two procedures are best contrasted in the presence of heteroscedasticity. As for the number of structures, Figure 4.7 illustrates that KMEANZ is a really good alternative to PECOK, and as in the variable clustering context the two clusters problem also reveals to be quite particular. Lastly Figure 4.8 shows that the other procedures are sooner affected by imbalance in the group size distribution.

Overall, KMEANZ and PECOK do consistently well and outperform the other methods in almost all studied context, with KMEANZ proving an excellent proxy to the PECOK objective. Next section will further prove that it is a grounded and greatly valuable algorithm to be considered for general purpose point clustering.

Point clustering : varying space dimension

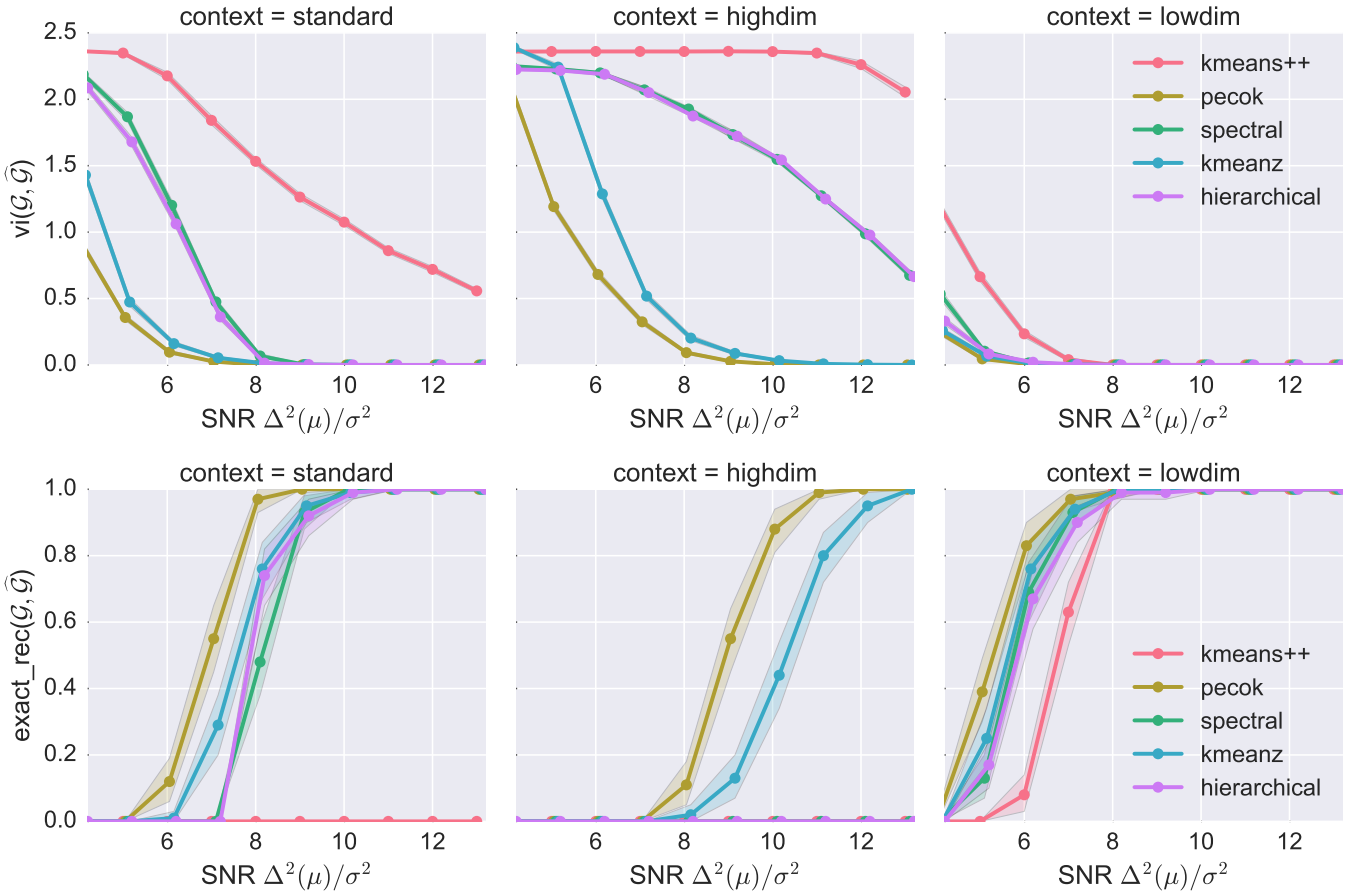


FIGURE 4.5 – Performance comparison with different space dimensions. (left, default) average dimension $p = 500$. (middle) high dimension $p = 2000$. (right) low dimension $p = 100$. Top row : approximate results, bottom row : exact results

Point clustering : varying heteroscedasticity

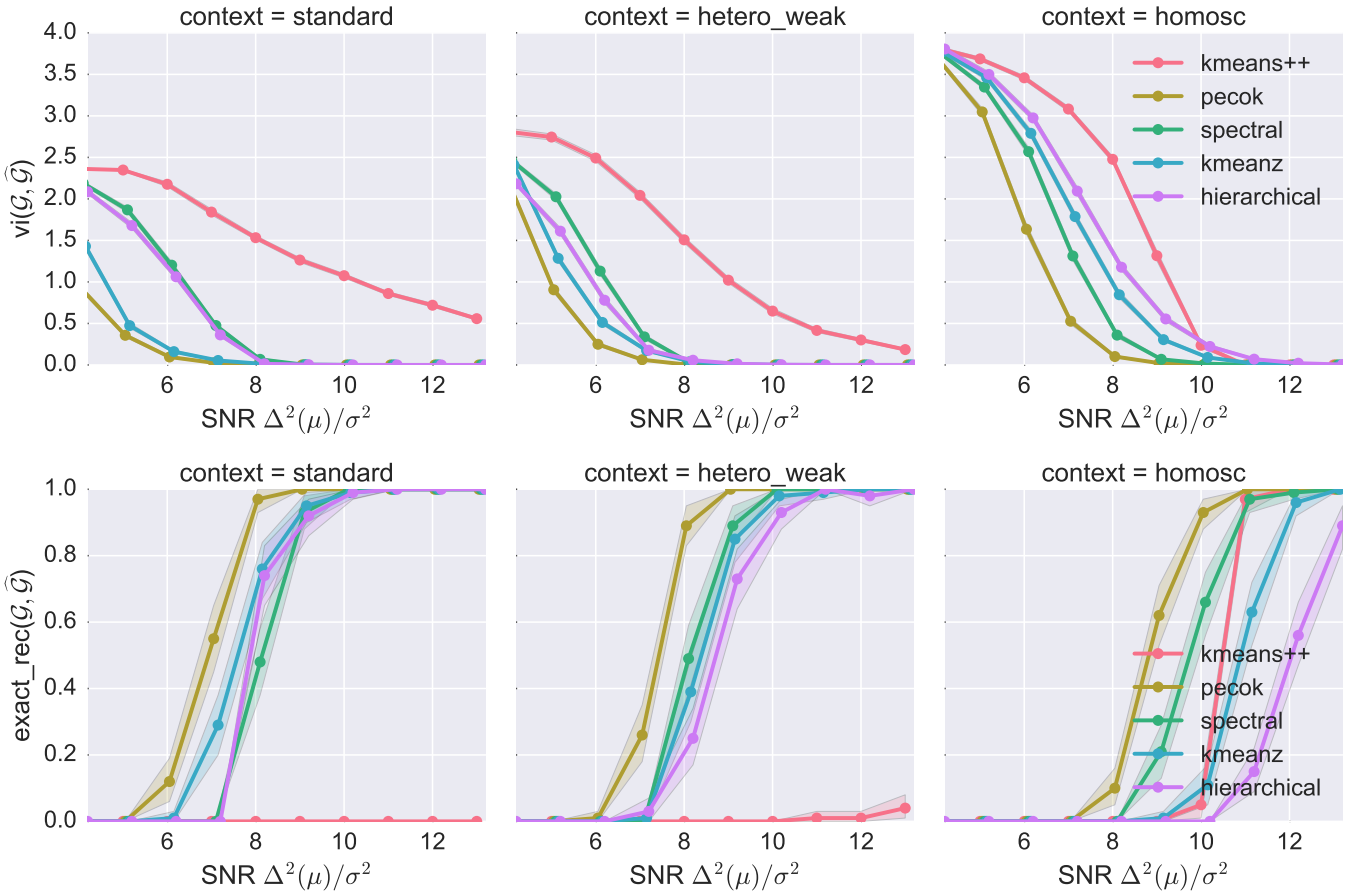


FIGURE 4.6 – Performance comparison with respect to heteroscedasticity. (left, default) standard deviations for groups : [0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.] (so for variable i in group 2, $\sqrt{\mathrm{Var}\, E_i} = 0.2$ and for variable j in group 10, $\sqrt{\mathrm{Var}\, E_j} = 1$). (middle) 10 deviations equally spaced between 0.5 and 1. (right) homoscedascetic variables. Top row : approximate results, bottom row : exact results

Point clustering : varying number of structures

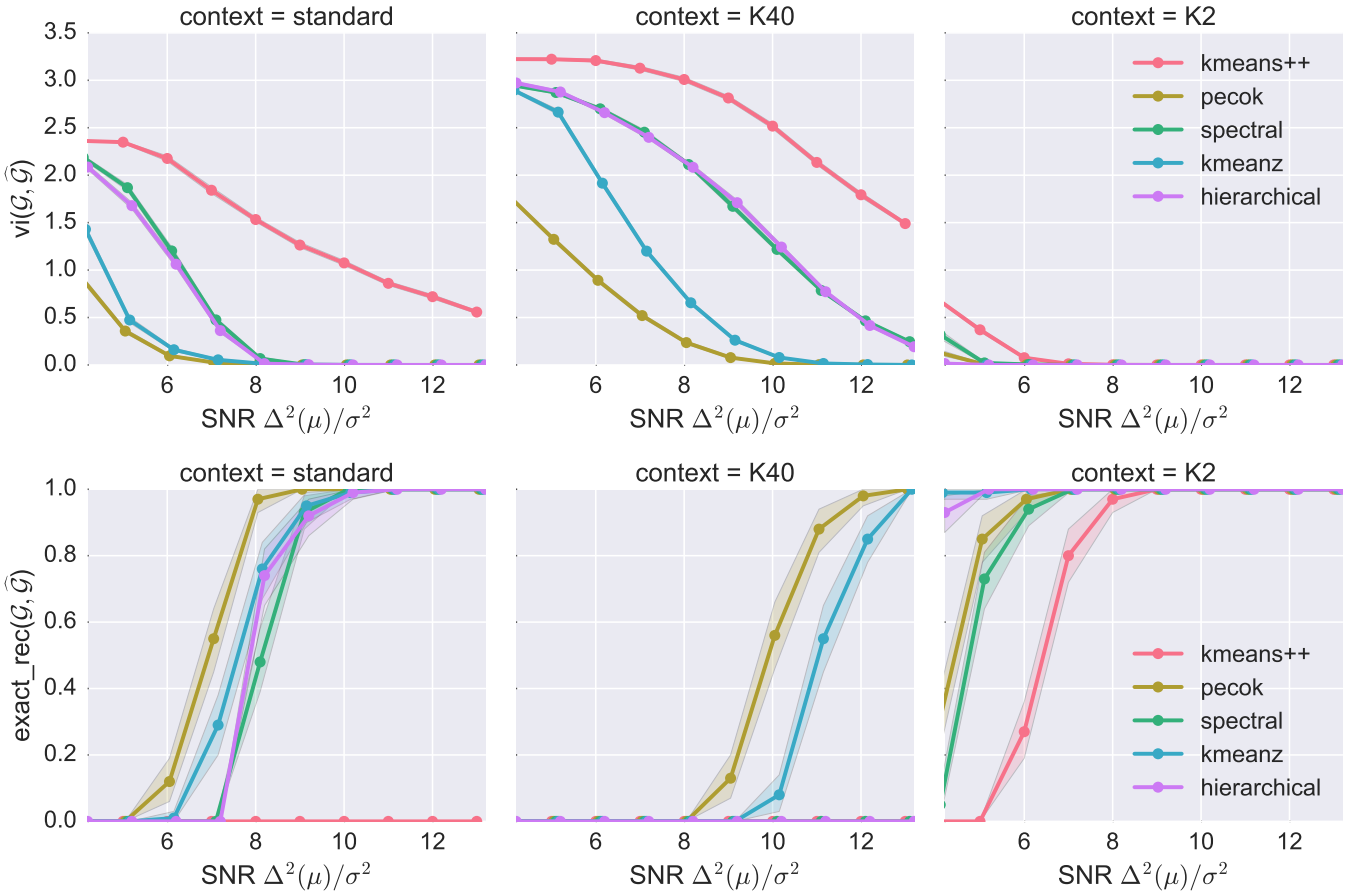


FIGURE 4.7 – Performance comparison for varying number of structures. (left, default) $K = 10$. (middle) K=40. (right) K=2. Top row : approximate results, bottom row : exact results

Point clustering : varying balance in group size

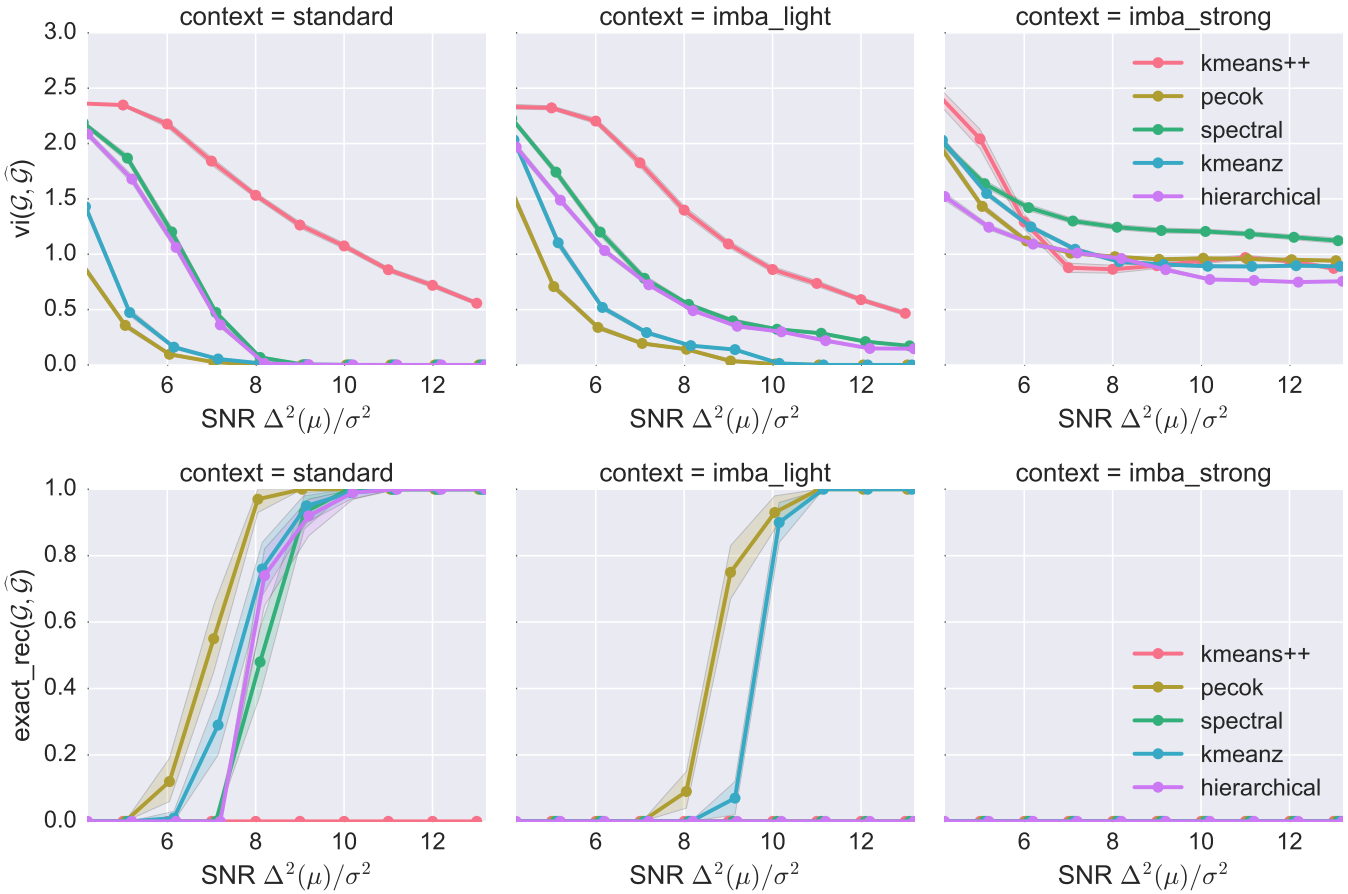


FIGURE 4.8 – Performance comparison with respect to balance in group size. (left, default) equally-sized groups ($K = 10$ groups of $m = 20$ points each). (middle) group size distribution : [1, 5, 10, 14, 18, 22, 26, 30, 35, 39]. (right) group size distribution : [1, 1, 1, 2, 2, 2, 2, 55, 63, 71]. Top row : approximate results, bottom row : exact results.

3.2 K-MEANS SDP surrogates for high-dimension clustering

We have noted excellent performances for the ADMM estimator, but as discussed previously, its main limitation lies in its inability to tackle large scale clustering problems. Here we showcase the performances of estimators designed with that limitation in mind : KMEANX, KMEANZ, KMEANZ+ and FACTPECOK. As FACTPECOK is still a time-consuming procedure, we use an early-stopping procedure after fixed number of a hundred iterations. All these estimators use some form of Γ estimation and we alluded to the fact that $\widehat{\Gamma}^{(4)}$ was much too costly an estimator for Γ, and presented several cheaper options.

$\widehat{\Gamma}$ substitutes experiments

The following experiments simply show (in the same standard setup as for the previous experiments) how well they approximate Γ in the point clustering context. This supports the idea that the surrogates $\widehat{\Gamma}^{(1)}$ and $\widehat{\Gamma}^{(3)}$ can and should be used for the higher scale problems in stead of $\widehat{\Gamma}^{(4)}$. The left display shows that even as the problem grows easier, one can misrepresent Γ for lack of a right estimating procedure : hence simply taking the diagonal of the relational matrix will not necessarily suffice in generally correcting for Γ. The right display shows that if the space dimension is very low, not or ill-estimating Γ could also be very costly for clustering performances.

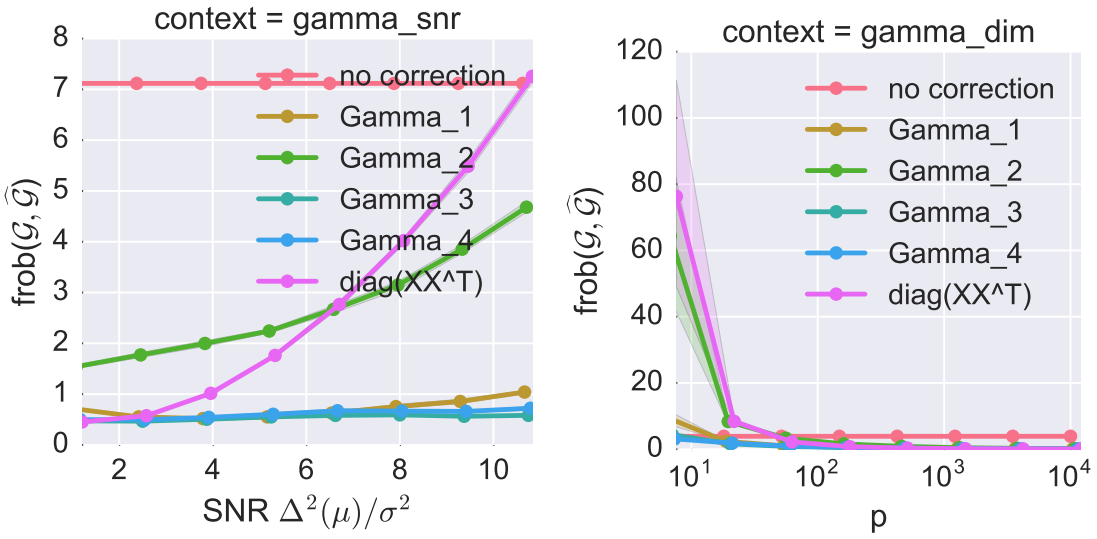


FIGURE 4.9 – Performance comparison for Γ surrogates measured in norm $|\widehat{\Gamma} - \Gamma|_2$. (left) standard setup (with $N = 200$) as the SNR is increased. (right) standard setup as the space dimension is increased.

Hence in what follows we chose to use the $\widehat{\Gamma}^{(1)}$ correction. We also make separate cases for the contexts of point and variable clustering, as it turns out that different procedures work best in each context.

	$\widehat{\Gamma}^{(1)}$	$\widehat{\Gamma}^{(2)}$	$\widehat{\Gamma}^{(3)}$	$\widehat{\Gamma}^{(4)}$
Time (s)	$4.0 * 10^{-3}$	$3.5 * 10^{-3}$	1.29 (0.03)	99.1 (0.5)

TABLE 4.1 – Average computing times (and standard deviations) for the SNR experiment Figure 4.9

Variable clustering

For variable clustering in Figure 4.10 we go on to cluster $p = 500$ variables from 20 different groups. The experiment showcases the strength of the low-rank factorisation FACTPECOK as well as that of approximation procedures KMEANX and corrected low-rank clustering mentioned above. Procedures KMEANZ and KMEANZ+ also do really well and at a very low computing cost (Table 4.2). FACTPECOK is much slower but we stress that the nature of this work is exploratory, hence the FACTPECOK algorithm could possibly benefit from some astute optimized implementation that we did not have in mind.

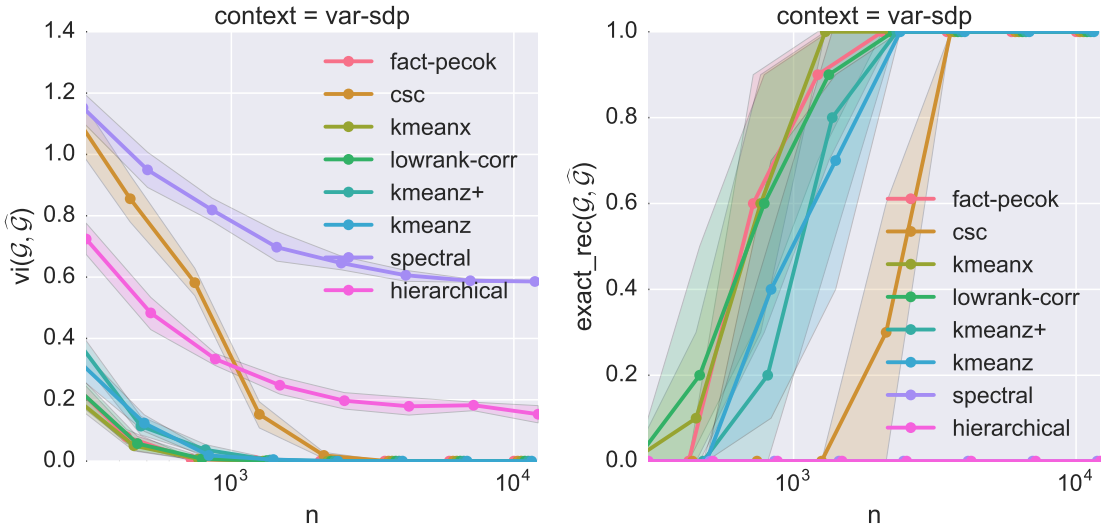


FIGURE 4.10 – Performance comparison for SDP surrogates, clustering $p = 500$ variables from $K = 20$ groups. (left) approximate results, (right) exact results

fact-pecok	CSC	lowrank-c	KMEANX	KMEANZ	KMEANZ+
4.99 (0.07)	0.08 (0.03)	0.16 (0.03)	0.22 (0.03)	0.22 (0.03)	0.07 (0.03)

TABLE 4.2 – Average computing times (and standard deviations) in seconds for the SNR experiment Figure 4.10

Point clustering

As far as point clustering goes in Figure 4.11, the FACTPECOK procedure does not work at all. On the other hand the KMEANZ and KMEANZ+ procedures of earlier have the best performances in clustering a thousand points from $K = 50$ groups in $\mathbb{R}^{2000}$, with CSC also equally performant. This time, KMEANX also fails at providing exact recovery results in time. Surprisingly the difference in behaviour is somewhat reverse from the variable clustering experiment. Table 4.3 shows how competitive the leading algorithms are time-wise.

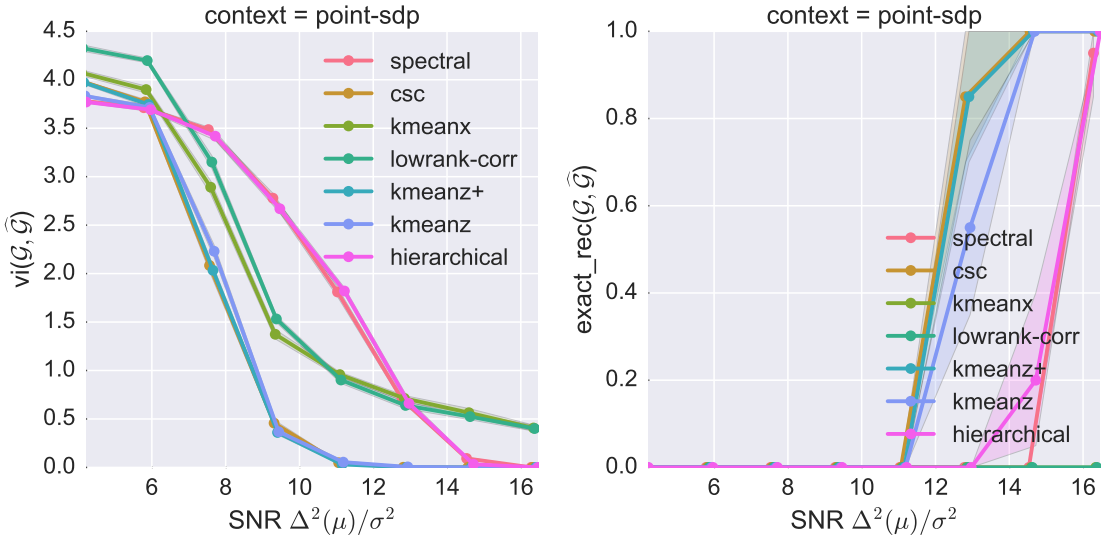


FIGURE 4.11 – Performance comparison for SDP surrogates, clustering $n = 1000$ points from $K = 50$ groups in $\mathbb{R}^{2000}$. (left) approximate results, (right) exact results

CSC	lowrank-c	KMEANX	KMEANZ	KMEANZ+	hierarchical
0.41 (0.05)	1.02 (0.05)	1.52 (0.04)	1.51 (0.04)	0.41 (0.04)	1.34 (0.04)

TABLE 4.3 – Average computing times (and standard deviations) in seconds for the SNR experiment Figure 4.10

Afterword

These experiments have shown that the good performances of ADMM for standard dimensional setups can be replicated in clustering problems of higher difficulty. For variable clustering good candidates would be methods such as KMEANZ, KMEANZ+, KMEANX, a factorized version of PECOK or the corrected low-rank algorithm. For point clustering this demonstrates the strength of KMEANZ, KMEANZ+ and the corrected spectral clustering CSC introduced in Bunea et al., 2018a.

We have implemented a low-rank factorization of the K-MEANS SDP problem (0.1), and our results show that in the context of variable clustering it has no trouble finding the global optimum, an empirical confirmation that such factorized problem still enjoy good theoretical properties as discussed above and intuited by the work of Burer et Monteiro, 2003.

All these methods are closely related to near-minimax estimator PECOK, and perhaps the closest of them is KMEANZ. KMEANZ is designed to emulate the optimal performance of the K-means estimator for when there is no bias in volume, see for instance section 3.1 page 9 in chapter 1. This *no-bias* tranformation of data matrix X into another data matrix $\widetilde{Z}$ is made possible by the $\widehat{\Gamma}$ correction that we developped in earlier chapters and enhanced through this one. Its low computing time and possible direct relaxation KMEANZ+ should make it a very attractive candidate for all sorts of clustering problems.

Annexe

A Itérations ADMM pour PECOK

On a séparé la variable B en trois parties X, Y, Z, qu'on contraindra à être égales grâce à $\chi \in \mathcal{D} = \{\chi = (x, y, z) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} | x = y = z\}$. Alors le problème (2.1) page 115 est équivalent à :

$$\begin{aligned} &\underset{X,Y,Z}{\operatorname{argmin}} f(X,Y,Z) + \delta_{\mathcal{D}}(\chi) \\ &\text{t.q. } (X,Y,Z) - \chi = 0 \end{aligned} \tag{A.1}$$

où $f(X,Y,Z) := -\langle A, X\rangle + \delta_{\mathcal{A}(X)=b} + \delta_{Y\succcurlyeq 0} + \delta_{Z\geqslant 0}$. On introduit la variable duale renormalisée $U = (U_X, U_Y, U_Z) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$ et le lagrangien augmenté : $L_\rho((X,Y,Z),\chi,U) = f(X,Y,Z) + \delta_{\mathcal{D}}(\chi) + (\rho/2)\|(X,Y,Z) - \chi + U\|_2^2$. ADMM est le processus de minimisation alternée du lagrangien augmenté :

$$(X,Y,Z)^{k+1} = \underset{(X,Y,Z)}{\operatorname{argmin}} f(X,Y,Z) + (\rho/2)\|(X,Y,Z) - \chi^k + U^k\|_2^2 \tag{A.2}$$

$$\chi^{k+1} = \underset{\chi}{\operatorname{argmin}}\, \delta_{\mathcal{D}}(\chi) + (\rho/2)\|(X,Y,Z)^{k+1} - \chi + U^k\|_2^2 \tag{A.3}$$

$$U^{k+1} = U^k + (X,Y,Z)^{k+1} - \chi^{k+1} \tag{A.4}$$

Après simplifications standards (voir Vandenberghe et Boyd, 1996) on peut éliminer χ et obtenir :

$$X^{k+1} = \Pi_{\mathcal{A}}(\chi_0^k - U_X^k + (1/\rho)A) \tag{A.5}$$

$$Y^{k+1} = \Pi_{\succcurlyeq 0}(\chi_0^k - U_Y^k) \tag{A.6}$$

$$Z^{k+1} = \Pi_{\geqslant 0}(\chi_0^k - U_Z^k) \tag{A.7}$$

$$U^{k+1} = U^k + (X,Y,Z)^{k+1} - (\bar{X},\bar{X},\bar{X})^{k+1} \tag{A.8}$$

Il reste à dériver l'opérateur de projection $\Pi_{\mathcal{A}}$:

$$\Pi_{\mathcal{A}}(Y) = Y - \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}[\mathcal{A}(Y) - b] \tag{A.9}$$

où $\mathcal{A}^*$ est l'adjoint de $\mathcal{A}$. On peut définir $[\mathcal{A}(X)]_i = \langle X, \widetilde{H}_i\rangle$ et $[\mathcal{A}(X)]_{N+1} = \langle X, I_p\rangle$, où $\widetilde{H}_i = (1_{k<i} + 1_{l\geqslant i})_{kl} \in \mathbb{R}^{N\times N}$. On a alors :

$$\mathcal{A}^*(b_1, ..., b_N, b_{N+1}) = \sum_{i=1}^{N} b_i \widetilde{H}_i + b_{N+1} I_N \tag{A.10}$$

et $(\mathcal{A}\mathcal{A}^*) = (N-1)I_{N+1} + 1_{N+1}$, with $1_{N+1} = (1)_{kl} \in \mathbb{R}^{N\times N}$. En constatant que $((\mathcal{A}\mathcal{A}^*) - (N-1)I_{N+1})^2 = (N+1)((\mathcal{A}\mathcal{A}^*) - (N-1)I_{N+1})$, on peut inverser $(\mathcal{A}\mathcal{A}^*)$:

$$(\mathcal{A}\mathcal{A}^*)^{-1} = \frac{1}{N-1} I_{N+1} - \frac{1}{2N(N-1)} 1_{N+1}. \tag{A.11}$$

Chapter 5

Latent model-based clustering for biological discovery

Contents

This chapter presents results obtained in collaborative work with Jishnu Das and Mike Bing. An article has been submitted to *iScience*.

Abstract

Numerous non-overlapping and overlapping clustering methods have been developed for a wide variety of specific biological applications but overall the field is lacking for more generic estimators with advantageous theoretical properties as well as the ability to adapt to different contexts. We put the LOVE method of Bing et al., 2017 to that challenge – a robust, highly scalable latent model-based clustering method for biological discovery. LOVE can be used across a range of datasets to generate both overlapping and non-overlapping clusters. We apply the algorithm to a gene-expression dataset and demonstrate that it detects biologically meaningful clusters. LOVE outperforms existing methods both in terms of the significance of the clusters, as well as correctly identifying overlaps corresponding to pleiotropic gene function. We also apply LOVE to a cohort of HIV controllers and chronic progressors and show that it is able to accurately cluster these two distinct clinical phenotypes in a non-overlapping fashion. Our results demonstrate that LOVE can be reliably used across a wide range of datasets for novel biological discovery.

1 Introduction

One of the most critical aspects of handling large biological datasets is identifying and accurately quantifying similarities and differences in the data. Clustering is one of the most popular ways to do this, and many clustering algorithms with specific biological applications have been developed over the last 2 decades. However, despite the availability of numerous clustering algorithms, 3 key issues still remain unaddressed. Most clustering methods use heuristics to assign clusters and do not come with rigorous statistical performance guarantees. Second, existing clustering methods work well only for specific datasets. A comprehensive benchmarking of 13 well-known methods across 24 datasets revealed that there was no universal best performer; rather, methods typically worked best for the datasets they were specifically designed for (Wiwie, Baumbach, and Röttger, 2015). Further, clustering methods typically work either for generating overlapping or non-overlapping clusters but both.

Hence biologists are in need for a clustering approach that comes with rigorous statistical guarantees regarding both cluster identification and assignment of variables to clusters. The method should be generically applicable across a wide range of datasets, as there would be no assumptions regarding data structure. Further, its design would enable the same method to be used to generate both overlapping and non-overlapping clusters. Here, we report on the performances of LOVE from Bing et al., 2017, a robust and highly scalable latent factor model-based clustering method that enjoys all the above properties. We apply LOVE to two datasets with very different structures and show that it generates stable, biologically meaningful and accurate clusters in both cases, outperforming current state-of-the-art methods.

LOVE considers the following latent factor model:

$$X = AZ + E, \tag{1.1}$$

where $X \in \mathbb{R}^p$ represents p genes, $Z \in \mathbb{R}^K$ denotes K latent factors corresponding to the K clusters based on K hidden latent factors. $A \in \mathbb{R}^{p\times K}$ represents the membership matrix assigning p variables to K clusters and E denotes an error term corresponding to random noise. To explicit the connexion with previous models from this manuscript, if we had $A \in \{0,1\}^{p\times K}$ with $A1_K = 1_p$ this would then be an instance of the more general G-block model (see (1.1) page 25 and (1.2) page 26 in chapter 2). A more rigorous mathematical presentation of the modelling is given in appendix Section A.

LOVE operates in three main steps: at first it identifies a specific subset of the variables that are called non-mixed, that have the specific property of belonging to a single cluster. This allows for an estimation of the precision matrix of the latent variables, that encodes information as to how the groups interact between each other. Then one determines an estimation of cluster association strength for the remaining mixed variables, and computes clusters based on that information. See appendix section A for a detailed algorithm, Figure 5.3 for visual intuition and we refer to Bing et al., 2017 for further details.

2 Overlapping clustering using LOVE

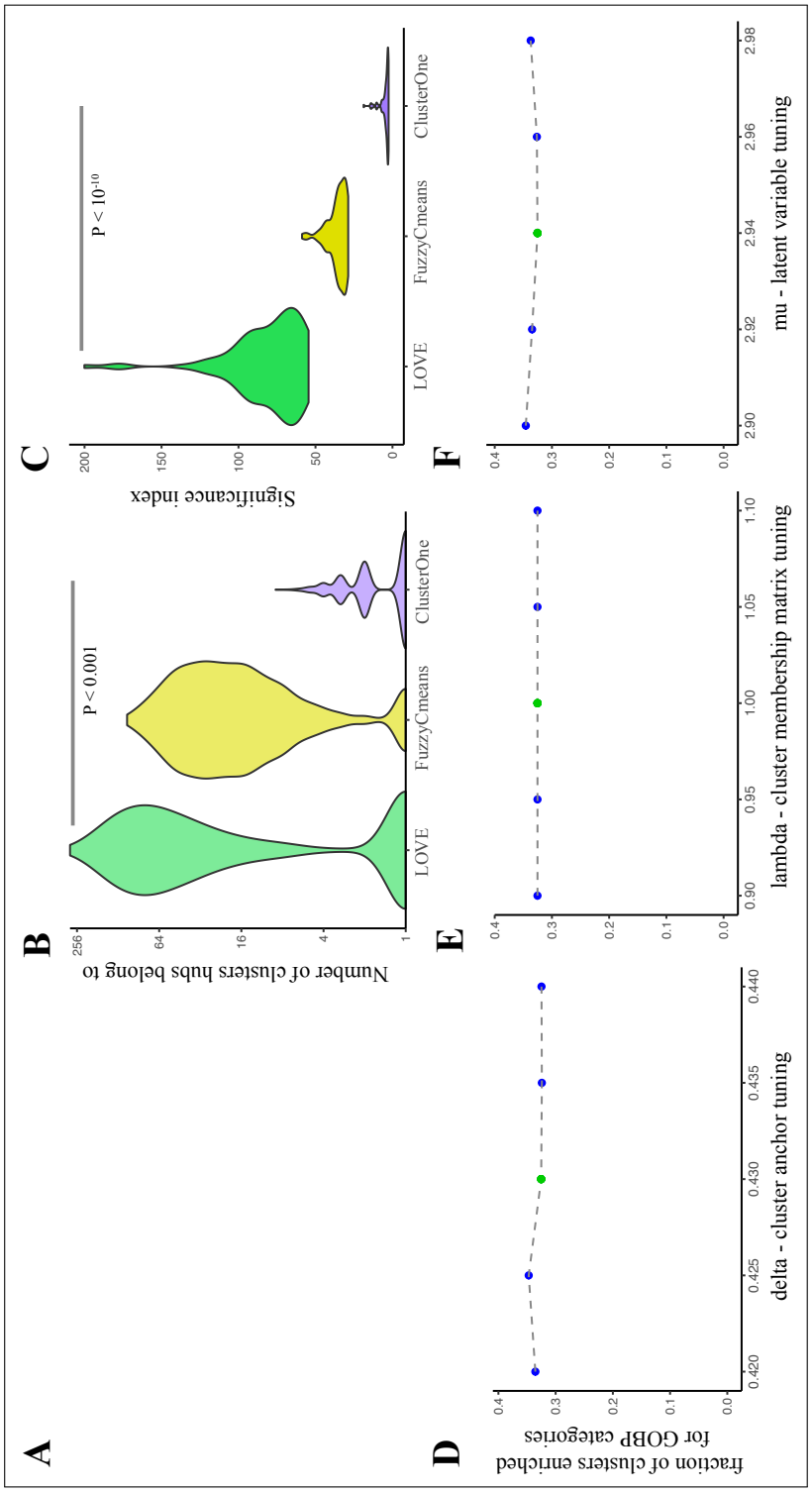
To test overlapping clustering using the LOVE algorithm, we used a previously described compendium of human gene expression data (Das, Jaaved, and Haiyuan, 2012). The dataset corresponds to expression measurements for 16,134 genes across 114 different points in the cell cycle.

Using LOVE, we obtained >1,000 overlapping clusters of varying sizes (Figure 5.1). In general, clusters corresponded to specific functions, illustrating that the latent factors in our modeling formulation are not merely mathematical entities, but also have underlying biological relevance. We systematically examine the nature of relationships between clusters and known biological functions later in this section.

First, we sought to evaluate whether the overlaps discovered by LOVE are biologically meaningful. If clusters indeed correspond to biological functions, one expects pleiotropic genes to be overlapping across clusters as these carry out several functions. We defined pleiotropic genes based on the network degree of the proteins encoded by these genes i.e., protein-protein interaction network hubs were defined as pleiotropic. This is a standard way to characterize multiplicity of function as proteins perform their functions by interacting with other proteins (Rolland et al., 2014; Vo et al., 2016; Yu et al., 2008) and previous studies have shown that network hubs are the most functionally important genes (Albert, Jeong, and Barabasi, 2000; Jeong et al., 2000; Yu et al., 2008). We used a consensus high-quality protein interaction network to define hubs (Das and Yu, 2012) and found that hubs belonged to significantly more clusters than non-hubs (Figure 5.1-B, $P < 0.01$). Thus, the assignment of overlapping clusters was consistent with biological expectation – pleotropic genes were more likely to be assigned to multiple clusters than non-pleotropic genes.

We then compared our results to two existing clustering methods – fuzzy k-means clustering (Bezdek, Ehrlich, and Full, 1984; Ashburner et al., 2000) and ClusterOne (Nepusz, Yu, and Paccanaro, 2012). Fuzzy k-means clustering is a well-established and widely used distance-metric based algorithm. ClusterOne is graph-based and has recently been demonstrated to be superior to several similar approaches (Nepusz, Yu, and Paccanaro, 2012). Thus fuzzy k-means clustering and ClusterOne are two state-of-the-art methods, use orthogonal concepts, and serve as excellent benchmarks to compare against. We found that hubs were assigned to significantly more clusters by LOVE than they were by k-means clustering or ClusterOne (Figure 5.1-C, $P < 0.001$ using a Mann-Whitney U test), suggesting that the overlaps detected by LOVE are more consistent with prior biological expectation than the overlaps detected by other methods.

Figure 5.1 – Cell-cycle experiment



While the above analysis shows that genes with multiple functions are correctly assigned by LOVE to multiple clusters, a good clustering method should also not assign genes with similar expression levels to multiple clusters. To test this, we looked at how housekeeping genes (Eisenberg and Levanon, 2013) were distributed across the clusters generated by LOVE. Since housekeeping genes are basally expressed i.e., have low variability in their expression levels, ideally these should only be assigned to one or a few clusters and not the other clusters. To systematically test this, we calculated the under-representation of housekeeping genes in the clusters generated by LOVE and quantified this using an under-representation index (see Supplementary Methods). We found that housekeeping genes were under-represented across most LOVE clusters (Figure 5.1-C) and the corresponding index was significantly higher ($P < 10\text{-}10$ using a Mann-Whitney U test) for LOVE compared to fuzzy k-means clustering and ClusterOne (Figure 5.1-C). These results illustrate that LOVE not only accurately identifies overlaps, it is also effective at discriminating between basally expressed genes and genes with specific expression profiles.

We then explored how biologically relevant the clusters discovered by LOVE are. To define biological relevance, we examined whether a cluster was over-represented for at least one known Gene Ontology biological process (GO BP) category (Ashburner et al., 2000). Significant over-representation was defined using an FDR cutoff of 0.05 (P value calculated from a hypergeometric test followed by Benjamini-Hochberg multiple-testing correction) and computed using WebGestalt (Wang et al., 2013). The number of biologically relevant clusters identified using this approach represents a lower bound on the actual number of biologically relevant clusters as current GO annotations are not complete. Thus, any cluster enriched for at least one GO BP category is definitely biologically relevant, while clusters not enriched for a GO BP category may still be meaningful.

Finally, we tested how the biological relevance of the clusters discovered by LOVE changed when key tuning parameters of the method are varied. We first performed a grid search around the optimal delta, the parameter that determines cluster anchors i.e., which non-mixed variables will serve to define clusters. We found that across a range of parameter values around the optimal delta, the fraction of clusters known to be biologically relevant remained stable (Figure 5.1-D). Next, we performed a similar analysis with lambda – the parameter used to tune the membership matrix i.e., how each variable is assigned to a cluster. Again, the fraction of clusters known to be biologically relevant remained stable around the optimal lambda (Figure 5.1-E). We also observed similar results with a grid search around the optimal mu – the parameter that determines tuning of the latent variables (Figure 5.1-F). Thus, LOVE discovers biologically meaningful clusters across a range of parameter choices.

3 Non-overlapping clustering using LOVE

Most methods are good at either generating overlapping or non-overlapping clusters (Wiwie, Baumbach, and Röttger, 2015). However, due to the inherent formulation of LOVE, it can be used for either purpose. To test the effectiveness of LOVE in generating non-overlapping clusters, we chose a recently published dataset of humoral immune measurements from human subjects from two distinct clinical phenotypes – long-term HIV controllers and chronic progressors (Sadanand et al., 2018). This dataset is different with regard to several key aspects from the earlier gene-expression dataset. First, the desired clusters here are non-overlapping as HIV controllers and chronic progressors are clinically distinct groups and are known to be very different in terms of

their humoral immune responses (Alter et al., 2018; Sadanand et al., 2018). Second, the sources of biological and technical variance in the two datasets are different. In terms of biological variability, the modulation of transcript expression levels across time-points in the cell cycle is structurally very different from variation across human subjects with different clinical phenotypes. The extent of technical noise is also different as microarray measurements are relatively noisy, while this dataset comprises humoral immune measurements collected using modern methods. Finally, the number of entities being clustered (number of input variables for LOVE) is also very different. The gene expression dataset had >16,000 genes profiled over 114 different points in the cell cycle. The dataset of controllers and progressors have 19 human subjects. For each subject, 18 different measurements of antibody-effector functions and titers were available at 4 different time-points, corresponding to a total of 72 measurements (Sadanand et al., 2018). The differences across these 2 datasets reveals the inherent variation across different biological datasets. Testing LOVE on two extremes provides an opportunity to benchmark how the clustering method performs at different ends of the spectrum.

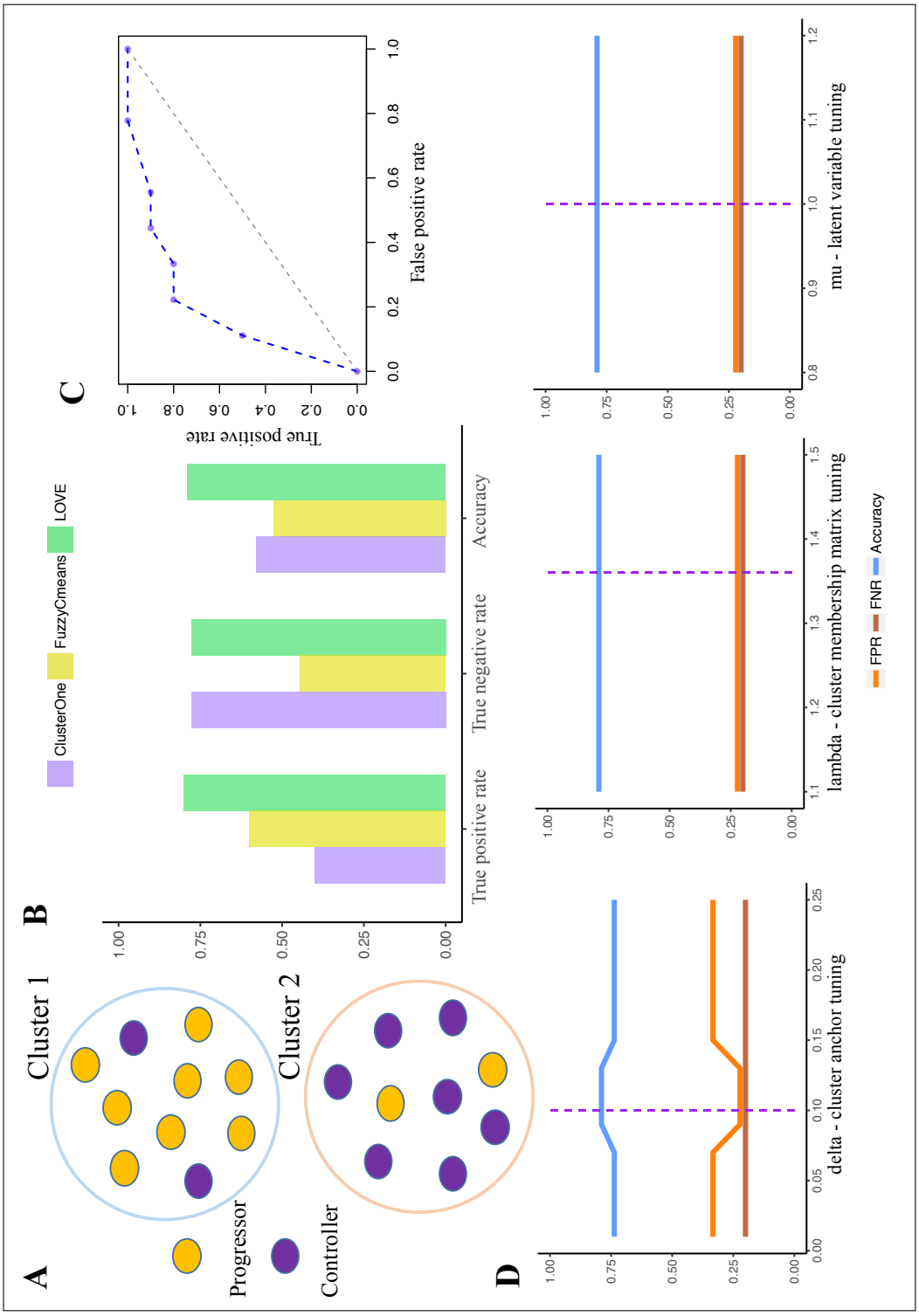
The 19 human subjects were separated into 2 clusters – one of which comprised 8 progressors and 2 controllers, while the other comprised 7 controllers and 2 progressors (Figure 5.2-A). These correspond to an accuracy, a true positive rate and a true negative rate each of 80% for LOVE (Figure 5.2-B). We also found that LOVE outperformed both k-means clustering and ClusterOne in terms of all 3 metrics – accuracy, true positive and true negative rate (Figure 5.2-B).

Next, we sought to explore how LOVE performs across a range of assignments. While the most optimal assignments correspond to all subjects being assigned to a cluster with high accuracy (as shown in Figure 5.2-B), we wanted to check how LOVE does across a range of assignment thresholds (including those where not all subjects are assigned to clusters). A ROC curve drawn across assignment thresholds revealed robust performance across thresholds (Figure 5.2-C, AUC = 0.85).

Finally, we wanted to evaluate how stable LOVE is across the 3 key parameters – delta (cluster anchor tuning), lambda (membership matrix tuning) and mu (latent variable tuning). We performed a grid search around the optimal parameters and found that the all 3 indicators of performance − accuracy, false positive rate ($1 -$ true positive rate) and false negative rate ($1 -$ true negative rate) are stable across a range of tuning parameters (Figures 5.2-D). These results demonstrate that LOVE is able to accurately cluster even if somewhat less than optimal parameter choices are made and are analogous to those observed for overlapping clustering.

Afterword. We have shown that LOVE, a variable clustering with statistical guarantees could be well-used on a gene-expression dataset. We demonstrated competitive performances with respect to state-of-the-art methods, based on several metrics, including the biological significance of the clusters as well as the correct identification of overlaps. Next, we used LOVE on a cohort of HIV controllers and chronic progressors and showed that it was able to accurately cluster these two distinct clinical phenotypes in a non-overlapping fashion. Thus, we successfully applied this method across two completely different datasets with inherently different structures, to generate overlapping and non-overlapping clusters respectively. Our findings demonstrate that LOVE can be used to address both basic sciences and clinical questions. We believe that LOVE will be widely adopted by many biological researchers because of its unique properties described above.

Figure 5.2 – Humoral immune measurements experiment



Appendix

A The LOVE method of Bing et al., 2017

LOVE considers the following latent factor model,

$$X = AZ + E \tag{A.1}$$

where $X \in \mathbb{R}^p$ represents p genes, $Z \in \mathbb{R}^K$ denotes K latent factors (biological functions), $A \in \mathbb{R}^{p\times K}$ is the membership matrix assigning p genes to K groups and E denotes some random error. In model (A.1), X, Z and E are considered random with $\mathbb{E}[E] = 0$, $\mathrm{Cov}(Z) = C$ and $\mathrm{Cov}(E) = \Gamma$. One also assume that $\Gamma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is diagonal and C is strictly positive definite. Without loss of generality, X and Z have mean zero since one can always subtract their means. Only X is observable and both the group number K and the membership matrix A are the parameters of interest. The final clusters of X are defined via matrix A. To be more specific, for each group $k \in \{1, \ldots, K\}$ X_j are clustered into group k satisfying $A_{jk} \neq 0$, that is:

$$G_k = \{j \in \{1, \ldots, p\} : A_{jk} \neq 0\}, \forall k \in 1 \ldots K. \tag{A.2}$$

Since each row of A is allowed to have more than one non-zero entries, groups G_k are expected to be overlapped.

Clearly, model (A.1) is not identifiable under the current specifications. The following model specifications are also assumed:

(i) $\sum_{k=1}^K |A_{jk}| \leqslant 1$ and $\forall j \in 1 \ldots p$, $A_{j\cdot}$ is sparse

(ii) For each $k \in 1 \ldots K$, there exists at least two indices $j \in 1 \ldots p$ such that $|A_{jk}| = 1$ and $|A_{jk'}| = 0$ for any $k \neq k'$

(iii) $\Delta(C) := \min_{k\neq k'}(C_{kk} \wedge C_{k'k'} - |C_{kk'}|) > 0$

where $a \wedge b = \min(a, b)$. Specification (i) rules out the scaling ambiguities between A and Z. It also allows one variable not associated with any group, that is $A_{jk} = 0$ for all $k \in 1 \ldots K$. This has particular biological meaning since many genes are not associated with any biological functions. (ii) requires each group (biological function) to have at least two variables (genes) that are solely associated with this group. In many areas of factor analysis, this assumption is arguably the most well-received. It has practical implication since if gene X_i is as in (ii), X_i is only related with biological function Z_k and it clarifies the property of this group, which makes the multi-clustering association meaningful. The variables in (ii) are named the non-mixed variables and its set is defined as

$$\mathcal{I} = \{I_1, \ldots, I_K\}, \quad I_k\{j \in \{1, \ldots, p\} : |A_{jk}| = 1, A_{jk'} = 0 \forall k' \neq k\} \tag{A.3}$$

Specification (iii) implies $|Z_k| \neq |Z_{k'}|$ almost surely for any two latent variables and can be viewed as the minimal assumption to make two latent variables distinguishable.

Under model (A.1) if (i) - (iii) hold, Theorems 1 and 2 in Bing et al., 2017 show that K and I are identifiable from $\Sigma := \mathrm{Cov}(X)$ up to a group permutation. Moreover, A is also identifiable from Σ

up to a $K \times K$ signed permutation. Notice that entries of A can have both positive and negative signs. The signs also have meaning for clustering X since if two gene variables X_i and X_j have the same sign for a biological function Z_k, it implies that X_i and X_j are associated with function Z_k in the same direction. However, the direction itself is not identifiable.

In practice, instead of having access to Σ, one only has access to n i.i.d. copies of $X = (X_1 \dots X_n)^T$. By using the empirical covariance estimator $\widehat{\Sigma} = \frac{1}{n} X^T X$, assuming X has sub-Gaussian tail, one can prove similar results of esimates of $\widehat{\mathcal{I}}$ and $\widehat{A}$ as shown in Theorems 3 and 4 Bing et al., 2017 under suitable conditions. After estimating A, an estimate the clusters G_k is given by

$$\widehat{G}_k := \{j \in 1 \dots p : \widehat{A}_{jk} \neq 0\}, \forall k \in 1 \dots \widehat{K} \tag{A.4}$$

Under suitable conditions, Part 3 of Remark 4 in Bing et al., 2017 guarantees that $\widehat{G}_k$ retrieves the underlying cluster for all $1 \leqslant k \leqslant K$ with high probability, up to label permutation.

Lastly we give the main steps of the LOVE algorithm:

1. Estimate the set of non-mixed variables $\widehat{\mathcal{I}}$, and its partition into non-overlapping clusters (see Figure 5.3-B)

2. For $\widehat{C}$ a moment estimator of $C = \mathrm{Cov}(Z)$ derived from the empirical covariance matrix and $\widehat{\mathcal{I}}$, λ tuning parameter, estimate the precision matrix C^{-1} by solving the linear program

$$(\widehat{\Omega}, \widehat{t}) := \underset{\text{sym.}\Omega, t \in \mathbb{R}_+}{\operatorname{argmin}} \; t \tag{A.5}$$

$$\text{s.t.} \quad \|\Omega \widehat{C} - I\|_\infty \leqslant \lambda t, \quad \|\Omega\|_{\infty,1} \leqslant t. \tag{A.6}$$

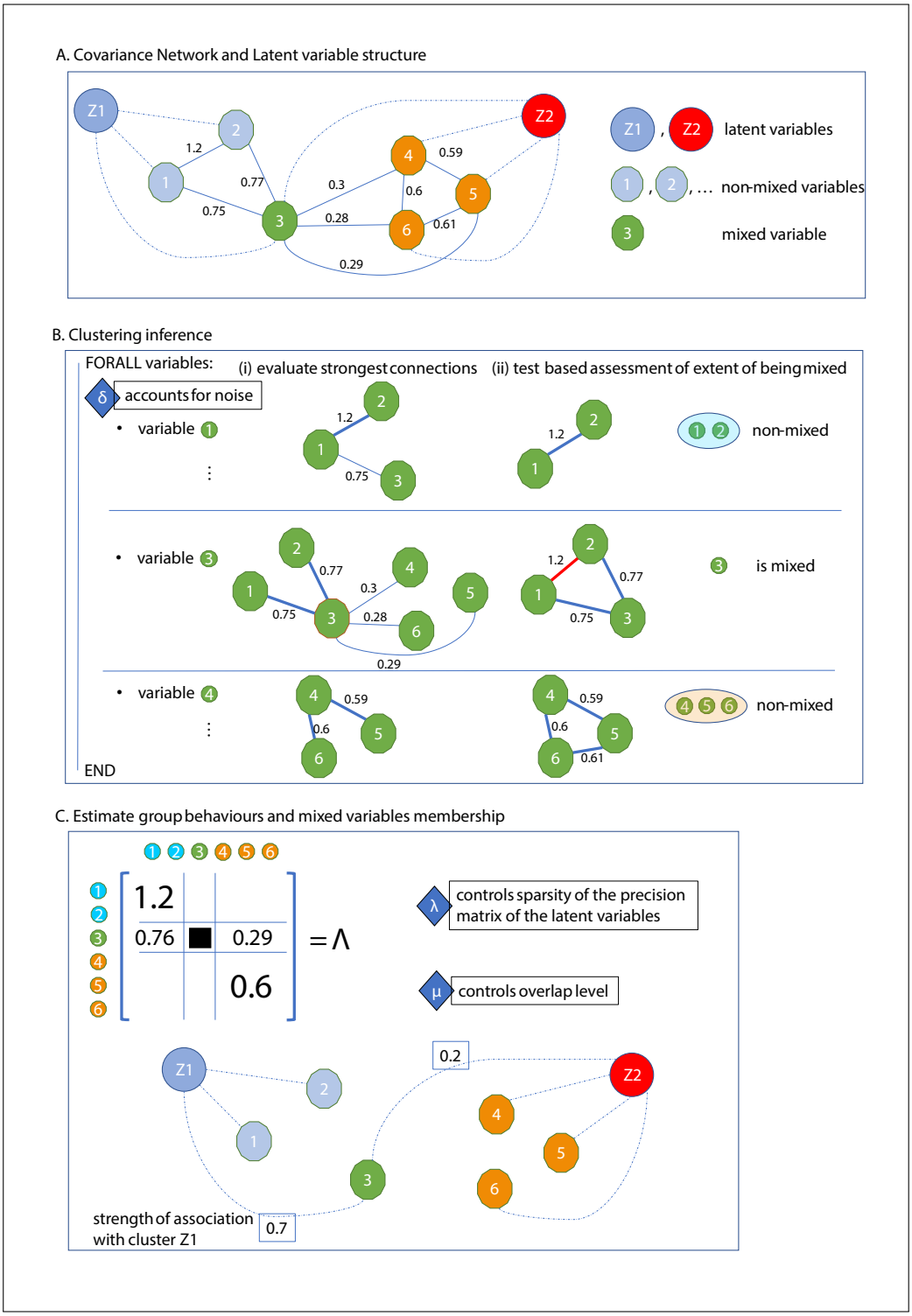
3. Estimate the strength of association of each node $j \in 1 \dots p$ to a cluster by quantity $\widehat{\Omega}\widehat{\beta}_j$, then for μ tuning parameter sparsify through a constrained optimization procedure:

$$\widehat{A}_{j\cdot} = \underset{\beta \in \mathbb{R}^{\widehat{K}}}{\operatorname{argmax}} \|\beta\|_1 \tag{A.7}$$

$$\text{s.t.} \|\beta - \widehat{\Omega}\widehat{\beta}_j\|_\infty \leqslant \mu \tag{A.8}$$

where $\widehat{\beta}_j$ is a moment estimator derived from the empirical covariance matrix and $\widehat{\mathcal{I}}$. The non-null components of $\widehat{A}_{j\cdot}$ indicate overlapping groups that share variable j.

Figure 5.3 – Schematics operations for the LOVE algorithm



Bibliographie

Abbe, Emmanuel, Afonso S. Bandeira et Georgina Hall (2016). « Exact Recovery in the Stochastic Block Model ». In : *IEEE Transactions on Information Theory* 62, p. 471–487.

Abbe, Emmanuel et Colin Sandon (2015). « Community Detection in General Stochastic Block models : Fundamental Limits and Efficient Algorithms for Recovery ». In : *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, p. 670–688.

Albert, R., H. Jeong et A.L. Barabasi (2000). « Error and attack tolerance of complex networks ». In : *Nature* 406.6794, p. 378–382.

Aloise, Daniel, Amit Deshpande, Pierre Hansen et Preyas Popat (2009). « NP-hardness of Euclidean Sum-of-squares Clustering ». In : *Mach. Learn.* 75.2, p. 245–248. ISSN : 0885-6125. DOI : 10.1007/s10994-009-5103-0. URL : http://dx.doi.org/10.1007/s10994-009-5103-0.

Alon, Noga et Assaf Naor (2006). « Approximating the Cut-Norm via Grothendieck's Inequality ». In : *SIAM Journal on Computing* 35.4, p. 787–803. ISSN : 0097-5397. DOI : 10.1137/S0097539704441629. URL : http://dx.doi.org/10.1137/S0097539704441629.

Alter, Galit, Karen G Dowell, Eric P Brown, Todd J Suscovich, Anastassia Mikhailova, Alison E Mahan, Bruce D Walker, Falk Nimmerjahn, Chris Bailey-Kellogg et Margaret E Ackerman (2018). « High-resolution definition of humoral immune response correlates of effective immunity against HIV ». In : *Molecular Systems Biology* 14.3. DOI : 10.15252/msb.20177881. eprint : http://msb.embopress.org/content/14/3/e7881.full.pdf. URL : http://msb.embopress.org/content/14/3/e7881.

Amini, Arash A. et Elizaveta Levina (2018). « On semidefinite relaxations for the block model ». In : *Ann. Statist.* 46.1, p. 149–179. DOI : 10.1214/17-AOS1545. URL : https://doi.org/10.1214/17-AOS1545.

Andersen, Martin Skovgaard, Joachim Dahl, Zhang Liu et Lieven Vandenberghe (2011). « Interior-point methods for large-scale cone programming ». In : *Optimization For Machine Learning*. MIT Press. ISBN : 9780262016469.

Arthur, David et Sergei Vassilvitskii (2007). « K-means++ : The Advantages of Careful Seeding ». In : *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '07. Philadelphia, PA, USA : Society for Industrial et Applied Mathematics, p. 1027–1035. ISBN : 978-0-898716-24-5. URL : http://dl.acm.org/citation.cfm?id=1283383.1283494.

Ashburner, M et al. (2000). « Gene ontology : tool for the unification of biology. The Gene Ontology Consortium ». In : *Nature genetics* 25.1, 25—29. ISSN : 1061-4036. DOI : 10.1038/75556. URL : http://europepmc.org/articles/PMC3037419.

Awasthi, Pranjal, Moses Charikar, Ravishankar Krishnaswamy et Ali Kemal Sinop (2015). « The Hardness of Approximation of Euclidean k-Means ». In : *Symposium on Computational Geometry*.

Azizyan, M., A. Singh et L. Wasserman (2013). « Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation ». In : *Proceedings of the 26th International Conference on Neural Information Processing Systems*. NIPS'13. Lake Tahoe, Nevada : Curran Associates Inc., p. 2139–2147. URL : http://dl.acm.org/citation.cfm?id=2999792.2999851.

Banerjee, Onureena, Laurent El Ghaoui et Alexandre d'Aspremont (2008). « Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data ». In : *Journal of Machine learning research* 9.Mar, p. 485–516.

Banks, J., C. Moore, R. Vershynin, N. Verzelen et J. Xu (2018). « Information-Theoretic Bounds and Phase Transitions in Clustering, Sparse PCA, and Submatrix Localization ». In : *IEEE Transactions on Information Theory* 64.7, p. 4872–4894. ISSN : 0018-9448. DOI : 10.1109/TIT.2018.2810020.

Bellec, Pierre, Vincent Perlbarg, Saâd Jbabdi, Mélanie Pélégrini-Issac, Jean-Luc Anton, Julien Doyon et Habib Benali (2006). « Identification of large-scale networks in the brain using fMRI ». In : *Neuroimage* 29.4, p. 1231–1243.

Bellec, Pierre C (2014). « Concentration of quadratic forms under a Bernstein moment assumption ». In : *Technical Report, Ecole Polytechnique*.

Bernardes, Juliana S., Fabio RJ Vieira, Lygia MM Costa et Gerson Zaverucha (2015). « Evaluation and improvements of clustering algorithms for detecting remote homologous protein families ». In : *BMC Bioinformatics* 16.1, p. 1–14. ISSN : 1471-2105. DOI : 10.1186/s12859-014-0445-4. URL : http://dx.doi.org/10.1186/s12859-014-0445-4.

Berthet, Q., P. Rigollet et P. Srivastava (2018). « Exact recovery in the Ising blockmodel ». In : *Annals of Statistics (to appear)*, arXiv :1612.03880.

Berthet, Quentin et Philippe Rigollet (2013). « Complexity Theoretic Lower Bounds for Sparse Principal Component Detection ». In : *Proceedings of the 26th Annual Conference on Learning Theory*. Sous la dir. de Shai Shalev-Shwartz et Ingo Steinwart. T. 30. Proceedings of Machine Learning Research. Princeton, NJ, USA : PMLR, p. 1046–1066. URL : http://proceedings.mlr.press/v30/Berthet13.html.

Bezdek, J. C., R. Ehrlich et W. Full (1984). « FCM : The fuzzy c-means clustering algorithm ». In : *Computers and Geosciences* 10, p. 191–203. DOI : 10.1016/0098-3004(84)90020-7.

Bing, Mike, Jishnu Das et Martin Royer (2018). « Latent model-based clustering for biological discovery ». In : *(to submit)*.

Bing, X., F. Bunea, Y. Ning et M. Wegkamp (2017). « Adaptive Estimation in Structured Factor Models with Applications to Overlapping Clustering ». In : *ArXiv e-prints*. arXiv : 1704.06977 [stat.ME].

Boumal, N, PA Absil et C Cartis (2018). « Global rates of convergence for nonconvex optimization on manifolds ». In : *IMA Journal of Numerical Analysis*.

Boumal, N., V. Voroninski et A. S. Bandeira (2018). « Deterministic guarantees for Burer-Monteiro factorizations of smooth semidefinite programs ». In : *ArXiv e-prints*. arXiv : 1804.02008 [math.OC].

Bouveyron, Charles et Camille Brunet-Saumard (2014). « Model-based clustering of high-dimensional data : A review ». In : *Computational Statistics & Data Analysis* 71, p. 52–78.

Boyd, S., N. Parikh, E. Chu, B. Peleato et J. Eckstein (2011). « Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers ». In : *Found. Trends Mach. Learn.* 3.1, p. 1–122. ISSN : 1935-8237. DOI : 10.1561/2200000016. URL : http://dx.doi.org/10.1561/2200000016.

Bunea, F., C. Giraud, M. Royer et N. Verzelen (2016). « PECOK : a convex optimization approach to variable clustering ». In : *arXiv e-prints arXiv :1606.05100*. arXiv : 1606.05100 [math.ST].

Bunea, F., C. Giraud, X. Luo, M. Royer et N. Verzelen (2018a). « Model assisted variable clustering : minimax-optimal recovery and algorithms ». In : *Preprint*.

— (2018b). « Supplement to : Model assisted variable clustering : minimax-optimal recovery and algorithms ». In : *Preprint*.

Bunea, Florentina, Christophe Giraud et Xi Luo (2015). « Minimax Optimal Variable Clustering in *G*-models via Cord ». In : *arXiv preprint arXiv :1508.01939*.

Bunea, Florentina, Yiyuan She et Marten H. Wegkamp (2011). « Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices ». In : *The Annals of Statistics* 39.2, p. 1282–1309. ISSN : 00905364. URL : http://www.jstor.org/stable/29783674.

Burer, S. et R. Monteiro (2003). « A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization ». In : *Math. Program., Ser. B 95 :329-357*.

Cai, T. Tony et Xiaodong Li (2015). « Robust and computationally feasible community detection in the presence of arbitrary outlier nodes ». In : *Ann. Statist.* 43.3, p. 1027–1059. DOI : 10.1214/14-AOS1290. URL : https://doi.org/10.1214/14-AOS1290.

Carson, Timothy, Dustin G. Mixon et Soledad Villar (2017). « Manifold optimization for k-means clustering ». In : *2017 International Conference on Sampling Theory and Applications (SampTA)*, p. 73–77.

Celeux, Gilles et Gérard Govaert (1992). « A Classification EM Algorithm for Clustering and Two Stochastic Versions ». In : *Comput. Stat. Data Anal.* 14.3, p. 315–332. ISSN : 0167-9473. DOI : 10.1016/0167-9473(92)90042-E.

Chen, Yudong et Jiaming Xu (2016). « Statistical-computational Tradeoffs in Planted Problems and Submatrix Localization with a Growing Number of Clusters and Submatrices ». In : *J. Mach. Learn. Res.* 17.1, p. 882–938. ISSN : 1532-4435. URL : http://dl.acm.org/citation.cfm?id=2946645.2946672.

Chong, M, C Bhushan, AA Joshi, S Choi, JP Haldar, DW Shattuck, RN Spreng et RM Leahy (2017). « Individual parcellation of resting fMRI with a group functional connectivity prior ». In : *NeuroImage* 156, p. 87–100.

Chrétien, S., C. Dombry et A. Faivre (2016). « A Semi-Definite Programming approach to low dimensional embedding for unsupervised clustering ». In : *CoRR* abs/1606.09190. URL : http://arxiv.org/abs/1606.09190.

Craddock, R Cameron, G Andrew James, Paul E Holtzheimer, Xiaoping P Hu et Helen S Mayberg (2012). « A whole brain fMRI atlas generated via spatially constrained spectral clustering ». In : *Human brain mapping* 33.8, p. 1914–1928.

Das, Jishnu, Mohammed Jaaved et Yu. Haiyuan (2012). « Genome-Scale Analysis of Interaction Dynamics Reveals Organization of Biological Networks. » In : *Bioinformatics 28.14*.

Das, Jishnu et Haiyuan Yu (2012). « HINT : High-quality protein interactomes and their applications in understanding human disease ». In : *BMC Systems Biology* 6.1, p. 92. ISSN : 1752-0509. DOI : 10.1186/1752-0509-6-92. URL : https://doi.org/10.1186/1752-0509-6-92.

Dasgupta, Sanjoy et Leonard Schulman (2007). « A Probabilistic Analysis of EM for Mixtures of Separated, Spherical Gaussians ». In : *J. Mach. Learn. Res.* 8, p. 203–226. ISSN : 1532-4435. URL : http://dl.acm.org/citation.cfm?id=1248659.1248666.

Davis, C. et W. M. Kahan (1970). « The Rotation of Eigenvectors by a Perturbation. III ». In : *SIAM Journal on Numerical Analysis* 7, p. 1–46. DOI : 10.1137/0707001.

Dempster, A. P., N. M. Laird et D. B. Rubin (1977). « Maximum Likelihood from Incomplete Data via the EM Algorithm ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, p. 1–38. ISSN : 00359246. URL : http://www.jstor.org/stable/2984875.

Eisenberg, Eli et Erez Y. Levanon (2013). « Human housekeeping genes, revisited ». In : *Trends in Genetics* 29.10. Human Genetics, p. 569 –574. ISSN : 0168-9525. DOI : https://doi.org/10.1016/j.tig.2013.05.010. URL : http://www.sciencedirect.com/science/article/pii/S0168952513000899.

Fei, Yingjie et Yudong Chen (2017). « Exponential error rates of SDP for block models : Beyond Grothendieck's inequality ». In : *ArXiv e-prints*. arXiv : 1705.08391 [stat.ML].

Fraley, C. et A. E. Raftery (2002). « Model-Based Clustering, Discriminant Analysis, and Density Estimation ». In : *Journal of the American Statistical Association* 97.458, p. 611–631. ISSN : 01621459. URL : http://www.jstor.org/stable/3085676.

Frey, Nicolas Frei dit et al. (2014). « Functional analysis of Arabidopsisimmune-related MAPKs uncovers a role for MPK3 as negative regulator of inducible defences ». In : *Genome Biology* 15.6, p. 1–22. ISSN : 1474-760X. DOI : 10.1186/gb-2014-15-6-r87. URL : http://dx.doi.org/10.1186/gb-2014-15-6-r87.

Friston, Karl J, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith et Richard SJ Frackowiak (1994). « Statistical parametric maps in functional imaging : a general linear approach ». In : *Human brain mapping* 2.4, p. 189–210.

Ge, Rong, Chi Jin et Yi Zheng (2017). « No Spurious Local Minima in Nonconvex Low Rank Problems : A Unified Geometric Analysis ». In : *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, p. 1233–1242. URL : http://proceedings.mlr.press/v70/ge17a.html.

Ge, Rong, Jason D. Lee et Tengyu Ma (2016). « Matrix Completion has No Spurious Local Minimum ». In : *Advances in Neural Information Processing Systems 29 : Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, p. 2973–2981. URL : http://papers.nips.cc/paper/6048-matrix-completion-has-no-spurious-local-minimum.

Giraud, C. et N. Verzelen (2018). « Partial recovery bounds for clustering with the relaxed Kmeans ». In : *ArXiv e-prints*. arXiv : 1807.07547 [math.ST].

Glasser, M.F. et al. (2016). « A Multi-modal parcelation of human cerebral cortex ». In : *Nature* 536, p. 171–178.

Goemans, Michel X. et David P. Williamson (1995). « Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming ». In : *J. ACM* 42.6, p. 1115–1145. ISSN : 0004-5411. DOI : 10.1145/227683.227684. URL : http://doi.acm.org/10.1145/227683.227684.

Guédon, Olivier et Roman Vershynin (2016). « Community Detection in Sparse Networks via Grothendieck's Inequality ». In : *Probability Theory and Related Fields*. URL : https://hal-upec-upem.archives-ouvertes.fr/hal-01262623.

Helmberg, C. et F. Rendl (2000). « A Spectral Bundle Method for Semidefinite Programming ». In : *SIAM Journal on Optimization* 10.3, p. 673–696. DOI : 10.1137/S1052623497328987. eprint : https://doi.org/10.1137/S1052623497328987. URL : https://doi.org/10.1137/S1052623497328987.

James, George Andrew, Onder Hazaroglu et Keith A Bush (2016). « A human brain atlas derived via n-cut parcellation of resting-state and task-based fMRI data ». In : *Magnetic resonance imaging* 34.2, p. 209–218.

Javanmard, Adel, Andrea Montanari et Federico Ricci-Tersenghi (2016). « Phase transitions in semidefinite relaxations ». In : *Proceedings of the National Academy of Sciences* 113.16, E2218–E2223. ISSN : 0027-8424. DOI : 10.1073/pnas.1523097113. eprint : http://www.pnas.org/content/113/16/E2218.full.pdf. URL : http://www.pnas.org/content/113/16/E2218.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai et A. L. Barabasi (2000). « The large-scale organization of metabolic networks ». In : *Nature* 407.6804, p. 651–654. ISSN : 00280836. DOI : 10.1038/35036627. eprint : cond-mat/0010278. URL : http://dx.doi.org/10.1038/35036627.

Jiang, Daxin, Chun Tang et Aidong Zhang (2004). « Cluster analysis for gene expression data : a survey ». In : *IEEE Transactions on Knowledge and Data Engineering* 16.11, p. 1370–1386. ISSN : 1041-4347. DOI : 10.1109/TKDE.2004.68.

Journée, M., F. Bach, P.-A. Absil et R. Sepulchre (2010). « Low-Rank Optimization on the Cone of Positive Semidefinite Matrices ». In : *SIAM J. on Optimization* 20.5, p. 2327–2351. ISSN : 1052-6234. DOI : 10.1137/080731359. URL : http://dx.doi.org/10.1137/080731359.

Koltchinskii, Vladimir et Karim Lounici (2017). « Concentration inequalities and moment bounds for sample covariance operators ». In : *Bernoulli* 23.1, p. 110–133. DOI : 10.3150/15-BEJ730. URL : https://doi.org/10.3150/15-BEJ730.

Kong, Ru, Jingwei Li, Nanbo Sun, Mert Sabuncu, Hesheng Liu, Andrew Schaefer, Xi-Nian Zuo, Avram Holmes, Simon Eickhoff et Thomas Yeo (2018). « Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality and Emotion ». In :

Kumar, A., Y. Sabharwal et S. Sen (2004). « A simple linear time $(1 + \epsilon)$-approximation algorithm for k-means clustering in any dimensions ». In : *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, p. 454–462. DOI : 10.1109/FOCS.2004.7.

Kumar, Amit et Ravindran Kannan (2010). « Clustering with Spectral Norm and the k-Means Algorithm ». In : *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. FOCS '10. Washington, DC, USA : IEEE Computer Society, p. 299–308. ISBN : 978-0-7695-4244-7. DOI : 10.1109/FOCS.2010.35. URL : http://dx.doi.org/10.1109/FOCS.2010.35.

Le, Can M, Elizaveta Levina et Roman Vershynin (2014). « Optimization via Low-rank Approximation for Community Detection in Networks ». In : *arXiv preprint arXiv :1406.0067*.

Lei, Jing et Alessandro Rinaldo (2015). « Consistency of spectral clustering in stochastic block models ». In : *Ann. Statist.* 43.1, p. 215–237. ISSN : 0090-5364. DOI : 10.1214/14-AOS1274. URL : http://dx.doi.org/10.1214/14-AOS1274.

Lei, Jing et Lingxue Zhu (2014). « A Generic Sample Splitting Approach for Refined Community Recovery in Stochastic Block Models ». In : *arXiv preprint arXiv :1411.1469*.

Lesieur, Thibault, Caterina De Bacco, Jess Banks, Florent Krzakala, Cristopher Moore et Lenka Zdeborová (2016). « Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering ». In : *Allerton*. IEEE, p. 601–608.

Li, Xiaodong, Yang Li, Shuyang Ling, Thomas Strohmer et Ke Wei (2017). « When Do Birds of a Feather Flock Together ? K-Means, Proximity, and Conic Programming ». In : *ArXiv e-prints*. arXiv : 1710.06008 [math.OC].

Lloyd, S. (1982). « Least Squares Quantization in PCM ». In : *IEEE Trans. Inf. Theor.* 28.2, p. 129–137. ISSN : 0018-9448. DOI : 10.1109/TIT.1982.1056489. URL : http://dx.doi.org/10.1109/TIT.1982.1056489.

Lovasz, L. (1979). « On the Shannon Capacity of a Graph ». In : *IEEE Trans. Inf. Theor.* 25.1, p. 1–7. ISSN : 0018-9448. DOI : 10.1109/TIT.1979.1055985. URL : https://doi.org/10.1109/TIT.1979.1055985.

Lu, Yu et Harrison H. Zhou (2016). « Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants ». In : *CoRR* abs/1612.02099.

Ma, Z. et Z. Ma (2017). « Exploration of Large Networks with Covariates via Fast and Universal Latent Space Model Fitting ». In : *ArXiv e-prints*. arXiv : 1705.02372 [stat.ME].

MacQueen, J. (1967). « Some methods for classification and analysis of multivariate observations ». In : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*. Berkeley, Calif. : University of California Press, p. 281–297. URL : http://projecteuclid.org/euclid.bsmsp/1200512992.

Massart, Pascal (2007). *Concentration inequalities and model selection*. T. 1896. Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Springer, Berlin, p. xiv+337. ISBN : 978-3-540-48497-4 ; 3-540-48497-3.

Matsushita, Ryosuke et Toshiyuki Tanaka (2013). « Low-rank matrix reconstruction and clustering via approximate message passing ». In : *Advances in Neural Information Processing Systems 26*. Sous la dir. de C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani et K. Q. Weinberger. Curran Associates, Inc., p. 917–925. URL : http://papers.nips.cc/paper/5074-low-rank-matrix-reconstruction-and-clustering-via-approximate-message-passing.pdf.

Mclachlan, G et K Basford (1988). *Mixture Models : Inference and Applications to Clustering*. T. 38.

McSherry, F. (2001). « Spectral Partitioning of Random Graphs ». In : *Proceedings of the 42Nd IEEE Symposium on Foundations of Computer Science*. FOCS '01. Washington, DC, USA : IEEE Computer Society, p. 529–. ISBN : 0-7695-1390-5. URL : http://dl.acm.org/citation.cfm?id=874063.875554.

Meilă, Marina (2007). « Comparing clusterings—an information based distance ». In : *Journal of Multivariate Analysis* 98.5, p. 873 –895. ISSN : 0047-259X. DOI : https://doi.org/10.1016/j.jmva.2006.11.013. URL : http://www.sciencedirect.com/science/article/pii/S0047259X06002016.

Mixon, D. G., S. Villar et R. Ward (2016). « Clustering subgaussian mixtures with k-means ». In : *2016 IEEE Information Theory Workshop (ITW)*, p. 211–215. DOI : 10.1109/ITW.2016.7606826.

Moitra, Ankur, William Perry et Alexander S. Wein (2016). « How Robust Are Reconstruction Thresholds for Community Detection ? » In : *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*. STOC '16. Cambridge, MA, USA : ACM, p. 828–841. ISBN : 978-1-4503-4132-5. DOI : 10.1145/2897518.2897573. URL : http://doi.acm.org/10.1145/2897518.2897573.

Montanari, Andrea et Subhabrata Sen (2016). « Semidefinite Programs on Sparse Random Graphs and Their Application to Community Detection ». In : *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*. STOC '16. Cambridge, MA, USA : ACM, p. 814–827. ISBN : 978-1-4503-4132-5. DOI : 10.1145/2897518.2897548. URL : http://doi.acm.org/10.1145/2897518.2897548.

Mossel, Elchanan, Joe Neeman et Allan Sly (2014). « Consistency thresholds for binary symmetric block models ». In : *arXiv preprint arXiv :1407.1591*.

Nepusz, Tamás, Haiyuan Yu et Alberto Paccanaro (2012). « Detecting overlapping protein complexes in protein-protein interaction networks ». In : *Nature Methods* 9, 471 EP –. URL : http://dx.doi.org/10.1038/nmeth.1938.

Nesterov, Yurii et Arkadii Nemirovskii (1994). *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial et Applied Mathematics. DOI : 10.1137/1.9781611970791. eprint : https://epubs.siam.org/doi/pdf/10.1137/1.9781611970791. URL : https://epubs.siam.org/doi/abs/10.1137/1.9781611970791.

Pearson, Karl (1894). « III. Contributions to the mathematical theory of evolution ». In : *Philosophical Transactions of the Royal Society of London A : Mathematical, Physical and Engineering Sciences* 185, p. 71–110. ISSN : 0264-3820. DOI : 10.1098/rsta.1894.0003.

Pedregosa, F. et al. (2011). « Scikit-learn : Machine Learning in Python ». In : *Journal of Machine Learning Research* 12, p. 2825–2830.

Peng, Jiming et Yu Wei (2007). « Approximating K-means-type Clustering via Semidefinite Programming ». In : *SIAM J. on Optimization* 18.1, p. 186–205. ISSN : 1052-6234. DOI : 10.1137/050641983. URL : http://dx.doi.org/10.1137/050641983.

Perry, William et Alexander S. Wein (2017). « A semidefinite program for unbalanced multisection in the stochastic block model ». In : *2017 International Conference on Sampling Theory and Applications (SampTA)*, p. 64–67.

Poldrack, Russell A (2007). « Region of interest analysis for fMRI ». In : *Social cognitive and affective neuroscience* 2.1, p. 67–70.

Power, Jonathan D, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar et al. (2011). « Functional network organization of the human brain ». In : *Neuron* 72.4, p. 665–678.

Ricci-Tersenghi, Federico, Adel Javanmard et Andrea Montanari (2016). « Performance of a community detection algorithm based on semidefinite programming ». In : *Journal of Physics : Conference Series* 699.1, p. 012015. URL : http://stacks.iop.org/1742-6596/699/i=1/a=012015.

Rigollet, P. (2015). *High-Dimensional Statistics*. Massachusetts Institute of Technology : MIT OpenCourseWare.

Rolland, Thomas et al. (2014). « A Proteome-Scale Map of the Human Interactome Network ». In : *Cell* 159.5, p. 1212–1226. ISSN : 0092-8674. DOI : 10.1016/j.cell.2014.10.050. URL : https://doi.org/10.1016/j.cell.2014.10.050.

Royer, Martin (2017). « Adaptive Clustering through Semidefinite Programming ». In : *Advances in Neural Information Processing Systems 30*. Sous la dir. d'I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett. Curran Associates, Inc., p. 1795–1803. URL : http://papers.nips.cc/paper/6776-adaptive-clustering-through-semidefinite-programming.pdf.

— (October, 2017). ADMM *implementation of* PECOK. https://github.com/martinroyer/pecok.

Rudelson, Mark et Roman Vershynin (2013). « Hanson-Wright inequality and sub-gaussian concentration ». In : *Electron. Commun. Probab.* 18, no. 82, 1–9. ISSN : 1083-589X. DOI : 10.1214/ECP.v18-2865. URL : http://ecp.ejpecp.org/article/view/2865.

Sadanand, Saheli et al. (2018). « Temporal variation in HIV-specific IgG subclass antibodies during acute infection differentiates spontaneous controllers from chronic progressors ». In : *AIDS* 32.4. ISSN : 0269-9370. URL : https://journals.lww.com/aidsonline/Fulltext/2018/02200/Temporal_variation_in_HIV_specific_IgG_subclass.4.aspx.

Steinhaus, H. (1956). « Sur la division des corp materiels en parties ». In : *Bull. Acad. Polon. Sci* 1, p. 801–804.

Tropp, J. A. (2015). « An Introduction to Matrix Concentration Inequalities ». In : *ArXiv e-prints*. arXiv : 1501.01571 [math.PR].

Vandenberghe, L. et S. Boyd (1996). « Semidefinite Programming ». In : *SIAM Review* 38.1, p. 49–95. DOI : 10.1137/1038003. eprint : https://doi.org/10.1137/1038003. URL : https://doi.org/10.1137/1038003.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*. Chapter 5 of : Compressed Sensing, Theory et Applications. Cambridge University Press.

Vershynin, Roman (2010). « Introduction to the non-asymptotic analysis of random matrices ». In : *arXiv preprint arXiv :1011.3027*.

Verzelen, N. et E. Arias-Castro (2014). « Detection and Feature Selection in Sparse Mixture Models ». In : *arXiv e-prints arXiv :1405.1478*. arXiv : 1405.1478 [math.ST].

Vo, Tommy ?V et al. (2016). « A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human ». In : *Cell* 164.1, p. 310–323. ISSN : 0092-8674. DOI : 10.1016/j.cell.2015.11.037. URL : https://doi.org/10.1016/j.cell.2015.11.037.

Wang, Jing, Dexter Duncan, Zhiao Shi et Bing Zhang (2013). « WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) : update 2013 ». In : *Nucleic Acids Research* 41.W1, W77–W83. DOI : 10.1093/nar/gkt439. eprint : /oup/backfile/content_public/journal/nar/41/w1/10.1093_nar_gkt439/3/gkt439.pdf. URL : http://dx.doi.org/10.1093/nar/gkt439.

Ward, J. H. (1963). « Hierarchical Grouping to Optimize an Objective Function ». In : *Journal of the American Statistical Association* 58.301, p. 236–244. DOI : 10.1080/01621459.1963.10500845. eprint : http://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500845. URL : http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

Wiwie, Christian, Jan Baumbach et Richard Röttger (2015). « Comparing the performance of biomedical clustering methods ». In : *Nature Methods* 12, 1033 EP –. URL : http://dx.doi.org/10.1038/nmeth.3583.

Xing, Eric P. et Michael I. Jordan (2003). *On Semidefinite Relaxation for Normalized k-cut and Connections to Spectral Clustering*. Rapp. tech. UCB/CSD-03-1265. EECS Department, University of California, Berkeley. URL : http://www2.eecs.berkeley.edu/Pubs/TechRpts/2003/6240.html.

Xue, Gui, Adam R Aron et Russell A Poldrack (2008). « Common neural substrates for inhibition of spoken and manual responses ». In : *Cerebral Cortex* 18.8, p. 1923–1932.

Yan, B. et P. Sarkar (2016). « Convex Relaxation for Community Detection with Covariates ». In : *arXiv e-prints arXiv :1607.02675*. arXiv : 1607.02675 [stat.ME].

Yang, Liuqin, Defeng Sun et Kim-Chuan Toh (2014). « SDPNAL+ : A Majorized Semismooth Newton-CG Augmented Lagrangian Method for Semidefinite Programming with Nonnegative Constraints ». In : *Mathematical Programming Computation*. T. 7.

Yeo, B.T. et al. (2011). « The organization of the human cerebral cortex estimated by intrinsic functional connectivity ». In : *Journal of Neurophysiology* 106, p. 1125 –1165.

Yu, Haiyuan et al. (2008). « High-Quality Binary Protein Interaction Map of the Yeast Interactome Network ». In : *Science* 322.5898, p. 104–110. ISSN : 0036-8075. DOI : 10.1126/science.

1158684. eprint : http://science.sciencemag.org/content/322/5898/104.full.pdf. URL : http://science.sciencemag.org/content/322/5898/104.

Zaag, Rim et al. (2015). « GEM2Net : from gene expression modeling to -omics networks, a new CATdb module to investigate *Arabidopsis thaliana* genes involved in stress response ». In : *Nucleic Acids Research* 43.Database-Issue, p. 1010–1017. DOI : 10.1093/nar/gku1155. URL : http://dx.doi.org/10.1093/nar/gku1155.

Zha, Hongyuan, Xiaofeng He, Chris Ding, Horst Simon et Ming Gu (2001). « Spectral Relaxation for K-means Clustering ». In : *Proceedings of the 14th International Conference on Neural Information Processing Systems : Natural and Synthetic*. NIPS'01. Vancouver, British Columbia, Canada : MIT Press, p. 1057–1064. URL : http://dl.acm.org/citation.cfm?id=2980539.2980675.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Optimalité statistique du partitionnement par l'optimisation convexe

Mots Clefs : partitionnement, K-moyennes, minimax, optimisation, semi-défini positif

Résumé : Ces travaux traitent de la problématique du partitionnement d'un ensemble d'observations ou de variables en groupes d'éléments similaires. Elle sert de nombreuses applications essentielles comme la classification de gènes en biologie. Les travaux modélisent la notion de similarité entre éléments pour analyser les propriétés statistiques d'algorithmes de partitionnement, comme l'estimateur des K-moyennes. Ce dernier est équivalent au maximum de vraisemblance quand les groupes considérés sont homoscedastiques ; dans le cas contraire, on s'aperçoit que l'estimateur est biaisé, en ce qu'il tend à séparer les groupes ayant une plus grande dispersion. En utilisant une formulation équivalente qui fait intervenir l'optimisation semi-définie positive, on propose une correction opérationnelle de ce biais. On construit et étudie ainsi des algorithmes de complexité polynomiale qui sont quasi-minimax pour le partitionnement exact dans les deux contextes étudiés. Ces résultats s'interprètent dans le cadre de modèles standards comme le modèle de mélange ou le modèle à variables latentes, et s'étendent à de nouveaux modèles plus généraux et plus robustes, les modèles G-block. Les contrôles sont appuyés par des expériences extensives sur données de synthèse, ainsi que sur des jeux de données réelles. Enfin lorsqu'on cherche à améliorer l'efficacité computationnelle des algorithmes étudiés, on peut utiliser une connexion forte avec le domaine de l'optimisation convexe et notamment exploiter des techniques de relaxation de faible rang motivées par des problématiques de grande dimension.

Title : Statistically optimal clustering through convex optimisation

Keys words : clustering, K-means, minimax, optimisation, semidefinite programs

Abstract : This work focuses on the problem of point and variable clustering, that is the grouping of either similar vectors or similar components of a vector in a metric space. This has applications in many relevant fields such as gene expression data classification. Through adequate modeling of the similarity between points or variables within a cluster we analyse the statistical properties of known clustering algorithms such as K-means. When considering homoscedastic elements for all groups the K-means algorithm is equivalent to a maximum-likelihood procedure. Otherwise the algorithm shows bias in the sense that it tends to separate groups with larger dispersion, regardless of actual group separation. By using a semi definite positive reformulation of the estimator, we suggest a pattern of correction for the algorithm that leads to the construction of computational algorithm with quasiminimax properties for hard clustering of points or variables. Those results can be studied under the classical mixture model or latent variables model, and can be extended to more general and robust class of G-block models. They are supported by extensive simulation studies as well as data analysis stemming from the biological field. When focus is brought on the computational aspect of those algorithms, we exploit ideas based on a strong connexion with the domain of convex optimisation and specifically the technique of low-rank relaxation, of importance when dealing with high dimensional problems.