MODEL ASSISTED VARIABLE CLUSTERING: MINIMAX-OPTIMAL RECOVERY AND ALGORITHMS

BY FLORENTINA BUNEA[1], CHRISTOPHE GIRAUD[2,*], XI LUO[3], MARTIN ROYER[2,**]
AND NICOLAS VERZELEN[4]

[1]*Department of Statistical Science, Cornell University, fb238@cornell.edu*

[2]*Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Sud, Université Paris-Saclay,*
[*]*christophe.giraud@math.u-psud.fr;* [**]*martin.royer@math.u-psud.fr*

[3]*Department of Biostatistics and Data Science, School of Public Health, University of Texas Health Science Center at Houston, xi.rossi.luo@gmail.com*

[4]*INRA, Montpellier SupAgro, MISTEA, Université de Montpellier, nicolas.verzelen@inra.fr*

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 compare our methods with another popular clustering method, spectral clustering. Extensive simulation studies, as well as our data analyses, confirm the applicability of our approach.

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 [19, 23, 41] or protein profile data [8]. 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 cluster-ing 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 moti-vate 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 sub-sumes that the covariance matrix Σ of a centered random vector $X \in \mathbf{R}^p$ 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, [18, 20, 25, 40]. We further support these findings in Section 7, 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 $A \in \mathbf{R}^{p \times K}$ 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

$$(1.1) \qquad\qquad \Sigma = AC^*A^t + \Gamma,$$

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 $X \in \mathbf{R}^p$ follows an (exact) G^*-block covariance model.

(i) *G-Latent model.* Such a structure arises, for instance, when components of X that be-long 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 asso-ciation with the same unobservable source. Specifically, the exact block-covariance model (1.1) holds, with a diagonal matrix Γ, when

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

with $\mathrm{Cov}(Z_{k(a)}, E_a) = 0$, $\mathrm{Cov}(Z) = C^*$, and the individual fluctuations E_a are uncorre-lated, and thus E has diagonal covariance matrix Γ. The index assignment function $k :$ $\{1, \ldots, p\} \to \{1, \ldots, K\}$ 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 [36], requires averaging the obser-vations 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 in-terpretable brain area. These new variables are then used for downstream analyses. From this perspective, model (1.2) can be used in practice (see, e.g., [7]) 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 func-tional systems, especially based on the covariance matrix of X, is becoming increasingly important; see, for example, [18, 20, 37, 40]. Accurate data-driven clustering methods also enable studying the cluster differences across subjects [17] or experimental conditions [22].

(ii) *The Ising block model.* The Ising block model has been proposed in [10] for modeling 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

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

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 [6]. 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 $C^* = (\omega_{\text{in}} - \omega_{\text{out}})I_K + \omega_{\text{out}}J$ and $\Gamma = (1 - \omega_{\text{in}})I$, 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 ω_{in}, ω_{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

$$(1.4) \qquad \Sigma = ACA^t + \Gamma \quad \text{where } \Gamma \text{ has small off-diagonal entries.}$$

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, $1 \le a \le p$.

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 of 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 $\max_{c \ne a,b} |\Sigma_{ac} - \Sigma_{bc}| > 0$ 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:

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

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

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

The measure $\mathrm{COD}(a, b)$ quantifies the similarity of the covariances that X_a and X_b have, respectively, with all other variables. From this perspective, the size of $\mathrm{MCOD}(\Sigma)$ 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 semipositive 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

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

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 semipositive definite. Further, we observe that $\Delta(C^*) = 0$ implies $Z_j = Z_k$ a.s. Conversely, we prove in Corollary 2.3 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 [18] for creating a human brain atlas by partitioning appropriate covariance matrices. We also present a neuroscience-based data example in Section 7.

Formally, the two metrics are connected via the following chain of inequalities, proved in Lemma B.1 of Section 2 of the Supplementary Material [14], and valid as soon as the size of the smallest cluster is larger than one, Γ and C^* is semipositive definite (for the last inequality)

$$(1.8) \qquad 2\lambda_K(C^*) \le \Delta(C^*) \le 2\mathrm{MCOD}(\Sigma) \le 2\sqrt{\Delta(C^*)} \max_{k=1,\dots,K} \sqrt{C_{kk}^*}.$$

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 $\mathrm{MCOD}(\Sigma)$ 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 $\mathrm{MCOD}^2(\Sigma)$ 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, for $K \ge 3$ and for some numerical constant $c > 0$, no algorithm can estimate consistently clusters defined by (1.1) uniformly over covariance matrices fulfilling

$$\mathrm{MCOD}(\Sigma) \ge c\sqrt{\frac{\log(p)}{n}}.$$

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,

$$m^* = \min_{1 \le k \le K} |G_k^*|.$$

Indeed, there exists a numerical constant $c > 0$, such that no algorithm can estimate consistently clusters defined by (1.1) uniformly over covariance matrices fulfilling

$$(1.9) \qquad \Delta(C^*) \ge c\left(\sqrt{\frac{\log(p)}{nm^*}} \vee \frac{\log(p)}{n}\right).$$

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 (1.9) 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 5.3 to be near-minimax optimal. For instance, for balanced clusters, there exists a constant c' such that exact recovery is guaranteed when $\Delta(C^*) \ge c'(\sqrt{\frac{K \vee \log p}{m^* n}} + \frac{K \vee \log(p)}{n})$. 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 [16] and in sparse PCA [9]. It has been conjectured that, when K increases with n, there exists a gap between the statistical boundary, that is, the minimal cluster separation for which a statistical method achieves perfect clustering with high probability, and the polynomial boundary, that is, 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 [16] and [9] for more details.

However, if we consider directly the metric $\text{MCOD}(\Sigma)$, and its corresponding, larger, minimax threshold, we derive the COD algorithm, which is minimax optimal with respect to $\text{MCOD}(\Sigma)$ when $K \ge 3$. 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 6.2 and Theorem 6.5, respectively.

Table 1 gives a snapshot 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.

TABLE 1
Algorithm performance relative to minimax thresholds of each metric

Metric	Minimax threshold	PECOK	COD
$d_1 =: \Delta(C^*)$	$\sqrt{\frac{\log p}{mn}} + \frac{\log p}{n}$	Minimax optimal w.r.t. d_1 when $K = O(\log(p))$.	Minimax optimal w.r.t. d_1 when m is constant.
$d_2 =: \text{MCOD}(\Sigma)$	$\sqrt{\frac{\log p}{n}}$ when $K \ge 3$	Minimax optimal w.r.t. d_2 when $m > n/\log(p)$ and $K = O(\log p)$.	Minimax optimal w.r.t. d_2 when $K \ge 3$.

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 [12] 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 8.3.

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 2.2 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 $\mathrm{MCOD}(\Sigma)$-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 Semidefinite Programing (SDP). Theorem 5.3 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 [10], 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 $\mathrm{MCOD}(\Sigma)$ cluster separation when $K \geq 3$, 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. Spectral clustering is less computationally involved than PECOK, but the theoretical guaranties that we can offer for it are weaker.

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 $\mathrm{MCOD}(\Sigma)$, 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 analyzes spectral clustering for variable clustering, and compares it with PECOK.

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

Section 7 presents their application to the clustering of putative brain areas using a real fMRI data.

Section 8 contains a discussion of our results and overall recommendations regarding the usage of our methods. Given the space constraints, all proofs and simulation results are included in the Supplementary Material.

The implementation of PECOK can be found at http://github.com/martinroyer/pecok/ and that of COD at http://CRAN.R-project.org/package=cord.

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

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

Given a vector v and $q \geq 1$, $|v|_q$ stands for the ℓ_q norm. For a generic matrix M: $|M|_q$ denotes its the entrywise ℓ_q norm, $\|M\|_{\mathrm{op}}$ denotes its operator norm and $\|M\|_F$ refers to the Frobenius norm. We use $M_{:a}$, $M_{b:}$, to denote the ath column or, respectively, bth row of a generic matrix M. The bracket $\langle \cdot, \cdot \rangle$ 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 seminorm 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 $c_1, c_2, \ldots$ 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 $\|Y\|_{\psi_2} = \sup_{t \in \mathbf{R}^p : \|t\|_2 = 1} \inf\{s > 0 : \mathbb{E}[e^{(Z^t t)/s^2} \leq 2]\}$. Throughout the paper, we will assume that X follows a sub-Gaussian distribution. Specifically, we use the following.

ASSUMPTION 1 (Sub-Gaussian distributions). The exists $L > 0$ such that the random vector $\Sigma^{-1/2}X$ satisfies $\|\Sigma^{-1/2}X\|_{\psi_2} \leq L$.

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 $\max_{i=1,\dots,p} |X_i| \leq M$ 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 covariance 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 subpartition 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*

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

PROOF. If decomposition $\Sigma = ACA^t + \Gamma$ holds with A related to a partition G, then we have $\text{COD}(a,b) = 0$ 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. $\square$

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 $\text{MCOD}(\Sigma, G) := \min_{a \overset{G}{\not\sim} b} \text{COD}(a,b)$, 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 2.2. *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 given in Section B of the Supplementary Material [14]. 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 2.3. *Let G be a partition such that $m = \min_k |G_k| \geq 2$, 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 distinct.

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 $1 \leq m \leq p/2$ and $\eta > 0$, we define $\mathcal{M}(m, \eta)$ as the set of covariance matrices Σ fulfilling $\mathrm{MCOD}(\Sigma) > \eta |\Sigma|_\infty$ and whose associated partition G^* has groups of equal size $m^* \geq m$. Similarly, for $\tau > 0$, we define $\mathcal{D}(m, \tau)$ as the set of covariance matrices Σ fulfilling $\Delta(C^*) > \tau |\Gamma|_\infty$ and whose associated partition G^* has groups of equal size $m^* \geq m$. 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_1 such that, for any $1 \leq m \leq p/3$ and any η such that*

$$(3.1) \qquad 0 \leq \eta < \eta^* := c_1 \sqrt{\frac{\log(p)}{n}},$$

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

When $2 \leq m = p/2$, the same result holds but with the Condition (3.1) replaced by

$$(3.2) \qquad 0 \leq \eta < \eta^* := c_1 \left[\sqrt{\frac{\log(p)}{np}} \vee \frac{\log(p)}{n} \right].$$

We also have the following.

THEOREM 3.2. *There exist positive constants c_1–c_3 such that the following holds for any $2 \leq m \leq p/2$. For any τ such that*

$$(3.3) \qquad 0 \leq \tau < \tau_* := c_1 \left[\sqrt{\frac{\log(p)}{n(m-1)}} \vee \frac{\log(p)}{n} \right],$$

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

Conversely, there exists a procedure $\widehat{G}$ satisfying $\sup_{\Sigma \in \mathcal{D}(m,\tau)} \mathbb{P}_\Sigma[\widehat{G} \neq G^*] \leq c_3/p$ for any τ such that

$$\tau > \tau^* := c_2 \left[\sqrt{\frac{\log(p)}{n(m-1)}} \vee \frac{\log(p)}{n} \right].$$

Theorems 3.2 and 3.1 show that if either metric falls below the thresholds in (3.3) and (3.1) or (3.2), 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 Section C of the Supplementary Material [14]. We note that the $\Delta(C^*)$ 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, as soon as there are at least 3 groups. The proof of (3.1) provides an example where we have $K = 3$ clusters, that are very large, of size $m^* = p/3$ each, and where the MCOD(Σ) threshold does not decrease with m^*.

Theorem 3.2 also provides a matching upper bound for the minimax threshold. Unfortunately, the procedure achieving this bound has an exponential computational complexity (see Section C.3 in the Supplementary Material [14] for further details and Section 5 for a near-minimax optimal algorithm with polynomial computational complexity).

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

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

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

$$\widehat{s\text{COD}}(a, b) := \max_{c \neq a,b} |\widehat{\mathbf{cor}}(X_a - X_b, X_c)|$$

(4.1)
$$= \max_{c \neq a,b} \left| \frac{\widehat{\Sigma}_{ac} - \widehat{\Sigma}_{bc}}{\sqrt{(\widehat{\Sigma}_{aa} + \widehat{\Sigma}_{bb} - 2\widehat{\Sigma}_{ab})\widehat{\Sigma}_{cc}}} \right|.$$

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,

$$\widehat{G}_l = \{j \in S : \widehat{s\text{COD}}(a_l, j) \vee \widehat{s\text{COD}}(b_l, j) \leq \alpha\}.$$

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 [13]. 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 $O(p^2(p \vee n))$. 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 $S \neq \varnothing$

1. $l \leftarrow l + 1$
2. If $|S| = 1$ Then $\widehat{G}_l = S$
3. If $|S| > 1$ Then
    (a) $(a_l, b_l) = \mathrm{argmin}_{a,b \in S, a \neq b} \widehat{sCOD}(a, b)$
    (b) If $\widehat{sCOD}(a_l, b_l) > \alpha$ Then $\widehat{G}_l = \{a_l\}$
        Else $\widehat{G}_l = \{j \in S : \widehat{sCOD}(a_l, j) \wedge \widehat{sCOD}(b_l, j) \leq \alpha\}$
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 minimax optimal* $\mathrm{MCOD}(\Sigma)$ *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 a constant times the threshold (3.1) of Theorem 3.1 of the previous section.

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

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

and $\mathrm{MCOD}(\Sigma) > 3\alpha |\Sigma|_\infty$, *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 Section D.1 of the Supplementary Material [14].

We observe that while the COD algorithm succeeds to recover G^* at the minimax separation rate (3.1) when $K \geq 3$, it does not offer guaranties at the minimax separation rate (3.2) when $K = 2$. In this last case ($K = 2$), we observe that

$$\frac{1}{2}\Delta(C^*) \leq \mathrm{MCOD}(\Sigma) \leq \Delta(C^*),$$

so the metric $\mathrm{MCOD}(\Sigma)$ is equivalent to $\Delta(C^*)$ and we refer to the Section 5 for an optimal algorithm.

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 sub-Gaussian 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. Let us consider two independent subsamples of the original sample, $\mathcal{D}^1$ and $\mathcal{D}^1$, each of size $n/2$.

We denote by $\widehat{\mathcal{G}}^{(1)}$ a collection of partitions computed from $\mathcal{D}^1$, for instance via the COD algorithm with a varying threshold α. For any $a < b$, we use $\mathcal{D}^i$, $i = 1, 2$, to calculate, respectively,

$$\widehat{\Delta}_{ab}^{(i)} =: [\widehat{\mathrm{Cor}}^{(i)}(X_a - X_b, X_c)]_{c \neq a,b}, \quad i = 1, 2.$$

Since $\Delta_{ab} := [\mathrm{Cor}(X_a - X_b, X_c)]_{c \neq a,b}$ equals zero if and only if $a \overset{G}{\sim} b$, we want to select a partition G such that $\widehat{\Delta}_{ab}^{(2)} \mathbf{1}_{a \overset{G}{\not\sim} b}$ 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 reuse

sample $\mathcal{D}^1$ 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 the data-splitting criterion $\mathcal{H}$:

$$\widehat{G} \in \underset{G \in \widehat{\mathcal{G}}^{(1)}}{\operatorname{argmin}} \mathcal{H}(G) \quad \text{with} \quad \mathcal{H}(G) =: \sum_{a<b} [|\widehat{\Delta}_{ab}^{(2)} \mathbf{1}_{a \overset{G}{\not\approx} b} - \widehat{\Delta}_{ab}^{(1)}|_{\infty}^2].$$

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

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

In general, the sample correlation is not an unbiased estimator of the population level correlation. Still, (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 criterion $\mathcal{H}$, averaged over $\mathcal{D}^2$, and denoted by $\mathbb{E}^{(2)}[\mathcal{H}(G)]$. The proof is given in Section D.2 of the Supplementary Material [14].

PROPOSITION 4.2. *Assume that the distributional Assumption* 1 *and* (P1) *hold. Then there exists a constant $c_1 > 0$ such that, when* $\operatorname{MCOD}(\Sigma) > c_1 |\Sigma|_\infty L^2 \sqrt{\log(p)/n}$, *we have*

$$(4.2) \qquad \mathbb{E}^{(2)}[\mathcal{H}(G^*)] \leq \min_{G \in \widehat{\mathcal{G}}^{(1)}} \mathbb{E}^{(2)}[\mathcal{H}(G)],$$

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

Under the condition $\operatorname{MCOD}(\Sigma) > c_1 |\Sigma|_\infty L^2 \sqrt{\log(p)/n}$, Theorem 4.1 ensures that G^* belongs to $\widehat{\mathcal{G}}^{(1)}$ with high probability, whereas (4.2) suggests that the criterion is minimized at G^*.

If we consider a data-splitting algorithm based on $\widehat{\operatorname{COD}}(a, b)$ instead of $\widehat{s\operatorname{COD}}(a, b)$, then we can obtain a counterpart of Proposition 4.2 without requiring the additional assumption (P1). Still, we favor the procedure based on $\widehat{s\operatorname{COD}}(a, b)$ 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 $\operatorname{MCOD}(\Sigma)$ metric for $K \geq 3$, 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 [30], which is NP-hard [5]. Following Peng and Wei [34], 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

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

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

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

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 [34] that the collection of such matrices B is described by the collection $\mathcal{O}$ of orthogonal projectors fulfilling $\operatorname{tr}(B) = K$, $B\mathbf{1} = \mathbf{1}$ and $B_{ab} \geq 0$ for all a, b.

Theorem 2.2 in Peng and Wei [34] shows that solving the K-means problem is equivalent to finding the global maximum

$$(5.3) \qquad \bar{B} = \underset{B \in \mathcal{O}}{\operatorname{argmax}} \langle \widehat{\Sigma}, B \rangle,$$

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

The set of orthogonal projectors is not convex, so, following Peng and Wei [34], we consider a convex relaxation $\mathcal{C}$ of $\mathcal{O}$ obtained by relaxing the condition "B orthogonal projector," by "B positive semidefinite," leading to

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

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

$$(5.5) \qquad \widetilde{B} = \underset{B \in \mathcal{C}}{\operatorname{argmax}} \langle \widehat{\Sigma}, B \rangle.$$

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*

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

$$|G_1^*| = |G_2^*| = |G_3^*| = m^*.$$

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 Section A.1 in the Supplementary Material [14]. As a consequence, when Γ is not proportional 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 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

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

as long as $\Delta(C^*) > 0$. This suggests the following PEnalized COnvex 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.

> Step 1. Estimate Γ by $\widehat{\Gamma}$.
>
> Step 2. Estimate B^* by $\widehat{B} = \underset{B \in \mathcal{C}}{\operatorname{argmax}} (\langle \widehat{\Sigma}, B \rangle - \langle \widehat{\Gamma}, B \rangle)$.
>
> Step 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 [30] seeded with K distinct centroids, kmeans++ [4] or any approximate K-means as defined in (5.13) in Section 5.4, fulfill this property.

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 nontrivial task, and needs to be done with care. We explain our estimation below and analyze it in Proposition A.10 in Section A.5. 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

$$(5.6) \qquad V(a, b) := \max_{c, d \in [p] \setminus \{a, b\}} \frac{|(\widehat{\Sigma}_{ac} - \widehat{\Sigma}_{ad}) - (\widehat{\Sigma}_{bc} - \widehat{\Sigma}_{bd})|}{\sqrt{\widehat{\Sigma}_{cc} + \widehat{\Sigma}_{dd} - 2\widehat{\Sigma}_{cd}}},$$

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

$$b_1(a) := \underset{b \in [p] \setminus \{a\}}{\operatorname{argmin}} V(a, b) \quad \text{and} \quad b_2(a) := \underset{b \in [p] \setminus \{a, b_1(a)\}}{\operatorname{argmin}} V(a, b),$$

to be two elements "close" to a, that is, two indices $b_1 = b_1(a)$ and $b_2 = b_2(a)$ such that the empirical covariance difference $\widehat{\Sigma}_{b_i c} - \widehat{\Sigma}_{b_i d}$, $i = 1, 2$, is most similar to $\widehat{\Sigma}_{ac} - \widehat{\Sigma}_{ad}$, for all variables c and d not equal to a or b_i, $i = 1, 2$. It is expected that $b_1(a)$ and $b_2(a)$ 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

$$(5.7) \qquad \widehat{\Gamma}_{aa} = \widehat{\Sigma}_{aa} + \widehat{\Sigma}_{b_1(a)b_2(a)} - \widehat{\Sigma}_{ab_1(a)} - \widehat{\Sigma}_{ab_2(a)} \quad \text{for } a = 1, \ldots, p.$$

Intuitively, $\widehat{\Gamma}_{aa}$ should be close to $\Sigma_{aa} + \Sigma_{b_1(a)b_2(a)} - \Sigma_{ab_1(a)} - \Sigma_{ab_2(a)}$, which is equal to Γ_{aa} in the favorable event where both $b_1(a)$ and $b_2(a)$ belong to the same group as a.

In general, $b_1(a)$ and $b_2(a)$ cannot be guaranteed to belong to the same group as a. Nevertheless, these two surrogates $b_1(a)$ and $b_2(a)$ 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.10 in Section A.5 of the Supplementary Material [14]. A slightly simpler estimator of Γ was proposed in Appendix A of a previous version of this work [15], but a bound on $|\widehat{\Gamma}_{aa} - \Gamma_{aa}|$ for that estimator contains a factor proportional to $|\Sigma|_\infty$, which is not desirable, and can be avoided by (5.7). In the next subsection, we show that our proposed $\widehat{\Gamma}$ leads to perfect recovery of G^*, via PECOK, under minimal separation conditions.

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 [15], Section 4. Alternatively, we propose in Section G of Supplementary Material [14] selection via a simple data-splitting 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 5.3. *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 $|\widehat{\Gamma} - \Gamma|_V \leq \delta_{n,p}$ with probability $1 - c_1/p$. Then, under Assumption 1, and when $L^4 \log(p) \leq c_3 n$ and*

$$(5.8) \qquad \Delta(C^*) \geq c_L \left[\|\Gamma\|_\infty \left\{ \sqrt{\frac{\log p}{m^* n}} + \sqrt{\frac{p}{nm^{*2}}} + \frac{\log(p)}{n} + \frac{p}{nm^*} \right\} + \frac{\delta_{n,p}}{m^*} \right],$$

then $\widehat{B} = B^*$ and $\widehat{G} = G^*$, 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 (5.7), the same conclusion holds with probability higher than $1 - c_2/p$ when

$$(5.9) \qquad \Delta(C^*) \geq c_L \|\Gamma\|_\infty \left\{ \sqrt{\frac{\log p}{m^* n}} + \sqrt{\frac{p}{nm^{*2}}} + \frac{\log(p)}{n} + \frac{p}{nm^*} \right\}.$$

The proof is given in Section A.3 of the Supplementary Material [14].

REMARK 1. We left the term $\delta_{n,p}$ explicit in (5.8) 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 nonzero (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 (5.9) shows that the separation condition can be much decreased with the correction. In particular, for balanced clusters, that is when $m^* = \frac{p}{K}$, exact recovery is guaranteed when

$$(5.10) \qquad \Delta(C^*) \geq c_L \left[\sqrt{\frac{K \vee \log p}{m^* n}} + \frac{K \vee \log p}{n} \right],$$

for an appropriate constant $c_L > 0$. In view of Theorem 3.2, when $m^* \geq cp/\log(p)$ the rate is minimax optimal, since in this case $K = p/m^* = O(\log(p))$. When $m^* = o(p/\log(p))$, the number K of clusters grows faster than $\log(p)$, and 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 [16] for a discussion in the related stochastic block model.

Bounded variables X also follow a sub-Gaussian distribution. Nevertheless, the corresponding sub-Gaussian norm L may be large and Theorem 5.3 can sometimes be improved, as in Theorem 5.4 below, proved in Section A.3 of the Supplementary Material [14].

THEOREM 5.4. *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 $|\widehat{\Gamma} - \Gamma|_V \leq \delta_{n,p}$ with probability $1 - c_1/p$. Then, under Assumption 1-bis, and*

$$\Delta(C^*) \geq c_2 \left[M \|\Gamma\|_\infty^{1/2} \sqrt{\frac{p \log(p)}{nm^{*2}}} + M^2 \frac{p \log(p)}{nm^*} + \frac{\delta_{n,p}}{m^*} \right]. \tag{5.11}$$

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

When we choose $\widehat{\Gamma}$ as in (5.7), the term $\delta_{n,p}/m^*$ can be simplified as under Assumption 1; see Proposition A.10 in Section A.5 of the Supplementary Material [14]. For balanced clusters, $m^* = \frac{p}{K}$, Condition (5.11) can be simplified in

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

In comparison to (5.10), the condition does no longer depend on the sub-Gaussian norm 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 $p = m^* K$, $C^* = (\omega_{\text{in}} - \omega_{\text{out}}) I_K + \omega_{\text{out}} J$ and $\Gamma = (1 - \omega_{\text{in}}) I_K$. 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 (5.11) simplifies to

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

In the specific case $K = 2$, we recover (up to numerical multiplicative constants) the optimal rate proved in [10]. Our procedure and analysis provide a generalization of these results, as they are valid for general K and Theorem 5.4 also allows for unbalanced clusters.

5.4. *A comparison between PECOK and spectral clustering.* In this section, we discuss connections between the PECOK algorithm introduced above and spectral clustering, a method that has become popular in network clustering.

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 [28]) that a and b belong to the same cluster if and only if $U_{a:} = U_{b:}$ and if and only if $[UU^t]_{a:} = [UU^t]_{b:}$. 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 $\widehat{V}$.

Arguing as in Peng and Wei [34], we have the following.

LEMMA 5.5.   *SC algorithm is equivalent to the following algorithm*:

Step 1. *Find* $\overline{B} = \operatorname{argmax}\{\langle \widehat{\Sigma}, B \rangle : \operatorname{tr}(B) = K, I \succcurlyeq B \succcurlyeq 0\}$.
Step 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 $\widetilde{B}$ defined in (5.5) involves the calculation of

$$(5.12) \qquad \widetilde{B} = \operatorname*{argmax}_{B}\{\langle \widehat{\Sigma}, B \rangle : B\mathbf{1} = \mathbf{1}, B_{ab} \geq 0, \operatorname{tr}(B) = K, B \succcurlyeq 0\}.$$

Since the matrices B involved in (5.12) are doubly stochastic, their eigenvalues are smaller than 1, and hence (5.12) is equivalent to $\widetilde{B} = \operatorname{argmax}_B\{\langle \widehat{\Sigma}, B \rangle : B\mathbf{1} = \mathbf{1}, B_{ab} \geq 0, \operatorname{tr}(B) = K, I \succcurlyeq B \succcurlyeq 0\}$. Note then that $\overline{B}$ can be viewed as a less constrained version of $\widetilde{B}$, in which $\mathcal{C}$ is replaced by $\overline{\mathcal{C}} = \{B : \operatorname{tr}(B) = K, I \succcurlyeq B \succcurlyeq 0\}$, where we have dropped the $p(p+1)/2$ constraints given by $B\mathbf{1} = \mathbf{1}$, and $B_{ab} \geq 0$. The proof of Lemma 5.5 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^*. Below, we adapt the arguments of [28] in order to provide some guarantees for a corrected version of spectral clustering.

In view of this connection between spectral clustering and unpenalized PECOK and of the fact that the population justification of spectral clustering deals with the spectral decomposition of $\Sigma - \Gamma$, we 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 (5.13).

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

$$(5.13) \qquad\qquad \operatorname{crit}(\widehat{U}^t, \widehat{G}) \leq \eta \min_{G} \operatorname{crit}(\widehat{U}^t, G),$$

with $\operatorname{crit}(\cdot, \cdot)$ the K-means criterion (5.1). Although solving K-means is NP-Hard [5], there exist polynomial time approximate K-means algorithms; see Kumar *et al.* [26]. As a consequence of the above discussion, the first step of CSC can be interpreted as a relaxation of the program associated to the PECOK estimator $\widehat{B}$.

To simplify the presentation of the results for CSC procedure, we assume in the following that all the groups have the same size $|G_1^*| = \cdots = |G_K^*| = m^* = p/K$. We emphasize that this information is not required by either PECOK or CSC, or in the proof of Proposition 5.6 below. We only use it here for simplicity. We denote by $\mathcal{S}_K$ the set of permutations on $\{1, \ldots, K\}$ and we denote by

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

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 $\overline{L}(\widehat{G}, G^*) \leq \rho$, for an appropriate quantity $\rho < 1$. We begin with a general theorem pertaining to partial partition recovery by CSC, under a "signal-to-noise ratio" involving the smallest eigenvalue $\lambda_K(C^*)$ of C^*.

PROPOSITION 5.6. *Let* $\mathrm{Re}(\Sigma) = \mathrm{tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$ *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*

$$(5.14) \qquad \lambda_K(C^*) \geq \frac{c_{\eta,L}\sqrt{K}\|\Sigma\|_{\mathrm{op}}}{m^*\sqrt{\rho}}\sqrt{\frac{\mathrm{Re}(\Sigma) \vee \log(p)}{n}},$$

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

The proof extends the arguments of [28], 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 $\|\widetilde{\Sigma} - AC^*A^t\|_{\mathrm{op}}$. We then employ the results of [24] to show that this operator norm can be controlled, with high probability, which leads to the conclusion of the theorem.

As n goes to infinity, the right-hand side of Condition (5.14) 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.

We observe that $\Delta(C^*) \geq 2\lambda_K(C^*)$, so we can compare the lower bound (5.14) on $\lambda_K(C^*)$ to the lower-bound (5.10) on $\Delta(C^*)$. 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 5.7 (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*

$$(5.15) \qquad \rho \geq c_{\eta,L}\left[\frac{K^2}{n} + \frac{K\log(p)}{n}\right] \quad and \quad \tau \geq c'_{\eta,L}\left[\frac{K^2}{\rho n} \vee \frac{K}{\sqrt{\rho n m^*}}\right],$$

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

Recall that Theorem 5.3 above states that when $\widehat{G}$ is obtained via the PECOK algorithm, and if $\tau \gtrsim \sqrt{\frac{K \vee \log p}{m^* n}} + \frac{\log(p) \vee K}{n}$, then $\overline{L}(\widehat{G}, G^*) = 0$, or equivalently, $\widehat{G} = G^*$, with high probability. We can therefore provide the following comparison (we refer to Section G of the Supplementary Material [14] for a numerical comparison):

- When $\tau \gtrsim \sqrt{\frac{K \vee \log p}{m^* n}} + \frac{\log(p) \vee K}{n}$, and under the additional condition that $n \gtrsim (K \vee \log(p))^2/K$, CSC algorithm satisfies $\overline{L}(\widehat{G}, G^*) \leq K^2/(K \vee \log(p))$. So, for $K = o(\log(p))$ and for a large enough sample size $n \gtrsim (\log(p))^2/K$, the fraction of misclassified variables by CSC is vanishing as $O(K/\log(p))$ for $\tau \gtrsim \sqrt{\frac{\log p}{m^* n}} + \frac{\log(p)}{n}$. This guaranty is slightly weaker than for PECOK which ensures exact recovery in this setting. This discrepancy may be an artifact of the proof technique. Very recent works [1, 31] (released during the reviewing process of this paper) present reconstruction error bounds tighter than those of [28], for (variants of) spectral clustering, when applied to two parameter SBMs, for network data, not the type of data analyzed in this work.

- When we move away from the case $C^* = \tau I_K$, the guaranties for CSC can degenerate. 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 discussed on page 126. 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 (5.14) requires a cluster separation at least

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

which is independent of m^*, unlike the condition $\tau \gtrsim \sqrt{\frac{K \vee \log p}{m^* n}} + \frac{\log(p) \vee K}{n}$ for PECOK. This unpleasant feature is induced by the inflation of $\|\widetilde{\Sigma} - AC^*A^t\|_{\mathrm{op}}$ with α. Again, this weakness in the guarantees may be an artifact of the proof, which relies on the Davis–Kahan inequality for controlling the alignment between the sample eigenvectors associated with the K largest eigenvalues and their population counterpart.

All the results of this section are proved in Section E of the Supplementary Material [14].

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 nondiagonal, where the nondiagonal 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 nondiagonal, we can decompose Σ using

different partitions. For instance,

$$\Sigma = \underbrace{\begin{bmatrix} 2r & 0 & 0 \\ 0 & r & r \\ 0 & r & r \end{bmatrix}}_{=A_1 C_1 A_1^t} + \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & r & -r \\ 0 & -r & r \end{bmatrix}}_{=\Gamma_1} + I_p$$

$$= \underbrace{\begin{bmatrix} r & r & 0 \\ r & r & 0 \\ 0 & 0 & 2r \end{bmatrix}}_{=A_2 C_2 A_2^t} + \underbrace{\begin{bmatrix} r & -r & 0 \\ -r & r & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{=\Gamma_2} + I_p.$$

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

For this, let us define $\mathcal{P}_j(\Sigma, K)$, $j = \{1, 2\}$ 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 $\min_k |G_K| \geq j$, 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

$$\rho_1(\Sigma, K) = \max\{\mathrm{MCOD}(\Sigma, G)/|R|_\infty :$$
$$(A, C, D, R) \in \mathcal{P}_1(\Sigma, K) \text{ and } G \text{ associated to } A\},$$
$$\rho_2(\Sigma, K) = \max\{\Delta(C)/|R|_\infty : (A, C, D, R) \in \mathcal{P}_2(\Sigma, K)\}.$$

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* $\mathrm{MCOD}(\Sigma, G) > 2|R|_\infty$. *We denote by $G_1[K]$ this partition.*

(ii) *Assume that $\rho_2(\Sigma, K) > 8$. Then there exists a unique partition G with $\min_k |G_k| \geq 2$, 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 Section B.3 of the Supplementary Material [14]. The proof of Proposition 6.1. is given in Section B.2 of [14].

The conclusion of Proposition 6.1 does essentially revert to that of Proposition 2.2 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 $m^* \geq 2$. 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 $\mathrm{MCOD}(\Sigma, G)$ (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 $\min_k |G_2[K]_k| \geq 2$. 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 6.2. *Under the distributional Assumption* 1, *there exist numerical constants* $c_1, c_2 > 0$ *such that the following holds for all* $\alpha \geq c_1 L^2 \sqrt{\frac{\log p}{n}}$. *If, for some partition G and decomposition* $\Sigma = ACA^t + R + D$, *we have*

$$(6.1) \qquad |R|_\infty \leq \frac{\lambda_{\min}(\Sigma)}{2\sqrt{2}}\alpha \quad \text{and} \quad \mathrm{MCOD}(\Sigma, G) > 3\alpha|\Sigma|_\infty,$$

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

The proof is given in Section D.1 of the Supplementary Material [14]. If G satisfies the assumptions of Theorem 6.2, 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 $\mathrm{MCOD}(\Sigma, G^*) \geq 3\alpha|\Sigma|_\infty$, then COD selects G^* with high probability. If $\mathrm{MCOD}(\Sigma, G^*)$ 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 $\mathrm{MCOD}(\Sigma, G)$ 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 $\Delta(C)/|R|_\infty$.

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

PROPOSITION 6.3. *If,* $\Delta(C^*) > \frac{7|D|_V + 2\|R\|_{\mathrm{op}}}{m} + 3|R|_\infty$, *then* $B^* = \mathrm{argmin}_{B \in \mathcal{C}}\langle \Sigma, B \rangle$.

COROLLARY 6.4. *If* $\Delta(C^*) > 3|R|_\infty + \frac{2\|R\|_{\mathrm{op}}}{m}$, *then* $B^* = \mathrm{argmin}_{B \in \mathcal{C}}\langle \Sigma - D, B \rangle$.

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 Section 6.1. In comparison to the necessary conditions discussed in Section 6.1, there is an additional $\|R\|_{\mathrm{op}}/m$ term. The proofs are given in Section A.2 in the Supplementary Material [14].

We now examine the behavior of PECOK when we specify the estimator $\widehat{\Gamma}$ to be as in (5.7). 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

$$(6.2) \qquad \Gamma_{aa} \geq 3 \max_{c:c\neq a} |\Gamma_{ac}| \quad \left(\text{or equivalently} \quad D_{aa} \geq 3 \max_{c:c\neq a} |R_{ac}| \right).$$

The next theorem states that PECOK estimator $\widehat{B}$ recovers the groups under similar conditions to that of Theorem 5.3 if R is small enough. The proof is given in Section A.3 of the Supplementary Material [14].

THEOREM 6.5. *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 $L^4 \log(p) \leq c_1 n$ and*

$$|R|_\infty + \frac{\sqrt{|R|_\infty |D|_\infty} + \|R\|_{\mathrm{op}}}{m}$$

$$(6.3) \qquad \leq c_L \|\Gamma\|_{\mathrm{op}} \left\{ \sqrt{\frac{\log p}{mn}} + \sqrt{\frac{p}{nm^2}} + \frac{\log(p)}{n} + \frac{p}{nm} \right\}$$

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

$$(6.4) \qquad \Delta(C^*) \geq c_L' \left[\|\Gamma\|_{op} \left\{ \sqrt{\frac{\log p}{mn}} + \sqrt{\frac{p}{nm^2}} + \frac{\log(p)}{n} + \frac{p}{nm} \right\} \right].$$

So, as a long as $|R|_\infty$ and $\|R\|_{\mathrm{op}}$ are small enough so that (6.3) are satisfied, PECOK algorithm will correctly identify the target partition G^* at the Δ-(near) optimal minimax level (6.4). A counterpart of Theorem 6.5 for Assumption 1-bis is provided in Section A.3 of the Supplementary Material [14].

7. Data analysis. Using functional MRI data, [37] 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 data set to illustrate the clusters recovered by different methods. The data set was originally published in [39] 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 [39], and we follow [37] to subsample the whole brain data using $p = 264$ putative areas; see Section A.3 of the Supplementary Material [14] 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 data-splitting 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 loss $\mathcal{H}(G)$ given in Section 4.3. Among our methods (COD, COD-CC and PECOK), COD yields the smallest loss when $K = 142$. We thus first focus on illustrating the COD clusters here. Table 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 1(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 [37] 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 meth-

TABLE 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

ods. 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:

$$(7.1) \qquad \left\| \hat{S}_2 - \Upsilon(\hat{S}_1, \hat{G}) \right\|_F,$$

where block averaging operator $\Upsilon(R, G)$ produces a G-block structured matrix based on $\hat{G}$. For any $a \in G_k$ and $b \in G_{k'}$, the output matrix entry $[\Upsilon(R, G)]_{ab}$ is given by

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

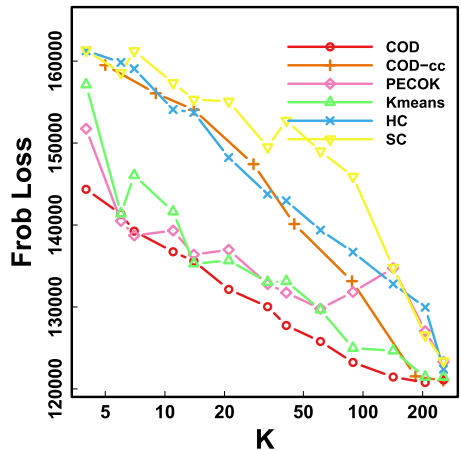(a) A brain cluster by COD   (b) Prediction loss
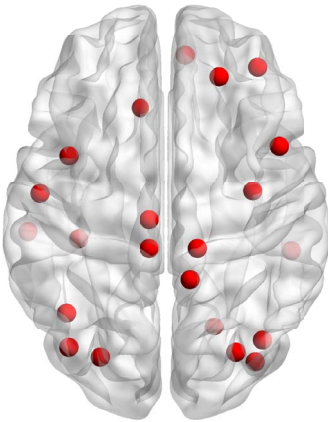
FIG. 1. (a) *Plot of the coordinates of the largest* COD *cluster overplayed over a standard brain template. The coordinates are shown as red balls.* (b) *Comparison of* COD, COD-CC, PECOK, *K-means, HC and SC using the Frobenius prediction loss criterion* (7.1) *where the groups are estimated by these methods, respectively.*

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 (7.1) while smoothing over different clusters will increase the loss.

Figure 1(b) compares the prediction loss values under different group sizes for each method. This shows that our data-splitting 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 data-splitting 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 1(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.

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

8.1. *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 [2, 16, 21, 27–29, 33]. 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 [28] or the SDP such as the ones in [3] 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 [3]) 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. [32] 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 nontrivial cross-product terms. Instead of relying on dual certificates arguments as in other work such as [35], we directly investigate the primal problem and combine different duality-norm bounds. The crucial step is the Lemma A.3 in the Supplementary Material [14] 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.

8.2. *Extension to other models.* The general strategy of correcting a convex relaxation of K-means can be applied to other models. In [38], one of the authors has adapted the PECOK algorithm to the clustering problem of mixture of sub-Gaussian distributions. In particular, in the high-dimensional setting where the correction plays a key role, [38] obtains sharper separation conditions dependencies than in state-of-the-art clustering procedures [32]. Extensions to model-based overlapping clustering are beyond the scope of this paper, but we refer to [11] for recent results.

8.3. *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.

SUPPLEMENTARY MATERIAL

Supplement to "Model assisted variable clustering: Minimax-optimal recovery and algorithms" (DOI: 10.1214/18-AOS1794SUPP; .pdf). This supplement contains proofs of the theoretical results, the simulation results and additional supporting information regarding the data analysis.

REFERENCES

[1] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. ArXiv Preprint arXiv:1709.09565.

[2] ABBE, E. and SANDON, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In 2015 *IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS* 2015 670–688. IEEE Computer Soc., Los Alamitos, CA. MR3473334

[3] AMINI, A. A. and LEVINA, E. (2018). On semidefinite relaxations for the block model. *Ann. Statist.* **46** 149–179. MR3766949 https://doi.org/10.1214/17-AOS1545

[4] ARTHUR, D. and VASSILVITSKII, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035. ACM, New York. MR2485254

[5] AWASTHI, P., CHARIKAR, M., KRISHNASWAMY, R. and SINOP, A. K. (2015). The hardness of approximations of Euclidean k-means. In 31*st International Symposium on Computational Geometry. LIPIcs. Leibniz Int. Proc. Inform.* **34** 754–767. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3392820

[6] BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res*. **9** 485–516. MR2417243

[7] BELLEC, P., PERLBARG, V., JBABDI, S., PÉLÉGRINI-ISSAC, M., ANTON, J.-L., DOYON, J. and BENALI, H. (2006). Identification of large-scale networks in the brain using fMRI. *NeuroImage* **29** 1231–1243.

[8] BERNARDES, J. S., VIEIRA, F. R., COSTA, L. M. and ZAVERUCHA, G. (2015). Evaluation and improvements of clustering algorithms for detecting remote homologous protein families. *BMC Bioinform*. **16** 1–14.

[9] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Proceedings of the* 26*th Annual Conference on Learning Theory* (S. Shalev-Shwartz and I. Steinwart, eds.). *Proceedings of Machine Learning Research* **30** 1046–1066. PMLR, Princeton, NJ.

[10] BERTHET, Q., RIGOLLET, P. and SRIVASTAVA, P. (2018). Exact recovery in the Ising blockmodel. *Ann. Statist*. To appear. arXiv:1612.03880.

[11] BING, M., BUNEA, F., NING, Y. and WEGKAMP, M. (2018). Adaptive estimation in structured factor models with applications to overlapping clustering. ArXiv E-prints.

[12] BOUVEYRON, C. and BRUNET-SAUMARD, C. (2014). Model-based clustering of high-dimensional data: A review. *Comput. Statist. Data Anal*. **71** 52–78. MR3131954 https://doi.org/10.1016/j.csda.2012.12.008

[13] BUNEA, F., GIRAUD, C. and LUO, X. (2015). Minimax optimal variable clustering in G-models via cord. ArXiv Preprint arXiv:1508.01939.

[14] BUNEA, F., GIRAUD, C., LUO, X., ROYER, M. and VERZELEN, N. (2020). Supplement to "Model assisted variable clustering: Minimax-optimal recovery and algorithms." https://doi.org/10.1214/18-AOS1794SUPP.

[15] BUNEA, F., GIRAUD, C., ROYER, M. and VERZELEN, N. (2016). PECOK: A convex optimization approach to variable clustering. ArXiv Preprint arXiv:1606.05100.

[16] CHEN, Y. and XU, J. (2016). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res*. **17** Paper No. 27, 57. MR3491121

[17] CHONG, M., BHUSHAN, C., JOSHI, A. A., CHOI, S., HALDAR, J. P., SHATTUCK, D. W., SPRENG, R. N. and LEAHY, R. M. (2017). Individual parcellation of resting fMRI with a group functional connectivity prior. *NeuroImage* **156** 87–100. https://doi.org/10.1016/j.neuroimage.2017.04.054

[18] CRADDOCK, R. C., JAMES, G. A., HOLTZHEIMER, P. E., HU, X. P. and MAYBERG, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp*. **33** 1914–1928.

[19] FREI, N., GARCIA, A. V., BIGEARD, J., ZAAG, R., BUESO, E., GARMIER, M., PATEYRON, S., DE TAUZIA-MOREAU, M. L., BRUNAUD, V. et al. (2014). Functional analysis of Arabidopsisimmune-related MAPKs uncovers a role for MPK3 as negative regulator of inducible defences. *Genome Biol*. **15** 1–22.

[20] GLASSER, M. F., COALSON, T. S., ROBINSON, E. C., HACKER, C. D., HARWELL, J., YACOUB, E., UGURBIL, K., ANDERSSON, J., BECKMANN, C. F. et al. (2016). A multi-modal parcelation of human cerebral cortex. *Nature* **536** 171–178.

[21] GUÉDON, O. and VERSHYNIN, R. (2016). Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Related Fields* **165** 1025–1049. MR3520025 https://doi.org/10.1007/s00440-015-0659-z

[22] JAMES, G. A., HAZAROGLU, O. and BUSH, K. A. (2016). A human brain atlas derived via n-cut parcellation of resting-state and task-based fMRI data. *Magn. Reson. Imaging* **34** 209–218.

[23] JIANG, D., TANG, C. and ZHANG, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Trans. Knowl. Data Eng*. **16** 1370–1386.

[24] KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

[25] KONG, R., LI, J., SUN, N., SABUNCU, M., LIU, H., SCHAEFER, A., ZUO, X.-N., HOLMES, A., EICKHOFF, S. et al. (2018). Spatial topography of individual-specific cortical networks predicts human cognition, personality and emotion. https://www.biorxiv.org/content/early/2018/01/31/213041.

[26] KUMAR, A., SABHARWAL, Y. and SEN, S. (2004). A simple linear time $(1 + \epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Foundations of Computer Science*, 2004. *Proceedings*. 45*th Annual IEEE Symposium on* 454–462.

[27] LE, C. M., LEVINA, E. and VERSHYNIN, R. (2016). Optimization via low-rank approximation for community detection in networks. *Ann. Statist.* **44** 373–400. MR3449772 https://doi.org/10.1214/15-AOS1360

[28] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

[29] LEI, J. and ZHU, L. (2014). A generic sample splitting approach for refined community recovery in stochastic block models. ArXiv Preprint arXiv:1411.1469.

[30] LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. MR0651807 https://doi.org/10.1109/TIT.1982.1056489

[31] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. ArXiv Preprint arXiv:1612.02099.

[32] MIXON, D. G., VILLAR, S. and WARD, R. (2017). Clustering subgaussian mixtures by semidefinite programming. *Inf. Inference* **6** 389–415. MR3764529 https://doi.org/10.1093/imaiai/iax001

[33] MOSSEL, E., NEEMAN, J. and SLY, A. (2014). Consistency thresholds for binary symmetric block models. ArXiv Preprint arXiv:1407.1591.

[34] PENG, J. and WEI, Y. (2007). Approximating K-means-type clustering via semidefinite programming. *SIAM J. Optim.* **18** 186–205. MR2299680 https://doi.org/10.1137/050641983

[35] PERRY, A. and WEIN, A. S. (2015). A semidefinite program for unbalanced multisection in the stochastic block model. ArXiv E-prints arXiv:1507.05605.

[36] POLDRACK, R. A. (2007). Region of interest analysis for fMRI. *Soc. Cogn. Affect. Neurosci.* **2** 67–70. https://doi.org/10.1093/scan/nsm006

[37] POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

[38] ROYER, M. (2017). Adaptive clustering through semidefinite programming. In *Advances in Neural Information Processing Systems* (*NIPS*).

[39] XUE, G., ARON, A. R. and POLDRACK, R. A. (2008). Common neural substrates for inhibition of spoken and manual responses. *Cereb. Cortex* **18** 1923–1932.

[40] YEO, B. T., KRIENEN, F. M., SEPULCRE, J., SABUNCU, M. R., LASHKARI, D., HOLLINSHEAD, M., ROFFMAN, J. L., SMOLLER, J. W., ZÖLLEI, L. et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106** 1125–1165.

[41] ZAAG, R., TAMBY, J., GUICHARD, C., TARIQ, Z., RIGAILL, G., DELANNOY, E., RENOU, J., BALZERGUE, S., MARY-HUARD, T. et al. (2015). GEM2Net: From gene expression modeling to -omics networks, a new CATdb module to investigate *Arabidopsis thaliana* genes involved in stress response. *Nucleic Acids Res.* **43** 1010–1017.