The Annals of Statistics
2020, Vol. 48, No. 4, 2055–2081
https://doi.org/10.1214/19-AOS1877

ADAPTIVE ESTIMATION IN STRUCTURED FACTOR MODELS WITH APPLICATIONS TO OVERLAPPING CLUSTERING

BY XIN BING[1,*], FLORENTINA BUNEA[1,**], YANG NING[1,†] AND MARTEN WEGKAMP[2]

[1]*Department of Statistics and Data Science, Cornell University,* [*]*xb43@cornell.edu;* [**]*fb238@cornell.edu;* [†]*yn265@cornell.edu*

[2]*Department of Mathematics & Department of Statistics and Data Science, Cornell University, mhw73@cornell.edu*

This work introduces a novel estimation method, called LOVE, of the entries and structure of a loading matrix A in a latent factor model $X = AZ + E$, for an observable random vector $X \in \mathbb{R}^p$, with correlated unobservable factors $Z \in \mathbb{R}^K$, with K unknown, and uncorrelated noise E. Each row of A is scaled, and allowed to be sparse. In order to identify the loading matrix A, we require the existence of pure variables, which are components of X that are associated, via A, with one and only one latent factor. Despite the fact that the number of factors K, the number of the pure variables and their location are all unknown, we only require a mild condition on the covariance matrix of Z, and a minimum of only two pure variables per latent factor to show that A is uniquely defined, up to signed permutations. Our proofs for model identifiability are constructive, and lead to our novel estimation method of the number of factors and of the set of pure variables, from a sample of size n of observations on X. This is the first step of our LOVE algorithm, which is optimization-free, and has low computational complexity of order p^2. The second step of LOVE is an easily implementable linear program that estimates A. We prove that the resulting estimator is near minimax rate optimal for A, with respect to the $\|\ \|_{\infty,q}$ loss, for $q \geq 1$, up to logarithmic factors in p, and that it can be minimax-rate optimal in many cases of interest.

The model structure is motivated by the problem of overlapping variable clustering, ubiquitous in data science. We define the population level clusters as groups of those components of X that are associated, via the matrix A, with the same unobservable latent factor, and multifactor association is allowed. Clusters are respectively anchored by the pure variables, and form overlapping subgroups of the p-dimensional random vector X. The *L*atent model approach to *OVE*rlapping clustering is reflected in the name of our algorithm, LOVE.

The third step of LOVE estimates the clusters from the support of the columns of the estimated A. We guarantee cluster recovery with zero false positive proportion, and with false negative proportion control. The practical relevance of LOVE is illustrated through the analysis of a RNA-seq data set, devoted to determining the functional annotation of genes with unknown function.

1. Introduction. In this work, we consider the problem of estimating the $p \times K$, possibly sparse, loading matrix A that parametrizes the factorization of a zero-mean observable random vector, $X \in \mathbb{R}^p$ as

$$X = AZ + E \tag{1.1}$$

from n i.i.d. realizations of X. The zero mean random vector $Z \in \mathbb{R}^K$ is unobservable, and can be viewed as a latent factor vector. $E \in \mathbb{R}^p$ is a zero-mean, unobservable random noise

Received March 2018; revised November 2018.
MSC2020 subject classifications. Primary 62H25, 62H30.
Key words and phrases. Overlapping clustering, latent model, identification, high-dimensional estimation, minimax estimation, pure variables, group recovery, support recovery, sparse loading matrix, matrix factorization, adaptive estimation.

vector, with uncorrelated entries. We assume E and Z are independent. The number of factors K is not known, and both p and K are allowed to grow, and be larger than n. Factor models have been used as dimension reduction devices in virtually any scientific discipline for nearly a century, and generated an enormous amount of literature. We refer to the classical monographs of Bollen (1989) and Anderson (2003) for earlier work, and to Izenman (2008) for a more recent survey and applications.

In this work, we revisit some of the open problems in factor model definition and estimation, and also consider one of their much less explored applications, to overlapping clustering. For the latter, we deem two components X_i and X_j of X similar if they have nonzero association, via the matrix A, with the same latent factor Z_a. Similar variables are placed in the same cluster, G_a:

$$G_a := \{j \in \{1, \dots, p\} : A_{ja} \neq 0\} \quad \text{for each } a \in \{1, \dots, K\}. \tag{1.2}$$

Since each X_j can be associated with multiple latent factors, the clusters will overlap. The problem of overlapping clustering is of wide-spread interest in virtually any scientific area, for instance, in neuroscience (Craddock et al. (2012, 2013)) and genetics (Jiang, Tang and Zhang (2004), Wiwie, Baumbach and Röttger (2015)), to give a very limited number of examples. The solutions are typically algorithmic in nature, and their quality is assessed against a ground scientific truth or via extensive simulation studies, for instance, Bezdek (1981), Krishnapuram et al. (2001), among many others. These problems have not received a systematic analysis in the statistical literature and, in particular, the problem of estimating overlapping clusters of variables, with theoretical guarantees, remains largely unexplored.

In this work, we propose model-based clustering via A. However, A cannot be uniquely defined in (1.1), without further restrictions, a phenomenon well understood over six decades ago. Most notably, Anderson and Rubin (1956) provided an in-depth analysis of this problem, and proved that in the absence of conditions on A and $C := \mathrm{Cov}(Z)$, A is not identifiable in model (1.1). We revisit some of these conditions here, with a view toward our application to overlapping clustering. We defer a detailed literature review of related identifiability conditions for model (1.1) to Section 4.4.

Using overlapping clustering as motivation, we formalize our first modeling assumption on A. We consider models (1.1) in which each row of A is scaled, to avoid scale ambiguities. Specifically, we assume that:

(i) $\sum_{a=1}^{K} |A_{ja}| \leq 1$.

The inequality in (i) allows for $\sum_a |A_{ja}| = 0$, which renders more flexibility to model (1.1), relative to the more commonly used equality conditions. If $\sum_a |A_{ja}| = 0$, then $X_j = E_j$, and X_j is not associated with any of the latent factors, via this model. The interpretation to clustering is that the corresponding $X_j = E_j$ does not belong to any cluster given by this model, which is a desired feature in many practical applications, including the one presented in this paper in Section 6. Furthermore, in order to use the model for clustering, we need to avoid the trivial situation in which each component X_j is associated with all latent factors. From this perspective, we allow the rows $A_{j\cdot} := (A_{j1}, \dots, A_{jK})$ to be sparse, for $j \in \{1, \dots, p\}$, but this property is not required for the identifiability of A.

Condition (i) alone cannot ensure that A in model (1.1) is uniquely defined, as one can still construct an invertible matrix Q such that $AZ = AQQ^{-1}Z$, with both A and AQ satisfying (i). Moreover, when A is sparse, A and AQ may not have the same sparsity pattern, creating ambiguity in the cluster definition. We introduce below two additional requirements that allow us to show, in Section 2 below, that A is identifiable.

We call (ii) given below *the pure variable assumption*. Informally, it postulates the existence of at least two *pure variables* X_j, which are components of X associated with one and only one latent factor. In Section 2, we provide examples that show that if pure variables do no exist, A in (1.1) is not uniquely defined.

(ii) For every $a \in \{1, \dots, K\}$, there exist at least two indices $j \in \{1, \dots, p\}$ such that $|A_{ja}| = 1$ and $A_{jb} = 0$ for all $b \neq a$.

In Remark 2 of Section 4 we discuss relaxations of this condition that allow each pure variable have a different scaling, possibly different than 1. We note that in the very particular case of known $\Gamma := \mathrm{Cov}(E)$, only *one* pure variable per group is required for identifiability, which follows from the proof of Theorem 2 in Section A.1. The pure variable assumption has an immediate practical implication to variable clustering. Since clusters G_a given by (1.2) are defined relative to the unobservable factor Z_a, a pure variable X_j is an observable proxy of Z_a, and that helps explain the otherwise unclear nature of G_a.

For future reference, we let I denote the index set corresponding to pure variables. In psychology, these variables are called factorially simple items (McDonald (1999)). A similar condition can be traced back to the econometrics literature, and an early reference is Koopmans and Reiersøl (1950), further discussed in Anderson and Rubin (1956), who called it "zero elements in *specified* positions." These works prove that (ii) corresponding to a *known* set I is a sufficient condition for identifying A, for latent factors with arbitrary correlations. However, full generality on the positive definite covariance matrix C of the latent factors comes at the steep price of knowing I a priori, which is often unrealistic in practice. Appropriate conditions on C that guarantee identifiability of I in (ii), in the general case when I is not known and, moreover, K is unknown, and have not been investigated for the general model (1.1), to the best of our knowledge. To this end, we introduce the following condition on the covariance matrix C:

(iii) $\Delta(C) := \min_{a \neq b}(C_{aa} \wedge C_{bb} - |C_{ab}|) > 0$ and C positive definite,

where $a \wedge b := \min(a, b)$. If (iii) holds, then $\mathrm{Cov}(Z_a \pm Z_b) = \mathrm{Var}(Z_a) + \mathrm{Var}(Z_b) \pm 2 \cdot \mathrm{Cov}(Z_a, Z_b) \geq C_{aa} + C_{bb} - 2|C_{ab}| > 0$, which implies that the latent factors are different, up to signs, that is, $|Z_a| \neq |Z_b|$ a.s. for any $a \neq b$.

Condition (iii) holds trivially under the much stronger assumption that the latent factors are independent, or have a slight departure from independence, corresponding to diagonal dominance in C. These types of assumptions are commonly made in latent factor models, but may often be unrealistic; see, for instance, Anderson (2003), Anderson and Rubin (1956), Bollen (1989), Everitt (1984), Izenman (2008) and our discussion in Section 4.4. Condition (iii) therefore relaxes the independent factor assumption, and we comment further on it below.

Condition (iii) is a companion of our conditions (i) and (ii). When the last two are being made, condition (iii) admits relaxations, which have been established only in special set-ups.

Under the pure variable assumption (ii), if I *is known* in advance, the arguments employed in the proof of our Theorem 2 of Section 2 show that (iii) is not required, and the assumption that C is a positive definite covariance matrix suffices. This is consistent with the classical literature on general latent models; see, for instance, Anderson and Rubin (1956).

Identifiability results corresponding to the realistic situation when I is not known are scarce, and correspond to particular instances of the model we consider in this work. In the limit case of our model, when all p variables are pure variables, which corresponds to nonoverlapping clustering, Bunea et al. (2018) showed that, once again, C being positive definite suffices for identifiability.

The problem of identifying A under (ii), with I unknown, has been revived more recently, in the particular case of modeling random vectors X with only nonnegative values, when A and Z also have only nonnegative entries. This set-up corresponds to the area known as nonnegative matrix factorization (NMF), in which one studies positive matrix factorizations of the type $\boldsymbol{X} = A\boldsymbol{Z} + \boldsymbol{E}$, where the observed data $\boldsymbol{X}$ is a $p \times n$ matrix, $\boldsymbol{Z}$ is the $K \times n$ unobservable matrix of the latent vectors and $\boldsymbol{E}$ is the $p \times n$ noise matrix. In this context, when $\boldsymbol{E} = 0$, and conditioning on $\boldsymbol{Z}$, Donoho and Stodden (2004) was among the first works to propose a

condition similar to (ii), with I unknown, coupled with appropriate conditions on $\boldsymbol{Z}$, leading to an NMF decomposition with unique factors. Moreover, the unique determination of I under (ii), for $\boldsymbol{E} \neq 0$, but with very small componentwise variances, was solved in Bittorf et al. (2012), for known K, and for scaled NMF models, in which the columns of $\boldsymbol{X}$, $\boldsymbol{Z}$ and A sum up to 1. These results were proved under the assumption that no row of a scaled version of $\boldsymbol{Z}$ is a convex combination of the other rows. Conditioning on Z, this requirement is weaker than our condition (iii), should we impose it on $n^{-1}\boldsymbol{Z}\boldsymbol{Z}^T$, but it is not readily generalizable outside the NMF framework.

In light of this discussion, our condition (iii) on C is a key ingredient in the identification of I, in the context of the more general model (1.1), when E is not negligible, and K is not known. The details are given in Section 2 below. If all the latent variables have the same variance, then condition (iii) becomes the very mild requirement that the correlations between pairs of latent variables are strictly less than 1, $\mathrm{Cor}(Z_i, Z_j) < 1$, for $1 \leq i < j \leq K$. When the factors have unequal variances, condition (iii) may still hold, but it becomes stronger. We view this as the price to pay for the identifiability of I, and consequently of A in the general model (1.1).

Summarizing, this work is devoted to estimation in model (1.1) with A, C satisfying (i)–(iii). The number of factors K is not known, and both K and p are allowed to grow and be larger than n. In Section 1.1 below, we present our contributions and the structure of this paper. A detailed contrast with existing literature is presented in Section 4.4.

1.1. *Our contributions.*

1.1.1. *Identifiability of the allocation matrix A in sparse latent models with pure variables.* We show, in Proposition 2 of Section 2, that the allocation matrix A, which is allowed to have entries of arbitrary signs, is uniquely defined, up to trivial orthogonal transformations, namely signed permutation matrices. This is a consequence of one of our main results, Theorem 1 of Section 2. In this result, we highlight and resolve the main difficulty in this problem, that of distinguishing between the pure variables and the nonpure variables. Both proofs are constructive, and show that the pure variable set I and allocation matrix A can be determined uniquely from $\Sigma := \mathrm{Cov}(X)$. Moreover, the number of factors K is not assumed to be known, and its determination is also a consequence of Theorem 1. To the best of our knowledge, these are new results in both the latent factors literature and other related matrix factorization literature. We comment on connections to related results in Section 4.4.

1.1.2. *Estimation of the allocation matrix A and of the overlapping clusters. The LOVE algorithm.* We provide an estimator $\widehat{A}$ of the sparse and structured matrix A that is tailored to our model specifications. Our approach follows the constructive techniques used in our identifiability proofs. We first construct $\widehat{I}$, an estimator of the pure variable set I, and $\widehat{K}$, an estimator of the number of clusters, K. These are used to estimate the rows in A corresponding to pure variables. The remaining rows of A are estimated via an easily implementable linear program that is tailored to this problem. As part of our procedure, we also develop a novel estimator (3.7) and (3.8) of a precision matrix, C^{-1}. Our procedure is presented in Sections 3.1, 3.2 and 3.3, respectively. To the best of our knowledge, our estimation strategy is new, and complements the large body of literature in factor models. In particular, we do not resort to optimizing a complicated quasi likelihood function via computationally demanding EM algorithms. These algorithms require, in addition, a notoriously delicate initialization, especially in high dimensions, and typically only convergence to a stationary point can be guaranteed; see Rubin and Thayer (1982). Moreover, as our procedure is not Bayesian, we do not employ distributional assumptions to construct our estimator. In Section 3.4, we build

a collection of overlapping clusters $\widehat{\mathcal{G}}$, using the estimated allocation matrix $\widehat{A}$. The combined procedure is summarized in a new algorithm, LOVE, highlighting our *L*atent model approach to *OVE*rlapping clustering.

1.1.3. *Statistical guarantees*. Our estimation procedure does not depend on distributional assumptions, but for the purpose of our statistical analysis, and in particular our minimax analysis, we assume that $X \in \mathbb{R}^p$ has a sub-Gaussian distribution with $\log p = o(n)$ as $n \to \infty$. LOVE, for appropriate choices of tuning parameters, recovers the population level clusters with a zero false positive proportion and generally low false negative proportion, with high probability, and under a mild condition on the cluster separation as measured by the quantity $\Delta(C)$. This is a direct consequence of a number of results regarding estimation of identifiable loading matrices in factor models satisfying (i)–(iii) and, to the best of our knowledge, they are all new:

(1) Consistent estimation of the number of factors K;
(2) Control of the relationship between $\widehat{I}$ and I for A with entries of arbitrary strength. In particular, we show $I \subseteq \widehat{I} \subseteq I \cup J_1$, where we carefully define and characterize J_1 as the set of quasi-pure variables.
(3) Minimax lower bounds on the norms $L_q(\widehat{A}, A)$, defined below, for all $q \geq 1$, in particular for $q = +\infty$, for A given by model (1.1) under (i)–(iii).
(4) Attainment of these bounds, showing that our procedure is minimax optimal and adaptive.
(5) Control of the relationship between the support of A and the support of $\widehat{A}$.
(6) Control of cluster recovery.

The details are given in Sections 4.2 and 4.3. In particular, we emphasize that (2) above, proved in Theorem 3 of Section 4.1, guarantees recovery of I with minimal mistakes. This result does not require the necessary, yet unpleasant, signal strength restrictions encountered in the typical exact support recovery literature. However, under such restrictions, we also obtain $\widehat{I} = I$, with high probability, in Remark 3 of Section 4.1. Since placing restrictions on the entries in A reduces the number of configurations of interest, the more general result (2) is a new and practically relevant result for pure variable recovery.

Results (3) and (4) are given in Theorems 4, 5 and 6 of Section 4.2. We consider the loss function

$$L_q(\widehat{A}, A) := \min_P \|\widehat{A}P - A\|_{\infty,q}, \quad 1 \leq q \leq \infty,$$

with the minimum taken over all $K \times K$ signed permutation matrices P and

$$\|A\|_{\infty,q} := \max_{1\leq i\leq p} \|A_{i\cdot}\|_q = \max_{1\leq i\leq p} \left(\sum_{j=1}^{K} |A_{ij}|^q \right)^{1/q},$$

is the maximum ℓ_q norm of the rows of A. We let $s = \max_{i\in[p]} \|A_{i\cdot}\|_0$ be the row-sparsity index.

We show that the error of estimation with respect to the L_q loss function, for each q, is proportional to $s^{1/q} n^{-1/2}$, multiplied by $\|C^{-1}\|_{\infty,1}$. This is consistent with the most recent results regarding error rates expressed in terms of the ℓ_q-sensitivity of C in Gautier and Tsybakov (2011) and Belloni, Rosenbaum and Tsybakov (2017), as discussed in Section 4.2. The results hold up to logarithmic factors in p and s.

Results (5) and (6) are presented in Theorem 7 of Section 4.3. Moreover, we can further partition the variables in each cluster into two signed sub-groups consistently. In our model formulation, A is allowed to have positive and negative entries. Since A can only be identified up to signed permutations, one cannot expect sign consistency for $\widehat{A}$. However, we can

TABLE 1
Comparisons

Model (1.1) under (i)–(iii)	Our results	Existing results in comparable factor models
Identifiability conditions	Existence of I with I and K *unknown*. C is positive definite and satisfies (iii).	Existence of *known* I and K. C is positive definite.
Estimation: I	Runs in $O(p^2)$ time; optimization-free.	×
Estimation: A	Not MLE-based approach. Unique solution. Linear program; runs in $O(p^2 + pK)$.	MLE-based approach. Multiple solutions. EM algorithm; computationally involved.
Guarantees: I	Recovered	×
Guarantees: A	Finite sample $\|\ \|_{\infty,q}$ lower bounds. Adaptive finite sample upper bounds. Both p and K can grow with n.	Row-wise asymptotic normality of MLE. Only p can grow with n and K is *fixed*.
Cluster recovery	Guaranteed	×

identify consistently the two subgroups of each cluster that contain variables that are associated with the common latent factor in the same direction, although the direction itself is not identifiable. These results are presented in Section 4.3.

We conduct an extensive simulation study in Section 5 to assess the numerical performance of our proposed strategy. The study confirms our theoretical findings. We conclude the validation of our approach with a data analysis, devoted to determining the functional annotation of genes with unknown function. Our analysis confirms existing biological ground truths, as our procedure tends to cluster together genes with the same Gene Ontology (GO) biological process, molecular function or cellular component terms.

We summarize our contributions in Table 1, restricting attention to estimation in general latent models (1.1) under (i)–(iii), without any further restrictions on the signs or scales of X and Z.

In Section 4.4, we discuss our results further, and provide a detailed comparison between our work and related contributions. All proofs are deferred to Section A of the Supplementary Material (Bing et al. (2019)).

1.2. *Notation*. We use the following notation throughout this paper. For the n consecutive integer set starting from 1, we write $[n] = \{1, \dots, n\}$. The sign of any generic number N is denoted by $\operatorname{sign}(N)$. For any $m \times d$ matrix M and index sets $I \subseteq \{1, \dots, m\}$ and $J \subseteq \{1, \dots, d\}$, we write M_I to denote the $|I| \times d$ submatrix $(M_{ij})_{i\in I, 1\le j\le d}$ of M consisting of the rows in the index set I, while we denote by M_{IJ} the $|I| \times |J|$ submatrix with entries M_{ij}, $i \in I$ and $j \in J$. The ith row of M is denoted by $M_{i\cdot}$, and the jth column of M is denoted by $M_{\cdot j}$. Let $\|M\|_\infty = \max_{1\le j\le m, 1\le k\le d} |M_{jk}|$, $\|M\|_1 = \sum_{1\le j\le m, 1\le k\le d} |M_{jk}|$, $\|M\|_F = (\sum_{j=1}^m \sum_{k=1}^d M_{jk}^2)^{1/2}$, $\|M\|_{\infty,1} = \max_{1\le j\le m} \sum_{k=1}^d |M_{jk}|$ and $\|M\|_{1,\infty} = \max_{1\le k\le d} \sum_{j=1}^m |M_{jk}|$ denote the matrix max norm, matrix ℓ_1 norm, matrix Frobenius norm, matrix 1 norm and matrix ∞ norm. We denote by $\langle\cdot\rangle$ the Frobenius scalar product. For a vector $v \in \mathbb{R}^d$, define $\|v\|_q = (\sum_{i=1}^d |v_j|^q)^{1/q}$ for $1 \le q < \infty$, $\|v\|_\infty = \max_{1\le j\le d} |v_j|$ and $\|v\|_0 = |\operatorname{supp}(v)|$, where $\operatorname{supp}(v) = \{j : v_j \ne 0\}$ and $|S|$ is the cardinality of the set S. For a vector $v \in \mathbb{R}^d$, we denote by v_S the vector $w \in \mathbb{R}^d$ that has the same coordinates $w_i = v_i$ as v on the index set $S \subseteq \{1, \dots, d\}$ and zero coordinates otherwise ($w_i = 0$ for all $i \in \bar{S} := [d] \setminus S$). We write M^T for the transpose of M and $\operatorname{diag}(m_1, \dots, m_d)$

for the $d \times d$ diagonal matrix with elements $m_1, \ldots, m_d$ on its diagonal, while $\operatorname{diag}(M)$ is the diagonal matrix obtained from the diagonal elements of a square matrix M. The identity matrix in $\mathbb{R}^{d\times d}$ is denoted by $\boldsymbol{I}_d$, the vector in $\mathbb{R}^d$ with all entries equal to one is denoted by $\mathbf{1}_d$ and a vector/matrix with all zero entries is denoted by $\mathbf{0}$ whose dimension might vary line by line. We use $c_0, c_1, \ldots$ to denote generic constants. Finally, a signed permutation matrix is an orthogonal matrix that permutes the index and switches the sign within each column. We write $\mathcal{H}_K$ as the hyperoctahedral group of $K \times K$ signed permutation matrices.

2. Identifiability. In this section, we show that the allocation matrix A given by Model (1.1) and (i)–(iii) is identifiable, up to multiplication with a signed permutation matrix.

For any $A \in \mathbb{R}^{p\times K}$ which satisfies model (1.1), we can partition the set $[p] = \{1, \ldots, p\}$ into two disjoint parts: I and its complement $J := [p] \setminus I$ such that for each row $A_{i\cdot}$ of A_I, there exists only one $a \in [K]$ such that $|A_{ia}| = 1$. We name I the pure variable set and J the nonpure variable set. Specifically, for any given A, the pure variable set I is defined as

$$I(A) := \bigcup_{a=1}^{K} I_a, \quad I_a := \{i \in [p] : |A_{ia}| = 1, A_{ib} = 0, \text{ for any } b \neq a\}. \tag{2.1}$$

We write $I(A)$ in (2.1) to emphasize that the pure variable set is defined relative to A. In the following, we will not write this explicitly when there is no confusion. We also note that the sets $\{I_a\}_{1\le a\le K}$ form a partition of I.

To show the identifiability of A, it suffices to show that A_I and A_J are identifiable, respectively, up to signed permutation matrices. By the definition of A_I, this matrix is identifiable provided the partition of the pure variable set I is. The identifiability of I, and thus the problem of distinguishing between the sets I and J, on the basis of the distribution of X alone, is the central challenge in this problem. We meet this challenge in Theorem 1 below: part (a) offers a necessary and sufficient characterization of I; part (b) shows that, as a consequence, I and its partition $\mathcal{I} := \{I_a\}_{1\le a\le K}$ are identifiable. Let

$$M_i := \max_{j\in[p]\setminus\{i\}} |\Sigma_{ij}| \tag{2.2}$$

be the largest absolute value of the entries of row i of Σ excluding $|\Sigma_{ii}|$. Let S_i be the set of indices for which M_i is attained:

$$S_i := \{j \in [p] \setminus \{i\} : |\Sigma_{ij}| = M_i\}. \tag{2.3}$$

THEOREM 1. *Assume that model* (1.1) *and* (*i*)–(*iii*) *hold. Then*:

(a) $i \in I \iff M_i = M_j$ *for all* $j \in S_i$.

(b) *The pure variable set* I *can be determined uniquely from* $\Sigma := \operatorname{Cov}(X)$. *Moreover, its partition* $\mathcal{I} := \{I_a\}_{1\le a\le K}$ *is unique and can be determined from* Σ *up to label permutations*.

The identifiability of the allocation matrix A and that of the collection of clusters $\mathcal{G} = \{G_1, \ldots, G_k\}$ in (1.2) use the results from Theorem 1 in crucial ways. We state the result in Theorem 2 below.

THEOREM 2. *Assume that Model* (1.1) *with* (*i*)–(*iii*) *holds. Then there exists a unique matrix* A, *up to a signed permutation, such that* $X = AZ + E$. *This implies that the associated overlapping clusters* G_a, *for* $1 \le a \le K$, *are identifiable, up to label switching*.

REMARK 1. We show below that the pure variable assumption (ii) is needed for the identifiability of A, up to a signed permutation. Assume that $X = AZ + E$ satisfies (i) and

(iii), but not (ii). We construct an example in which X can also be written as $X = \tilde{A}\tilde{Z} + E$, where $\tilde{A}$ and $\tilde{Z}$ satisfy the same conditions (i) and (iii), respectively, but $\tilde{A} \neq AP$ for any $K \times K$ signed permutation matrix P and $\tilde{A}$ may have a sparsity pattern different from A. To this end, we construct $\tilde{A}$ and $\tilde{Z}$ such that $\tilde{A}\tilde{Z} = AZ$. Let $\tilde{A} = AQ$ and $\tilde{Z} = Q^{-1}Z$, for some $K \times K$ invertible matrix Q to be chosen such that $\mathrm{Cov}(\tilde{Z}) = Q^{-1}C(Q^{-1})^T$ satisfies (iii). In addition, we need to guarantee that $\tilde{A} = AQ$ satisfies (i). For simplicity, we set $K = 3$. The following example satisfies all our requirements:

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \qquad Q = \begin{bmatrix} 1 & 1/3 & 0 \\ 1/3 & 2 & 1/2 \\ 0 & 1/2 & 2 \end{bmatrix}.$$

It is easy to verify that $\mathrm{Cov}(\tilde{Z}) = Q^{-1}C(Q^{-1})^T$ satisfies (iii). For any $1 \le j \le p$, consider

$$A_{j\cdot}^T = (1/8, -3/8, 0)$$

then

$$\tilde{A}_{j\cdot}^T = A_{j\cdot}^T Q = (0, -17/24, -3/16)$$

which also satisfies condition (i). However, $A_{j\cdot}$ and $\tilde{A}_{j\cdot}$ have different sparsity patterns. Thus, if the matrix A does not satisfy (ii), A is generally not identifiable.

3. Estimation. We develop estimators from the observed data, which is assumed to be a sample of n i.i.d. copies $X^{(1)}, \ldots, X^{(n)}$ of $X \in \mathbb{R}^p$, where p is allowed to be larger than n. Our estimation procedure consists of the following four steps:

(1) Estimate the pure variable set I, the number of clusters K and the partition $\mathcal{I}$;
(2) Estimate A_I, the submatrix of A with rows $A_{i\cdot}$ that correspond to $i \in I$;
(3) Estimate A_J, the submatrix of A with rows $A_{j\cdot}$ that correspond to $j \in J$;
(4) Estimate the overlapping clusters $\mathcal{G} = \{G_1, \ldots, G_K\}$.

3.1. *Estimation of I and $\mathcal{I}$.* Given the different nature of their entries, we estimate the submatrices A_I and A_J separately. For the former, we first estimate I and its partition $\mathcal{I} = \{I_1, \ldots, I_K\}$, which can be both uniquely constructed from Σ, as shown by Theorem 1. We use the constructive proof of Theorem 1 for this step, replacing the unknown Σ by the sample covariance matrix

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

Specifically, we iterate through the index set $\{1, 2, \ldots, p\}$, and use the sample version of part (**a**) of Theorem 1 to decide whether an index i is pure. If it is not deemed to be pure, we add it to the set that estimates J. Otherwise, we retain the estimated index set $\widehat{S}_i$ of S_i defined in (2.3), which corresponds to an estimator of M_i given by (2.2). We then use the constructive proof of part (**b**) of Theorem 1 to declare $\widehat{S}_i \cup \{i\} := \widehat{I}^{(i)}$ as an estimator of one of the partition sets of $\mathcal{I}$. The resulting procedure has complexity $O(p^2)$, and we give all the specifics in Algorithm 1 of Section 3.5. The algorithm requires the specification of a tuning parameter δ, which will be discussed in Section 5.1.

3.2. *Estimation of the allocation submatrix A_I.* Given the estimators $\widehat{I}$, $\widehat{K}$ and $\widehat{\mathcal{I}} = \{\widehat{I}_1, \ldots, \widehat{I}_{\widehat{K}}\}$ from Algorithm 1, we estimate the matrix A_I by a $|\widehat{I}| \times \widehat{K}$ matrix with rows $i \in \widehat{I}$ consisting of $\widehat{K} - 1$ zeros and one entry equal to either $+1$ or -1 as follows. For each $a \in [\widehat{K}]$,

(1) Pick an element $i \in \widehat{I}_a$ at random, and set $\widehat{A}_{ia} = 1$. Note that $\widehat{A}_{ia}$ can only be $+1$ or -1 by the definition of a pure variable.

(2) For the remaining $j \in \widehat{I}_a \setminus \{i\}$, we set $\widehat{A}_{ja} = \operatorname{sign}(\widehat{\Sigma}_{ij})$.

This procedure induces a partition of $\widehat{I}_a = \widehat{I}_a^1 \cup \widehat{I}_a^2$, where $\widehat{I}_a^1$ and $\widehat{I}_a^2$ are defined below:

$$\begin{cases} \widehat{A}_{ka} = \widehat{A}_{la}, & \text{for } k, l \in \widehat{I}_a^1 \text{ or } k, l \in \widehat{I}_a^2, \\ \widehat{A}_{ka} \neq \widehat{A}_{la}, & \text{for } k \in \widehat{I}_a^1 \text{ and } l \in \widehat{I}_a^2, \end{cases} \tag{3.1}$$

3.3. *Estimation of the allocation submatrix A_J.* We continue by estimating the matrix A_J, row by row. To motivate our procedure, we begin by highlighting the structure of each row $A_{j\cdot}$ of A_J, for $j \in J$. We recall that $A_{j\cdot}$ is sparse, with $\|A_{j\cdot}\|_1 \leq 1$, for each $j \in J$, as specified by assumption (i). In addition, model (1.1) subsumes a further constraint on each row $A_{j\cdot}$ of A, as explained below. To facilitate notation, we rearrange Σ, A and Γ as follows:

$$\Sigma = \begin{bmatrix} \Sigma_{II} & \Sigma_{IJ} \\ \Sigma_{JI} & \Sigma_{JJ} \end{bmatrix}, \qquad A = \begin{bmatrix} A_I \\ A_J \end{bmatrix} \quad \text{and} \quad \Gamma = \begin{bmatrix} \Gamma_{II} & 0 \\ 0 & \Gamma_{JJ} \end{bmatrix}.$$

Model (1.1) implies the following decomposition of the covariance matrix Σ of X:

$$\Sigma = \begin{bmatrix} \Sigma_{II} & \Sigma_{IJ} \\ \Sigma_{JI} & \Sigma_{JJ} \end{bmatrix} = \begin{bmatrix} A_I C A_I^T & A_I C A_J^T \\ A_J C A_I^T & A_J C A_J^T \end{bmatrix} + \begin{bmatrix} \Gamma_{II} & 0 \\ 0 & \Gamma_{JJ} \end{bmatrix}.$$

In particular, $\Sigma_{IJ} = A_I C A_J^T$. Thus, for each $i \in I_a$ with some $a \in [K]$ and $j \in J$, we have

$$A_{ia}\Sigma_{ij} = A_{ia}^2 \sum_{b=1}^{K} A_{jb} C_{ab} = \sum_{b=1}^{K} A_{jb} C_{ab} = C_{a\cdot}^T A_{j\cdot}. \tag{3.2}$$

Averaging display (3.2) over all $i \in I_a$ yields

$$\frac{1}{|I_a|} \sum_{i \in I_a} A_{ia}\Sigma_{ij} = C_{a\cdot}^T A_{j\cdot} \quad \text{for each } a \in [K]. \tag{3.3}$$

For each $j \in J$, we let

$$\beta^j := A_{j\cdot}$$

and

$$\theta^j = \left(\frac{1}{|I_1|} \sum_{i \in I_1} A_{i1}\Sigma_{ij}, \ldots, \frac{1}{|I_K|} \sum_{i \in I_K} A_{iK}\Sigma_{ij} \right)^T. \tag{3.4}$$

Since $A_{ia} \in \{-1, 1\}$, for each $i \in I_a$ and $a \in [K]$, the entries of θ^j are respective averages of the sign corrected entries of Σ corresponding to the partition of the pure variable set. Summarizing, modeling assumption (i) and equation (3.3) above show that the estimation of A_J reduces to estimating, for each $j \in J$, a K-dimensional vector β^j that is sparse, with norm $\|\beta^j\|_1 \leq 1$, and that satisfies the equation

$$\theta^j = C\beta^j.$$

Both C and θ^j, for each $j \in J$, can be estimated directly from the data as follows. For each $j \in \widehat{J}$, we estimate the ath entry of θ^j by

$$\widehat{\theta}_a^j = \frac{1}{|\widehat{I}_a|} \sum_{i \in \widehat{I}_a} \widehat{A}_{ia}\widehat{\Sigma}_{ij}, \quad a \in [\widehat{K}], \tag{3.5}$$

and compute

$$\begin{aligned}\widehat{C}_{aa} &= \frac{1}{|\widehat{I}_a|(|\widehat{I}_a|-1)} \sum_{i,j\in\widehat{I}_a, i\neq j} |\widehat{\Sigma}_{ij}|,\\ \widehat{C}_{ab} &= \frac{1}{|\widehat{I}_a||\widehat{I}_b|} \sum_{i\in\widehat{I}_a, j\in\widehat{I}_b} \widehat{A}_{ia}\widehat{A}_{ib}\widehat{\Sigma}_{ij},\end{aligned} \tag{3.6}$$

for each $a \in [\widehat{K}]$ and $b \in [\widehat{K}] \setminus \{a\}$ to form the estimator $\widehat{C}$ of C. The estimates (3.5) and (3.6) rely crucially on having first estimated the pure variables and their partition, according to the steps described in Sections 3.1 and 3.2 above.

We have developed a computationally efficient method to estimate β^j. We exploit the fact that the square matrix C is invertible and take the equation $\beta^j = C^{-1}\theta^j$ as our starting point. The idea is to first construct a pre-estimator $\bar{\beta}^j = \widehat{\Omega}\widehat{\theta}^j$, based on an appropriate estimator $\widehat{\Omega}$ of the precision matrix $\Omega := C^{-1}$, followed by a sparse projection of $\bar{\beta}^j$. Alternatively, and recommended to speed up the computation, we could use a simple hard threshold operation in the second step as described in Remark 5, item 4. We first motivate our proposed estimator of Ω. From the decomposition,

$$\begin{aligned}\bar{\beta}^j - \beta^j &= \widehat{\Omega}\big(\widehat{\theta}^j - \theta^j\big) + (\widehat{\Omega} - \Omega)\theta^j\\ &= \widehat{\Omega}\big(\widehat{\theta}^j - \theta^j\big) + (\widehat{\Omega}C - I)\beta^j,\end{aligned} \tag{3.7}$$

we immediately have

$$\big\|\bar{\beta}^j - \beta^j\big\|_\infty \le \|\widehat{\Omega}\|_{\infty,1}\big\|\widehat{\theta}^j - \theta^j\big\|_\infty + \|\widehat{\Omega}C - I\|_\infty\big\|\beta^j\big\|_1. \tag{3.8}$$

Since we can show in Lemma 12 of the Supplementary Material that $\|\widehat{\theta}^j - \theta^j\|_\infty$ has optimal convergence rate, and since $\|\beta^j\|_1 \le 1$ under our model, our estimator $\widehat{\Omega}$ should ideally render values for $\|\widehat{\Omega}\|_{\infty,1}$ and $\|\widehat{\Omega}C - I\|_\infty$ that are as small as possible. With this in mind, we propose the linear program

$$(\widehat{\Omega}, \widehat{t}) = \arg \min_{t\in\mathbb{R}^+, \Omega\in\mathbb{R}^{\widehat{K}\times\widehat{K}}} t \tag{3.9}$$

subject to

$$\Omega = \Omega^T, \qquad \|\Omega\widehat{C} - I\|_\infty \le \lambda t, \qquad \|\Omega\|_{\infty,1} \le t, \tag{3.10}$$

with tuning parameter λ. This linear programming problem is clearly tailored to our purpose, and its optimal solution $\widehat{\Omega}$ adds a novel estimator for C^{-1} to the rich literature on precision matrix estimation (Cai, Liu and Luo (2011), Cai, Liu and Zhou (2016), Friedman, Hastie and Tibshirani (2008), Meinshausen and Bühlmann (2006), Yuan and Lin (2007), to name a few). Its novelty consists in (a) the usage of the matrix $\|\cdot\|_{\infty,1}$ norm, instead of the commonly used matrix $\|\cdot\|_1$ norm, and (b) the fact that this norm appears in the upper bound of the restriction (3.10). After we compute $\bar{\beta}^j = \widehat{\Omega}\widehat{\theta}^j$, for each $j \in \widehat{J}$, we solve the following optimization problem:

$$\widehat{\beta}^j = \arg \min_{\beta\in\mathbb{R}^{\widehat{K}}} \|\beta\|_1 \tag{3.11}$$

subject to

$$\big\|\beta - \bar{\beta}^j\big\|_\infty \le \mu, \tag{3.12}$$

for some tuning parameter μ that is proportional to $\|C^{-1}\|_{\infty,1}$ to obtain our final estimate $\widehat{\beta}^j$ as the optimal solution of this linear program. This solution is also sparse and properly

scaled, in accordance to our model specification (i). Then $\widehat{A}_{\widehat{J}}$ is the matrix with rows $\widehat{\beta}^j$, for $j \in \widehat{J}$. Our final estimator $\widehat{A}$ of A is obtained by concatenating $\widehat{A}_{\widehat{I}}$ and $\widehat{A}_{\widehat{J}}$. Its statistical property is analyzed in Section 4, along with precise forms of the tuning parameters needed for its construction.

An alternative way to estimate β^j is by the following Dantzig-type estimator. Starting with the equation $\theta^j = C\beta^j$, we can consider, for each $j \in \widehat{J}$, the linear program

$$\min_{\beta \in \mathbb{R}^{\widehat{K}}} \|\beta\|_1 \tag{3.13}$$

subject to

$$\|\widehat{C}\beta - \widehat{\theta}^j\|_\infty \le \lambda', \tag{3.14}$$

with tuning parameter λ'. The solution is sparse and properly scaled, in accordance to our model specification (i). Our final goal of support recovery of β^j still requires an additional hard thresholding step of the solution of this linear program. In this case, the appropriate threshold μ is proportional to the ℓ_∞-sensitivity of the matrix C, introduced by Gautier and Tsybakov (2011). The latter quantity depends on the unknown support of the different rows θ^j, but can be upper bounded by $\|C^{-1}\|_{\infty,1}$. The statistical properties of this procedure are analyzed in Section 4 as well.

Both procedures require, in practice, the estimation of the quantity $\|C^{-1}\|_{\infty,1}$. The procedure in (3.11)–(3.12) recovers the support of β automatically while the procedure in (3.13)–(3.14), even though it renders a sparse solution, requires a further hard-thresholding step for the support recovery.

3.4. *Estimation of the overlapping groups.* Recalling the definition of groups in (1.2), the overlapping groups are estimated by

$$\widehat{\mathcal{G}} = \{\widehat{G}_1, \dots, \widehat{G}_{\widehat{K}}\}, \quad \widehat{G}_a = \{i \in [p] : \widehat{A}_{ia} \neq 0\}, \text{ for each } a \in [\widehat{K}]. \tag{3.15}$$

Variables X_i that are associated (via $\widehat{A}$) with the same latent factor Z_a are therefore placed in the same group $\widehat{G}_a$. To accommodate potential pure noise variables, we further define

$$G_0 := \{j \in \{1, \dots, p\} : A_{ja} = 0, \text{ for all } a \in \{1, \dots, K\}\} \tag{3.16}$$

as the pure noise cluster. We can estimate G_0 in (3.16) by

$$\widehat{G}_0 = \{i \in [p] : \widehat{A}_{ia} = 0, \text{ for all } a \in [\widehat{K}]\}. \tag{3.17}$$

However, our main focus is on $\mathcal{G}$ because it completely determines G_0.

In many applications, it may be of interest to identify the subgroups of variables that are all either positively or negatively associated with the same latent factor. To this end, we define

$$\begin{aligned} \mathcal{G}^s &:= \{G_1^s, \dots, G_K^s\}, \\ G_a^s &:= \{G_a^1, G_a^2\} := \{\{i \in G_a : A_{ia} > 0\}, \{i \in G_a : A_{ia} < 0\}\}, \end{aligned} \tag{3.18}$$

for each $a \in [K]$, and they are estimated by

$$\begin{aligned} \widehat{\mathcal{G}}^s &= \{\widehat{G}_1^s, \dots, \widehat{G}_{\widehat{K}}^s\}, \\ \widehat{G}_a^s &= \{\{i \in \widehat{G}_a : \widehat{A}_{ia} > 0\}, \{i \in \widehat{G}_a : \widehat{A}_{ia} < 0\}\}, \end{aligned} \tag{3.19}$$

for each $a \in [\widehat{K}]$. The fact that A is only identifiable up to a signed permutation matrix has the repercussion that the labels of the two subgroups in G_a^s are not identifiable. Thus, variables placed in the subgroups G_a^1 and G_a^2 are respectively associated with Z_a in the same direction. The directions between two subgroups, henceforth called direction subgroups, are opposite. This can be identified, although the direction itself cannot. We show in Section 4 that the direction subgroups can be identified, and well estimated.

Algorithm 1 Estimate the partition of the pure variables $\mathcal{I}$ by $\widehat{\mathcal{I}}$

procedure PUREVAR($\widehat{\Sigma}$, δ)
  $\widehat{\mathcal{I}} \leftarrow \varnothing$.
  for all $i \in [p]$ **do**
    $\widehat{I}^{(i)} \leftarrow \{l \in [p] \setminus \{i\} : \max_{j \in [p]\setminus\{i\}} |\widehat{\Sigma}_{ij}| \le |\widehat{\Sigma}_{il}| + 2\delta\}$
    Pure(i) $\leftarrow$ *True*.
    for all $j \in \widehat{I}^{(i)}$ **do**
      if $||\widehat{\Sigma}_{ij}| - \max_{k \in [p]\setminus\{j\}} |\widehat{\Sigma}_{jk}|| > 2\delta$ **then**
        Pure(i) $\leftarrow$ *False*,
        break
    if Pure(i) **then**
      $\widehat{I}^{(i)} \leftarrow \widehat{I}^{(i)} \cup \{i\}$
      $\widehat{\mathcal{I}} \leftarrow$ MERGE($\widehat{I}^{(i)}, \widehat{\mathcal{I}}$)
  return $\widehat{\mathcal{I}}$ and $\widehat{K}$ as the number of sets in $\widehat{\mathcal{I}}$

function MERGE($\widehat{I}^{(i)}, \widehat{\mathcal{I}}$)
  for all $G \in \widehat{\mathcal{I}}$ **do** ▷ $\widehat{\mathcal{I}}$ is a collection of sets
    if $G \cap \widehat{I}^{(i)} \neq \varnothing$ **then**
      $G \leftarrow G \cap \widehat{I}^{(i)}$ ▷ Replace $G \in \widehat{\mathcal{I}}$ by $G \cap \widehat{I}^{(i)}$
      return $\widehat{\mathcal{I}}$
  $\widehat{I}^{(i)} \in \widehat{\mathcal{I}}$ ▷ add $\widehat{I}^{(i)}$ in $\widehat{\mathcal{I}}$
  return $\widehat{\mathcal{I}}$

3.5. *LOVE*: *A* L*atent variable model approach for* OVE*rlapping clustering*. We give below the specifics of Algorithm 1, motivated in Section 3.1, and summarize our final algorithm, LOVE in Algorithm 2.

4. Statistical guarantees. We provide in this section statistical guarantees for:

(1a) The estimated number of clusters $\widehat{K}$;
(1b) The estimated pure variable set $\widehat{I}$ and its estimated partition $\widehat{\mathcal{I}}$;
(2) The estimated allocation matrix $\widehat{A}$ and its adaptation to the unknown row sparsity of A.
(3) The individual Group False Positive Proportion (GFPP), the individual Group False Negative Proportion(GFNP), the Total False Positive Proportion (TFPP) and the Total False Negative Proportion (TFNP) for the estimated overlapping groups.

Algorithm 2 The LOVE procedure for overlapping clustering

Require: $\widehat{\Sigma}$ from I.I.D. data $(X^{(1)}, \ldots, X^{(n)})$, the tuning parameters δ, λ and μ.
1: Apply Algorithm 1 to obtain the number of clusters $\widehat{K}$, the estimated set of pure variables $\widehat{I}$ and its partition of $\widehat{\mathcal{I}}$.
2: Estimate A_I by $\widehat{A}_{\widehat{I}}$ from (3.1).
3: Estimate C^{-1} by $\widehat{\Omega}$ from (3.9) and $\bar{\beta}^j$ for each $j \in \widehat{J}$.
4: Estimate A_J by $\widehat{A}_{\widehat{J}}$ from (3.11). Combine $\widehat{A}_{\widehat{I}}$ with $\widehat{A}_{\widehat{J}}$ to obtain $\widehat{A}$.
5: Estimate overlapping groups $\widehat{\mathcal{G}} = \{\widehat{G}_1, \ldots, \widehat{G}_{\widehat{K}}\}$ and its direction subgroups $\widehat{\mathcal{G}}^s = \{\widehat{G}^s_1, \ldots, \widehat{G}^s_{\widehat{K}}\}$ from (3.15)–(3.19) by using $\widehat{A}$.
6: Output $\widehat{A}$, $\widehat{\mathcal{G}}$ and $\widehat{\mathcal{G}}^s$.

We make the blanket assumption for the remainder of this paper that X is *sub-Gaussian*, that is, the Orlicz norm $\|X_j\|_{\psi_2}$ of each X_j is bounded by a common constant σ_*.[1] The sub-Gaussian condition implies $\max_{j\in[p]} \Sigma_{jj} \le 2\sigma_*^2$ and $\|C\|_\infty \le 2\sigma_*^2$. Let

$$\mathcal{E} = \mathcal{E}(\delta) := \Big\{ \max_{1\le i<j\le p} |\widehat{\Sigma}_{ij} - \Sigma_{ij}| \le \delta \Big\}. \tag{4.1}$$

We assume throughout that $\delta = c_0 \|\Sigma\|_\infty \sqrt{\log(p\vee n)/n}$, for some absolute constant c_0, and $\log p = o(n)$, so that $\delta = o(1)$, for n large enough, where $a \vee b = \max(a, b)$. Taking $c_0 > 0$ large enough, Lemma 2 in Bien, Bunea and Xiao (2016) guarantees that $\mathcal{E}$ holds with high probability:

$$\mathbb{P}(\mathcal{E}) \ge 1 - c_1 (p\vee n)^{-c_2} \tag{4.2}$$

for some positive, finite constants c_1 and c_2. Apart from δ, the quantity

$$\Delta(C) := \nu > 0, \tag{4.3}$$

plays an important role in our analysis. Indeed, assumption (iii) requires that $\nu > 0$ in order to guarantee that the latent factors are distinguishable from one another. We can view ν as a measure of their separation, and naturally therefore, the size of ν impacts the quality of all our estimators, in addition to the magnitude of δ.

REMARK 2. It is common practice to standardize the data in a pre-processing step, and perform statistical analyses on the standardized data. Our model can be easily adapted to this case by assuming that the latent variable model holds for a standardized version of X, specifically for $\widetilde{X} := (\mathrm{diag}(\Sigma))^{-1/2}(X - \mathbb{E}(X))$, leading to

$$\widetilde{X} = AZ + E \tag{4.4}$$

with A, Z and E satisfying the same conditions (i), (ii) and (iii). Recall that in model (1.1) we have already assumed that X has mean zero. Transforming model (4.4) back to the original scale, we have $X - \mathbb{E}[X] = [(\mathrm{diag}(\Sigma))^{1/2}A]Z + [(\mathrm{diag}(\Sigma))^{1/2}E]$. We note that the new allocation matrix $\widetilde{A} := [(\mathrm{diag}(\Sigma))^{1/2}A]$ has the same support as A. Moreover, a pure variable j in cluster a satisfies $|\widetilde{A}_{ja}| = \Sigma_{jj}^{1/2}$. Therefore, pure variables are given different weights, proportional to their respective standard deviations, which relaxes the equal weight restriction in Condition (ii). The caveat is that under (4.4), we have $1 = \mathrm{Cov}(\widetilde{X}_j) = A_{j\cdot}^T C A_{j\cdot} + \Gamma_{jj}$ for any $1 \le j \le p$. This further implies that $\Gamma_{jj} = \Gamma_{j'j'}$ for any $j, j' \in I_a$, that is, model (4.4) subsumes that the random noise has the same variance for all pure variables in each cluster. Depending on what modeling assumptions best fit a particular problem, either (1.1) or (4.4) can be considered. The identifiability of model (4.4) follows directly from the proof of Theorem 2. The LOVE algorithm, presented in the next subsection, is also applicable, provided we replace the sample covariance matrix $\widehat{\Sigma}$ with the sample correlation matrix $\widehat{R}$ with entries

$$\widehat{R}_{jk} = \frac{1}{n}\sum_{i=1}^n (X_j^{(i)} - \bar{X}_j)(X_k^{(i)} - \bar{X}_k) / \big(\mathrm{sd}(X_j)\,\mathrm{sd}(X_k)\big),$$

with $\bar{X}_j = n^{-1}\sum_{i=1}^n X_j^{(i)}$ and $\mathrm{sd}(X_j) = \{n^{-1}\sum_{i=1}^n (X_j^{(i)} - \bar{X}_j)^2\}^{1/2}$. Then all our theoretical guarantees hold unchanged on the new event

$$\mathcal{E} = \mathcal{E}(\delta) := \Big\{ \max_{1\le i<j\le p} |\widehat{R}_{ij} - R_{ij}| \le \delta \Big\}.$$

[1] The Orlicz norm of X_j is defined as $\|X_j\|_{\psi_2} = \inf\{c > 0 : \mathbb{E}[\psi_2(|X_j|/c)] < 1\}$, based on the Young function $\psi_2(x) = \exp(x^2) - 1$.

Since Bunea, Giraud and Luo (2016) showed that $\mathcal{E}$ holds with high probability by choosing $\delta = c_0\sqrt{\log(p \vee n)/n}$, for some constant c_0, we can obtain the same statistical guarantees under the model (4.4).

4.1. *Statistical guarantees for $\widehat{K}$, $\widehat{I}$ and $\widehat{\mathcal{I}}$.* We first analyze the performance of our estimator $\widehat{I}$ of I, and its corresponding partition. This problem belongs to the general class of pattern recovery problems, and it is well understood that under strong enough signal conditions one can expect $\widehat{I} = I$, with high probability. This turns out to be indeed the case for our problem, but we obtain this as a corollary of a more general result. We set out to quantify when our estimated set contains the least taxing type of errors, under minimal assumptions. To make this precise, we introduce the concept of *quasi-pure* variables. A quasi-pure variable X_i has very strong association with only one latent factor, say Z_a, in that $|A_{ia}| \approx 1$, and very low association with the rest: $|A_{ib}| \approx 0$, for all $b \neq a$. Formally, we define the set of quasi-pure variables as

$$J_1 := \big\{ j \in J : \text{there exists } a \in [K], \text{ such that } |A_{ja}| \geq 1 - 4\delta/\nu \big\}. \tag{4.5}$$

For each $a \in [K]$, we further define the set of quasi-pure variables associated with the same factor:

$$J_1^a := \big\{ j \in J_1 : |A_{ja}| \geq 1 - 4\delta/\nu \big\}. \tag{4.6}$$

When ν is a strictly positive constant, $\epsilon := 4\delta/\nu = o(1)$. The lower bound $|A_{ja}| \geq 1 - \epsilon$ in (4.6) implies, under condition (ii), that $|A_{jb}| \leq \epsilon$, for any $b \neq a$ and $j \in J_1^a$, justifying the name quasi-pure variables for those components of X with indices in J_1. We observe, for future reference, that $\{J_1^1, \ldots, J_1^K\}$ forms a partition of J_1.

We show in Theorem 3 that, with very high probability, the estimated $\widehat{I}$ contains the pure variable set I, and is in turn contained in a set that includes all pure variables and quasi-pure variables. Importantly, $\widehat{I}$ will not include indices of variables X_j that are associated with multiple latent factors at a level higher than ϵ. Equally importantly, if a quasi-pure variable X_i is included in $\widehat{I}$, then this variable will have the corresponding $|A_{ia}| \approx 1$, and it will be placed together with the pure variables associated with the same factor Z_a, for some a, and not in a new cluster. This is crucial for ensuring that the number of clusters K is consistently estimated, and also for establishing the cluster misclassification proportion in Section 4.3 below.

THEOREM 3. *Assume Model* (1.1) *with* (*i*)–(*iii*), *and*

$$\nu > 2\max\big(2\delta, \sqrt{2\|C\|_\infty \delta}\big). \tag{4.7}$$

Then:

(a) $\widehat{K} = K$;
(b) $I \subseteq \widehat{I} \subseteq I \cup J_1$.

Moreover, there exists a label permutation π of the set $\{1, \ldots, K\}$, such that the output $\widehat{\mathcal{I}} = \{\widehat{I}_a\}_{a \in [K]}$ from Algorithm 1 satisfies:

(c) $I_{\pi(a)} \subseteq \widehat{I}_a \subseteq I_{\pi(a)} \cup J_1^{\pi(a)}$.

All results hold with probability larger than $1 - c_1(n \vee p)^{-c_2}$, for c_1, c_2 positive constants defined in (4.2).

The conclusion of Theorem 3 holds only under condition (4.7), which stipulates that the separation between the latent factors, as measured by ν, is not only strictly positive, which was needed for identifiability, but slightly above a quantity that depends on the estimation error δ, and which becomes $o(1)$ for n large enough. From the inspection of the proof, condition (4.7) can be relaxed to $\nu > 4\delta$ when $J_1 = \varnothing$.

REMARK 3. Let $e_1 = (1, 0, \ldots, 0)^T$ and $\mathcal{H}_K$ be the hyperoctahedral group of signed permutation matrices. If A_I and A_J are well separated in the sense that

$$\min_{j \in J, P \in \mathcal{H}_K} \|A_{j\cdot} - Pe_1\|_1 > 8\delta/\nu,$$

then $J_1 = \varnothing$, and Theorem 3 yields exact recovery of the pure variable set and of its partition: $\widehat{I} = I$ and $\widehat{\mathcal{I}} = \mathcal{I}$, with high probability. However, we expect $J_1 \neq \varnothing$, as we expect quasi-pure variables to be present in a high-dimensional model, which is the context for which Theorem 3 has been established.

4.2. *Statistical guarantee for* $\widehat{A}$. In this section, we state and comment on the statistical properties of the estimate $\widehat{A}$ obtained in Sections 3.2 and 3.3. Recall that $\delta = O(\sqrt{\log(p \vee n)/n})$ was given in (4.1) above, and the estimation of A_J made use of two tuning parameters: λ, in (3.10) and μ, in (3.12). Theorem 4 establishes the properties of our estimates relative to the theoretically optimal values of these tuning parameters, both of which are functions of δ, while their data adaptive calibration is discussed in Section 5.1 below. We let $\lambda = 2\delta'$ and $\mu = 5\|C^{-1}\|_{\infty,1}\delta'$, with

$$\delta' = \left(\frac{8}{\nu}\|C\|_\infty - 3\right)\delta, \tag{4.8}$$

for ν defined in (4.3) above. When ν and $\|C\|_\infty$ are strictly positive constants, we thus have $\lambda = O(\sqrt{\log(p \vee n)/n})$ and $\mu = O(\|C^{-1}\|_{\infty,1}\sqrt{\log(p \vee n)/n})$. We consider the loss function for two $p \times K$ matrices A, A' as

$$L_q(A, A') := \min_{P \in \mathcal{H}_K} \|AP - A'\|_{\infty,q}, \quad 1 \leq q \leq \infty. \tag{4.9}$$

Here, $\mathcal{H}_K$ is the hyperoctahedral group of all $K \times K$ signed permutation matrices and

$$\|A\|_{\infty,q} := \max_{1 \leq i \leq p} \|A_{i\cdot}\|_q = \max_{1 \leq i \leq p} \left(\sum_{j=1}^K |A_{ij}|^q\right)^{1/q},$$

for a generic matrix $A \in \mathbb{R}^{p \times K}$.

THEOREM 4. *Assume the conditions in Theorem 3 hold. Let λ and μ be as defined above, and set $s = \max_{i \in [p]} \|A_{i\cdot}\|_0$. Then*

$$L_q(\widehat{A}, A) \leq 10 s^{1/q} \|C^{-1}\|_{\infty,1} \delta', \quad 1 \leq q \leq \infty,$$

with probability larger than $1 - c_1(n \vee p)^{-c_2}$, for c_1, c_2 positive constants defined in (4.2), provided that $(2\mu + 4\delta/\nu) < 1$. We use the convention that $s^{1/q} = 1$ for $q = +\infty$.

REMARK 4.

1. In fact, we prove the stronger result

$$\min_{P \in \mathcal{H}_K} \|\widehat{A}_{i\cdot} - (AP)_{i\cdot}\|_q \leq 10(s_i)^{1/q} \|C^{-1}\|_{\infty,1} \delta', \quad 1 \leq q \leq \infty,$$

with sparsity index $s_i = \|A_{i\cdot}\|_0$ for each row $A_{i\cdot}$, $i \in [p]$ of A. The signed permutation matrix P that achieves the minimum is determined by the alignment of the pure variables and is the same for each $i \in [p]$.

2. Inspection of the proof of this result quickly reveals that $\|\widehat{A}_{i\cdot}\|_1 \le 1$, for each $i \in [p]$, with high probability, in accordance with our model requirement (i).

3. The size of $\|C^{-1}\|_{\infty,1}$ ranges from the constant $\|C^{-1}\|_\infty$, when all latent factors are independent, to the fully general case of $\|C^{-1}\|_{\infty,1} = O(K)$. In the latter case, the bounds become meaningful when $K < O(\sqrt{n/\log p})$. However, if C^{-1} is sparse, then $\|C^{-1}\|_{\infty,1}$ may be considerably smaller than K. In particular, if Z has a multivariate normal distribution and many factors Z_i are conditionally independent, then $\|C^{-1}\|_{\infty,1}$ is small. We do not make any of these assumptions here, and regardless of the situation, Theorem 4 shows that our estimation procedure adapts automatically to it.

Our primary focus is the bound for $q = +\infty$, as this leads to inference on support recovery of A. More generally, for any $q \ge 1$, it is well understood that the quality of estimating a sparse vector in high-dimensional regression-type models depends on the interplay between its sparsity and the behavior of the appropriate Gram matrix associated with the model, which reduces to $C = \mathbb{E}[ZZ^T]$ in our case. The concept of ℓ_q-sensitivity, introduced by Gautier and Tsybakov (2011), is the most general characterization of this interplay to date. It offers a link between the ℓ_q-norm of sparse vectors β and the ℓ_∞-norm of the product between the Gram matrix and β, uniformly over vectors β of sparsity s, ranging over a collection of cones. Formally, the ℓ_q-sensitivity of the matrix C is defined as

$$\kappa_q(C, s) := \inf_{|S| \le s} \inf_{v \in \mathcal{C}_S} \frac{\|Cv\|_\infty}{\|v\|_q}, \tag{4.10}$$

with $\mathcal{C}_S := \{v \in \mathbb{R}^K : \|v_{\bar{S}}\|_1 \le \|v_S\|_1\}$ and $S \subseteq [K]$ with $|S| \le s$. In our context, that of a square, *invertible* matrix C, the reciprocal of the ℓ_∞-sensitivity $\kappa_\infty(C, s)$ becomes essentially $\|C^{-1}\|_{\infty,1}$ with $[\kappa_\infty(C, K)]^{-1} = \|C^{-1}\|_{\infty,1}$, which indeed links $\|\beta\|_\infty$ to $\|C\beta\|_\infty$. Similarly, the quantities $(2s)^{1/q}\|C^{-1}\|_{\infty,1}$ provide concrete substitutes of the reciprocals of the ℓ_q-sensitivities of C, and all of our rates in Theorem 4 match the lower bounds in Theorem 6, up to a logarithmic factor, and the quantities $\|C^{-1}\|_{\infty,1}$ and $\lambda_1(C)$.

Another possible estimation procedure is the linear program (3.13)–(3.14) with tuning parameter $\lambda' = 3\delta'$. We denote its solution by $\widehat{A}_D$.

THEOREM 5. *Assume the conditions in Theorem 3 hold. Let $\lambda' = 3\delta'$ and set $s = \max_{i\in[p]} \|A_{i\cdot}\|_0$. Then*

$$L_q(\widehat{A}_D, A) \le 6\big[\kappa_q(C, s)\big]^{-1}\delta', \tag{4.11}$$

$$\le 6\big\|C^{-1}\big\|_{\infty,1}(2s)^{1/q}\delta', \quad 1 \le q \le \infty, \tag{4.12}$$

with probability larger than $1 - c_1(n \vee p)^{-c_2}$, for c_1, c_2 positive constants defined in (4.2). We use the convention that $s^{1/q} = 1$ for $q = +\infty$.

As discussed in Section 3.3, we would need to further threshold $\widehat{A}_D$ in order to build the desired clusters. The thresholding level is proportional to $\|\widehat{A}_D - A\|_\infty$, and its practical implementation would require an estimator of $[\kappa_\infty(C, s)]^{-1}$, which cannot be computed. One can however bound $[\kappa_\infty(C, s)]^{-1}$ by $\|C^{-1}\|_{\infty,1}$ as in (4.12), leading to an estimate with the same rate of convergence as that of $\widehat{A}$, given in Theorem 4.

We now show that the rates of convergence in Theorems 4 and 5 are optimal (up to a logarithmic factor in p) in a minimax sense for all estimators over the parameter space

$$\mathcal{A}_s := \Big\{A \in [-1, 1]^{p\times K} : A \text{ satisfies (i) and (ii) and } \max_{1\le i\le p} \|A_{i\cdot}\|_0 \le s\Big\}.$$

For our purpose of establishing a minimax lower bound, it suffices to consider a particular sub-Gaussian distribution of X and a particular covariance matrix C. We choose to take the multivariate Gaussian $N_p(\mathbf{0}, ACA^T + \sigma^2 \boldsymbol{I}_p)$ with $A \in \mathcal{A}_s$, any positive definite C and some constant $\sigma^2 > 0$, satisfying (4.13) below.

THEOREM 6. *Assume* $X \sim N_p(\mathbf{0}, ACA^T + \sigma^2 \boldsymbol{I}_p)$. *Let* $K \geq 2$, $p \geq 2K + 1$, $1 \leq s \leq 4K/5$ *and*

$$s\sqrt{\frac{\sigma^2}{\lambda_1(C)}}\sqrt{\frac{\log(K/s)}{n}} \leq c_1, \tag{4.13}$$

for some constant $c_1 > 0$. *Then, for all* $1 \leq q \leq \infty$,

$$\inf_{\widehat{A}} \sup_{A \in \mathcal{A}_s} \mathbb{P}_A \left\{ L_q(\widehat{A}, A) \geq c_2 s^{1/q} \sqrt{\frac{\sigma^2}{\lambda_1(C)}}\sqrt{\frac{\log(K/s)}{n}} \right\} \geq c_3, \tag{4.14}$$

for some positive constants c_2, c_3 *depending solely on* c_1. *The infimum is taken over all estimators* $\widehat{A}$ *of* A *and we use the convention* $s^{1/q} = 1$ *for* $q = +\infty$.

We attain this bound, up to logarithmic factors, even when I and its partition are not known, for suitable covariance matrices C. Indeed, Theorems 4, 5 and 6 immediately imply that our procedures are not only adaptive in s, but minimax optimal over $A \in \mathcal{A}_s$, up to a logarithmic $\log(K/s)$ and $\log(p \vee n)$, for any covariance matrix C with bounded (constant) ν, $\lambda_1(C)$ and $\|C^{-1}\|_{\infty,1}$. We note that if Z were observed, then an ℓ_0 penalized least squares estimator of A would have an error upper bound containing the factor $\log(K/s)$. From this perspective, the factor $\log(K/s)$ in the lower bound (4.14), derived for unobservable Z, is sharp. The $\log(p)$-term in the upper bound of our estimator stems directly from our choice of δ in (4.1) that controls $\|\widehat{\Sigma} - \Sigma\|_\infty$, for sub-Gaussian distributions, and cannot be dispensed with in our estimation procedure of I and A. Finally, our bounds are established over large classes $\mathcal{A}_s$, without additional assumptions on A, at the expense of placing conditions on C. Even in the classical linear regression model, there is a mismatch—for instance, in terms of largest and smallest eigenvalues of the Gram matrix—between minimax lower bounds for estimating the vector of regression coefficients and achievable upper bounds. Our rates coincide with the minimax rates obtained by Belloni, Rosenbaum and Tsybakov (2017) in the errors in variables context, where just like in our case, the design is not observed.

4.3. *Statistical guarantee for* $\widehat{\mathcal{G}}$ *and* $\widehat{\mathcal{G}}^s$. For easy of presentation, and without loss of generality, throughout this section, we continue to write A for its orthonormal transformation AP that uses the optimal signed permutation matrix $P \in \mathcal{H}_K$ from Theorem 4 to align the columns and signs of A with that of $\widehat{A}$.

We define two criteria to evaluate the estimated clusters $\widehat{\mathcal{G}}$ on the event $\widehat{K} = K$. The latter holds with high probability by Theorem 3. We first define the individual Group False Positive Proportion (GFPP) and the individual Group False Negative Proportion (GFNP) as

$$\mathrm{GFPP}(\widehat{G}_a) := \frac{|(G_a)^c \cap \widehat{G}_a|}{|(G_a)^c|}, \qquad \mathrm{GFNP}(\widehat{G}_a) := \frac{|G_a \cap (\widehat{G}_a)^c|}{|G_a|}, \tag{4.15}$$

for each $a \in [K]$, where $(G_a)^c := [p] \setminus G_a$ and $(\widehat{G}_a)^c := [p] \setminus \widehat{G}_a$, with the convention $\mathrm{GFPP}(\widehat{G}_a) = 0$ if $|(G_a)^c| = \varnothing$. GFPP and GFNP quantify the misclassification proportion within each group $\widehat{G}_a$. Furthermore, with the same convention, we can define the Total False

Positive Proportion (TFPP) and Total False Negative Proportion (TFNP) to quantify the overall misclassification proportion of $\widehat{\mathcal{G}}$:

$$\text{TFPP}(\widehat{\mathcal{G}}) := \frac{\sum_{a=1}^K |(G_a)^c \cap \widehat{G}_a|}{\sum_{a=1}^K |(G_a)^c|}, \qquad \text{TFNP}(\widehat{\mathcal{G}}) := \frac{\sum_{a=1}^K |G_a \cap (\widehat{G}_a)^c|}{\sum_{a=1}^K |G_a|}. \tag{4.16}$$

Finally, given $\mu = 5\|\Omega\|_{\infty,1}\delta'$ with δ' specified in (4.8), we define

$$J_2 := \big\{i \in J : \text{for any } a \text{ with } A_{ia} \neq 0, |A_{ia}| > (2\mu) \vee (4\delta/\nu)\big\} \tag{4.17}$$

and $J_3 := J \setminus (J_1 \cup J_2)$. J_2 can be viewed as the set where every nonzero entry of $A_{j\cdot}$ is separated away from 0 for each $j \in J_2$. The following theorem shows that J_2 plays a critical role in quantifying both the support recovery of $\widehat{A}$ and the misclassification proportion of $\widehat{\mathcal{G}}$. Let $\widehat{S} := \text{supp}(\widehat{A})$.

THEOREM 7. *Under the conditions of Theorem 4, with probability greater than* $1 - c_1(n \vee p)^{-c_2}$ *for some positive constant* c_1 *and* c_2 *defined in* (4.2), *we have*:

(a) $\text{supp}(A_{J_2}) \subseteq \text{supp}(\widehat{A}) \subseteq \text{supp}(A)$, $\text{sign}(\widehat{A}_{\widehat{S}}) = \text{sign}(A_{\widehat{S}})$.

(b) *Let* $s_j^a = 1\{|A_{ja}| \neq 0\}$ *and* $t_j^a = 1\{|A_{ja}| \leq (2\mu) \vee (4\delta/\nu)\}$, *for each* $j \in J$ *and* $a \in [K]$.

$$\text{GFPP}(\widehat{G}_a) = 0; \qquad \text{GFNP}(\widehat{G}_a) \leq \frac{\sum_{j \in J_1 \cup J_3 \setminus J_1^a} t_j^a}{\sum_{j \in J} s_j^a + |I_a|}. \tag{4.18}$$

(c) *Let* $s_j = \sum_{a=1}^K 1\{|A_{ja}| \neq 0\}$ *and* $t_j = \sum_{a=1}^K 1\{|A_{ja}| \leq (2\mu) \vee (4\delta/\nu)\}$, *for each* $j \in J$.

$$\text{TFPP}(\widehat{\mathcal{G}}) = 0; \qquad \text{TFNP}(\widehat{\mathcal{G}}) \leq \frac{\sum_{j \in J_1 \cup J_3} t_j}{\sum_{j \in J} s_j + |I|}. \tag{4.19}$$

REMARK 5.

1. From our proof of Theorem 7, it is easy to verify that the expression of TFNP in (4.19) continues to hold for the Direction False Positive Proportion (DFPP) and the Direction False Negative Proportion (DFNP) defined in (5.2) below with s_j replaced by $\sum_{a=1}^K \mathbf{1}\{A_{ja} < 0\}$ or $\sum_{a=1}^K \mathbf{1}\{A_{ja} > 0\}$, t_j replaced by $\sum_{a=1}^K 1\{-(2\mu) \vee (4\delta/\nu) \leq A_{ja} < 0\}$ or $\sum_{a=1}^K 1\{0 < A_{ja} \leq (2\mu) \vee (4\delta/\nu)\}$ and I replaced by I^+ or I^-, where $I^\pm := \bigcup_{a \in [K]}\{i \in I_a : A_{ia} = \pm 1\}$.
2. According to display (4.18), it is easy to see that $\text{GFNP}(\widehat{G}_a)$ will be small if either t_j^a is small for $j \in J_1 \cup J_3$ or $|J_1| + |J_3| - |J_1^a|$ is dominated by $|I_a| + \sum_{j \in J} s_j^a$. Moreover, from display (4.19), TFNP will be small in the following two cases:

   - $|J_1| + |J_3|$ is dominated by $|I| + |J_2|$;
   - t_j is small relative to s_j, for $j \in J_1 \cup J_3$.

   To illustrate this, consider $t_j \equiv t$ and $s_j \equiv s$, for each $j \in J$, to simplify the expressions a bit, and assume $|J_1| + |J_3| = \alpha(|I| + |J_2|)$, for some $\alpha \geq 0$. We show in the Supplementary Material that

$$\text{TFNP}(\widehat{\mathcal{G}}) \leq t \Big/ \left\{ s + \frac{1}{\alpha}\left(1 + \frac{(s-1)|J_2|}{|I| + |J_2|}\right)\right\},$$

   Thus, when either t or α is small, that is, when $|J_1| + |J_3|$ is dominated by $|I| + |J_2|$, then TFNP will be small. Note that even when t itself is large but bounded by some constant, TFNP might also be small since s can be close to K which is allowed to grow as $O(\sqrt{n/\log p})$.

3. If $J_2 = J$ with $\mu = 3\|C^{-1}\|_{\infty,1}\delta$, from noting that $J_2 \subseteq J \setminus J_1$, Remark 3 in Section 4.1 yields $\widehat{I} = I$. We can choose $\lambda = \delta$ in (3.10) and $\mu = 3\|C^{-1}\|_{\infty,1}\delta$ in (3.12), and follow the proof of Theorems 4 and 7 to arrive at the following conclusions:

$$\operatorname{supp}(\widehat{A}) = \operatorname{supp}(A), \qquad \operatorname{sign}(\widehat{A}) = \operatorname{sign}(A).$$

Moreover, we get exact cluster recovery:

(a) $\operatorname{GFPP}(\widehat{G}_a) = \operatorname{GFNP}(\widehat{G}_a) = 0$, for each $a \in [K]$.
(b) $\operatorname{TFPP}(\widehat{\mathcal{G}}) = \operatorname{TFNP}(\widehat{\mathcal{G}}) = 0$.

This immediately yields $\widehat{G}_0 = G_0$. Again, all statements hold with probability greater than $1 - c_1(n \vee p)^{-c_2}$.

4. We prove that Theorem 7 also holds for the hard threshold estimator $\widetilde{A}$ in which we combine $\widehat{A}_{\widehat{I}}$ with $\widetilde{A}_{\widehat{J}}$. Each row of $\widetilde{A}_{\widehat{J}}$ is estimated by $\widetilde{\beta}_a^j = \bar{\beta}_a^j 1\{|\bar{\beta}_a^j| > \mu\}$ of $\beta_a^j = A_{ja}$, $a \in [\widehat{K}]$, using the same $\mu = 5\|C^{-1}\|_{\infty,1}\delta'$ as before for the threshold μ. However, we cannot guarantee that the scaling restriction of condition (i) holds for this estimator.
5. Theorem 7 holds for the Dantzig-type procedure $\widehat{A}_D$, followed by the hard-threshold procedure described in the above item, using this time the threshold $\mu = 6\|C^{-1}\|_{\infty,1}\delta'$. In this case, the scaling restriction of condition (i) continues to hold as it holds for $\widehat{A}_D$, with high probability.

4.4. *Discussion and related work.* To the best of our knowledge, optimal estimation of identifiable sparse loading matrices A in model (1.1) satisfying (i)–(iii), when both I and K are unknown, and when the entries in X, Z and A are allowed to have arbitrary signs, has not been considered elsewhere and our results bridge this gap. There exists, however, a very large body of literature on related problems. We review the most closely related results below, and explain the differences with our work.

Results regarding the identifiability of A in general latent models, typically not sparse, are scattered throughout over more than six decades of literature. They all involve conditions on both A and C, and there is typically a trade-off between the restrictions on A versus those on C, as first summarized and proved in Anderson and Rubin (1956), reviewed in Lawley and Maxwell (1971) and later in Anderson and Amemiya (1988). We recall them briefly here for the convenience of the reader.

By far, the most commonly used assumption is that the latent factors are uncorrelated, so that C is either the identity or a diagonal matrix. In this case, it is typically further assumed that the scaled columns of A are orthogonal; see, for instance, the literature review in Izenman (2008). An alternative requirement is that A contain a $K \times K$ lower diagonal matrix (see, e.g., Geweke and Zhou (1996)) and, moreover, that the placement of this matrix within A is known, which requires careful justification (Carvalho et al. (2008)), and may be problematic from a practical perspective (Bhattacharya and Dunson (2011)).

In general, latent factors are correlated, which is our point of view in this work. Then, starting with Anderson and Rubin (1956), one places on the structure of A constraints that are different than those made when C is diagonal. The most common of those assumptions involves the existence of a pure variable set I, similar to our assumption (ii). If I is known, classical results in Anderson and Rubin (1956) and the proof of our Theorem 2 show that C can be an arbitrary positive definite matrix. When I is unknown, conditions on the latent factors also need to be imposed. Sufficient conditions on Z, with provable guarantees for the identification of I, are only known, to the best of our knowledge, in the NMF literature: the uniqueness of I follows from the uniqueness of the solution of an appropriate linear program, applied to population quantities, and tailored to matrices with nonnegative entries; see Bittorf et al. (2012). In contrast, the arguments of Section 2 above are optimization-free and can be

used for matrices that have entries of arbitrary sign. Therefore, we provide a new addition to the literature on pure-variable and loading matrix identification in general latent models, and also in the particular case of NMF. We continue this line of reasoning in Bing, Bunea and Wegkamp (2018) that adapts the LOVE procedure to search for the anchor words in the topic model.

A related, but different, identifiability question regards the covariance matrix Σ of X which, under (1.1), can be written as the sum between a rank K matrix and a diagonal matrix:

$$\Sigma = ACA^T + \Gamma, \tag{4.20}$$

and $\Gamma = \mathrm{Cov}(E)$ is a diagonal matrix with possibly different entries. In these models, the identifiability question is whether Σ can be decomposed uniquely as the sum between ACA^T and Γ. Answers to this question generated a large amount of literature. We refer the reader to Anderson and Rubin (1956), Bekker and ten Berge (1997), Ledermann (1937), Shapiro (1982), Shapiro (1985) for earlier results, and to Bai and Ng (2002), Candès et al. (2011), Chandrasekaran, Parrilo and Willsky (2012), Chandrasekaran et al. (2011), Hsu, Kakade and Zhang (2011), Fan, Liao and Mincheva (2013), Wegkamp and Zhao (2016) for more recent works that also address the problems of rank estimation and optimal estimation of high-dimensional covariance matrices. It is noteworthy that these works, relative to one another, give different types of sufficient conditions under which one can separate the low rank matrix ACA^T from Γ. However, since we always have $ACA^T = (AQ)(Q^TCQ)(Q^TA^T)$, for any orthonormal Q, they do not guarantee the identifiability of A itself. Conversely, we show in Theorem 2 in Section 2 that under conditions (i)–(iii), C and Γ are identified, and A is identified up to signed permutations. Therefore, we also identify uniquely the decomposition of Σ. Our conditions are not always comparable to those employed for the unique decomposition of Σ, but in special cases they imply them. Although the uniqueness of the decomposition of Σ is a by-product of our results, we do not pursue the covariance estimation problem in this work, but we included the above discussion for completeness.

Furthermore, we do not view the problem of estimating the number of factors K as that of estimating the rank of a matrix. This approach is taken in Bai and Ng (2002), via penalized least squares, but provided that either $C = I$ or $AA^T = I$ and that K is bounded by a fixed integer. Alternatively, we could adapt the criteria in Bing and Wegkamp (2018), Bunea, She and Wegkamp (2011), Wegkamp and Zhao (2016) to (1.1) to allow for $K \to \infty$ in the rank estimation problem. However, proving that such an estimator is consistent would ultimately require an unnecessary lower bound restriction on the Kth largest eigenvalue of ACA^T. In contrast, our Theorem 3 shows that such conditions can indeed be avoided. We estimate directly the set I and its partition via LOVE, and as a byproduct K, at a low computational cost of order p^2.

Estimation of A in identifiable factor models is typically based on iterative alternating least squares procedures or the EM algorithm; see, for instance, Bai and Li (2012), Rubin and Thayer (1982) and the references therein. As discussed in these works, the resulting algorithms are not suitable for large data sets due to their notoriously slow convergence to a solution that is typically not the global optimum. Bayesian estimation (see, e.g., Carvalho et al. (2008) and the references therein) offers an alternative approach which may become computationally very demanding in high dimensions, requires a likelihood framework and careful prior specification. Moreover, existing procedures do not estimate A under our model specifications (i)–(iii), and any adaptation would still require the challenging estimation of I. Our procedure offers a solution to the computational problem, as LOVE does not require a likelihood or other prior distributional specifications, is tailored to our model with unknown I, and has provable low computational complexity.

The statistical properties of estimators of A in model (1.1) (i)–(iii) have not been studied, and even particular cases of the model have received a very limited amount of attention, from a theoretical perspective. When I is known and K is fixed, Bai and Li (2012) established the asymptotic normality of the MLE in a model similar to ours, although the estimator they ultimately construct is not necessarily the MLE under this model, but rather an appropriate transformation of the stationary point of a quasi-likelihood for a different factor model. We give the specific details of their construction in Section C.1 of the Supplementary Material. If I is unknown, but K is known, and moreover, the columns of $\boldsymbol{X}$, A and $\boldsymbol{Z}$ have nonnegative entries that sum up to 1, Arora et al. (2013) provide a practical algorithm for the estimation of A and offer bounds on the ℓ_1 matrix norm loss of their estimator. The extra restrictions on this model are motivated by a specific model, the topic model, appropriate for vectors with discrete distributions, for instance multinomial. The construction and analysis of these estimates are not transferable to our general framework, as they depend heavily on these restrictions. Our results of Section 4.2 bridge this gap in the literature and offer lower and upper bounds for the performance of estimators of A in model (1.1)(i)–(iii).

Finally, to the best of our knowledge, overlapping clustering based on model (1.1) has not been analyzed. A particular case of this model, corresponding to a matrix A with binary entries, has been considered in Bunea, Giraud and Luo (2016), Bunea et al. (2016) for nonoverlapping clustering. According to their model, all p variables are pure variables, as the model assume that $X_j = Z_k + E_j$, for all $j \in G_k$ and $k \in \{1, \dots, K\}$, $\{G_k\}_{1 \le k \le K}$ form a partition of $\{1, \dots, p\}$. When C is positive definite, the nonoverlapping clusters are shown to be identifiable, and the work of Bunea, Giraud and Luo (2016), Bunea et al. (2016) is devoted to exact recovery of clusters with minimax optimal cluster separation, a very different problem than the one considered here.

5. Simulation studies. In this section, we first discuss our procedure for selecting the tuning parameters, then evaluate the performance of LOVE based on estimation error and overall clustering misclassification proportion. In the Supplementary Material, we compare LOVE with existing overlapping clustering algorithms and study the performance of LOVE for the non-overlapping clustering problem.

5.1. *Data driven choice of the tuning parameters.*

5.1.1. *Tuning parameter δ.* Proposition 3 specifies the theoretical rate of δ, but only up to constants that depend on the underlying data generating mechanism. We propose below a data-dependent way to select δ, based on data splitting. Specifically, we split the data set into two independent parts, of equal sizes. On the first set, we calculate the sample covariance matrix $\widehat{\Sigma}^{(1)}$. On the second set, we choose a fine grid of values $\delta_\ell = c_\ell \sqrt{\log p / n}$, with $1 \le \ell \le M$, for δ, by varying the proportionality constants c_ℓ. For each δ_ℓ, we obtain the estimated number of clusters $\widehat{K}(\ell)$ and the pure variable set $\widehat{I}(\ell)$ with its partition $\widehat{\mathcal{I}}(\ell)$. Then we construct the $|\widehat{I}(\ell)| \times \widehat{K}(\ell)$ submatrix $\widehat{A}_{\widehat{I}(\ell)}$ of $\widehat{A}$, and estimate $\widehat{C}(\ell)$ via formula (3.6). Finally, we calculate the $|\widehat{I}(\ell)| \times |\widehat{I}(\ell)|$ matrix $W_\ell = \widehat{A}_{\widehat{I}(\ell)} \widehat{C}(\ell) \widehat{A}^T_{\widehat{I}(\ell)}$. In the end, we have constructed a family $\mathcal{F} = \{W_1, \dots, W_M\}$ of the fitted matrices W_ℓ, each corresponding to different $\widehat{\mathcal{I}}(\ell)$ that depend in turn on δ_ℓ, for $\ell \in \{1, \dots, M\}$. Define

$$\mathrm{CV}(\widehat{\mathcal{I}}(\ell)) := \frac{1}{\sqrt{|\widehat{I}(\ell)|(|\widehat{I}(\ell)| - 1)}} \big\| \widehat{\Sigma}^{(1)}_{\widehat{I}(\ell)\widehat{I}(\ell)} - W_\ell \big\|_{\text{F-off}}, \tag{5.1}$$

where $\|B\|_{\text{F-off}} := \|B - \mathrm{diag}(B)\|_F$ denotes the Frobenius norm over the off-diagonal elements of a square matrix B. We choose δ^{cv} as the value δ_ℓ that minimizes $\mathrm{CV}(\widehat{\mathcal{I}}(\ell))$ over the grid $\ell \in [M]$. To illustrate how the selection procedure works, we provide an example in Section B of the Supplementary Material.

5.1.2. *Tuning parameters λ and μ.* The tuning parameter λ in the linear program (3.10) for estimating $\Omega = C^{-1}$ is specified by $\lambda = 2\delta'$ with δ' defined in (4.8). Since δ' is proportional to δ, we use $\lambda = c_0\delta^{\mathrm{cv}}$ where c_0 is some constant and could be tuned by a cross-validation strategy used in the related work on the precision matrix estimation, for instance, Cai, Liu and Luo (2011). More precisely, we randomly split the data into two parts. For a given grid of λ, we compute $\widehat{\Omega}$ on the first dataset for each value in the grid. Then we choose the one which gives the smallest likelihood loss from the second dataset, where the likelihood loss is defined by

$$L(\Omega, C) = \langle \Omega, C\rangle - \log\det(\Omega).$$

From Remark 5 (3) in Section 4.3, when $J_2 = J$, we can choose $\lambda = \delta$ which is the smallest λ we should consider. Therefore, we set the grid of λ equal to $[\delta^{\mathrm{cv}}, 3\delta^{\mathrm{cv}}]$. From our simulation, the selected λ is δ^{cv} in most cases. Hence we recommend to use $\lambda = \delta^{\mathrm{cv}}$ and our simulations are based on this choice.

Recall that $\mu = c_1\|C^{-1}\|_{\infty,1}\delta$ for some constant c_1, and that $\widehat{\Omega}$ estimates C^{-1}. Our extensive simulations show that the choice of $\mu = \|\widehat{\Omega}\|_{\infty,1}\delta^{\mathrm{cv}}$ yields stable performance, with $\widehat{\Omega}$ solved from (3.9) and δ^{cv} selected via cross-validation.

5.2. *Estimation error and cluster recovery with LOVE.* In this section, we study the numerical performance of LOVE in terms of clustering and estimation accuracy. To the best of our knowledge, there is no comparable algorithm with provable guarantees developed for our framework, especially if the set I is unknown, as explained in detail in Section 4.4 above, and further revisited in Section C.1 of the Supplementary Material.

We generate the data in the following way. We set the number of clusters K to be 20 and simulate the latent variables $Z = (Z_1, \ldots, Z_K)$ from $N(0, C)$. The diagonal elements of C is given by $C_{ii} = 2 + (i-1)/19$ for $i = 1, \ldots, 20$, and the off-diagonal elements are generated as $C_{ij} = (-1)^{(i+j)}0.3^{|i-j|}(C_{ii} \wedge C_{jj})$ for any $i \neq j$. In addition, the error terms $E_1, \ldots, E_p$ are independently sampled from $N(0, \sigma_p^2)$, where σ_p^2 itself is sampled from a uniform distribution on $[1, 3]$. Since the rows of A corresponding to pure variables in the same cluster are allowed to have different signs, we consider the following configuration of signs for pure variables in each cluster: $(3, 2)$, $(4, 1)$, $(2, 3)$, $(1, 4)$ and $(5, 0)$, with the convention that the first number denotes the number of positive pure variables in that group and the second one denotes the number of negative pure variables. Among the 20 groups, each sign pattern is repeated 4 times. To generate A_J, for any $j \in J$, we randomly assign the cardinality s_j of the support of $A_{j\cdot}$ to a number in $\{2, 3, 4, 5\}$, with equal probability. Then we randomly select the support from $\{1, 2, \ldots, K\}$ with cardinality equal to s_j. For A_{jk} which is nonzero, we set it as $A_{jk} = \mathrm{sign}\cdot(1/s_j)$ with sign randomly sampled from $\{-1, 1\}$. Thus, we can generate X according to the model $X = AZ + E$. In the simulation studies, we vary p from 200 to 1000 and n from 300 to 1000. Each simulation is repeated 50 times.

Recall that the true allocation matrix A and our estimator $\widehat{A}$ are not directly comparable, since they may differ by a permutation matrix. To evaluate the performance of our method, we consider the following mapping approach (Wiwie, Baumbach and Röttger (2015)). If A and $\widehat{A}$ have the same dimension, we first find the mapping (i.e., the signed permutation matrix $P \in \mathcal{H}_K$) such that $\|A - \widehat{A}P\|_F$ is minimized. Thus, we can compare the permuted estimator $\widetilde{A} = \widehat{A}P$ with A to evaluate the estimation and recovery error. Under this mapping approach, we can evaluate TFPP and TFNP defined in (4.16). Moreover, in order to account for the direction subgroups defined in (3.18), we can define Direction False Positive Proportion (DFPP) and Direction False Negative Proportion (DFNP) as follows:

$$\mathrm{DFPP} = \frac{\sum_{a=1}^K |G_a^1 \cap \widehat{G}_a^2|}{\sum_{a=1}^K |G_a^1|}, \qquad \mathrm{DFNP} = \frac{\sum_{a=1}^K |G_a^2 \cap \widehat{G}_a^1|}{\sum_{a=1}^K |G_a^2|}. \tag{5.2}$$

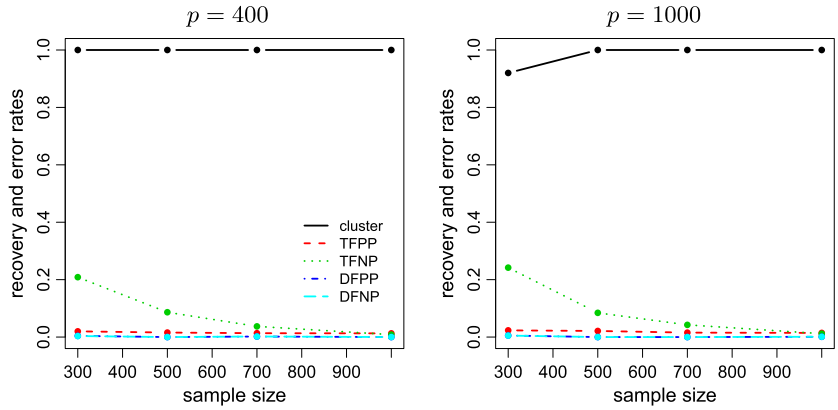

FIG. 1. *Percentage of exact recovery of number of clusters K (cluster), total false positive proportion (TFPP), total false negative proportion (TFNP), direction false positive proportion (DFPP) and direction false negative proportion (DFNP) for LOVE.*

Figure 1 shows the percentage of exact recovery of number of clusters K, TFPP, TFNP, DFPP and DFNP of LOVE. Since the last four measures are well-defined only if $\operatorname{rank}(\widehat{A}) = K$, we can compute them when the number of clusters is correctly identified. We can see that the proposed method correctly selects K and as long as the number of clusters is correctly selected, TFPP, TFNP, DFPP and DFNP of our method are very close to 0, which implies that the sign and sparsity pattern of A can be correctly recovered. We present the estimation error of $\widehat{A}$ as measured by the matrix ℓ_1 norm scaled by pK and the Frobenius norm scaled by $\sqrt{pK}$ in Table 2.

As expected, the estimation error decreases when the sample size increases from 300 to 1000, which is in line with our theoretical results. The simulations are conducted on a macOS Sierra system version 10.12.6 with 2.2 GHz Intel Core i7 CPU and 16 GB memory. Even with $p = 1000$ and $n = 1000$, the computing time of our method for each simulation is around 1 minute.

Moreover, we evaluated the performance of the LOVE procedure for K varying in a wide range, from 3 to 30, and when A_J contains many very small entries. The results are consistent with what we observed in this section and deliver the same message. The GFPP and GFNP are similar as TFPP and TFNP and the performance of the hard thresholding estimator $\widetilde{A}$,

TABLE 2

The average estimation error of $\widehat{A}$ as measured by the matrix ℓ_1 norm (ℓ_1) (divided by pK) and the Frobenius norm (ℓ_2) (divided by $\sqrt{pK}$). Numbers in parentheses are the simulation standard errors

	$n = 300$		$n = 500$		$n = 700$		$n = 1000$	
p	ℓ_1	ℓ_2	ℓ_1	ℓ_2	ℓ_1	ℓ_2	ℓ_1	ℓ_2
200	0.018	0.062	0.015	0.053	0.013	0.048	0.012	0.041
	(0.001)	(0.005)	(0.001)	(0.003)	(0.001)	(0.008)	(0.001)	(0.002)
400	0.026	0.075	0.023	0.064	0.021	0.059	0.018	0.051
	(0.002)	(0.007)	(0.001)	(0.003)	(0.001)	(0.006)	(0.001)	(0.003)
600	0.029	0.079	0.025	0.067	0.023	0.063	0.020	0.055
	(0.002)	(0.006)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)
800	0.031	0.083	0.026	0.068	0.024	0.064	0.022	0.057
	(0.002)	(0.006)	(0.001)	(0.004)	(0.001)	(0.004)	(0.001)	(0.004)
1000	0.032	0.083	0.027	0.069	0.025	0.065	0.022	0.057
	(0.002)	(0.006)	(0.001)	(0.003)	(0.001)	(0.004)	(0.001)	(0.004)

TABLE 3
Number of pure genes and total number of genes in each group

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
Number of pure genes	2	2	2	4	2	10	2	2	2	4	2	15
Total number of genes	58	35	67	105	80	104	28	43	44	74	94	108

defined in Remark 5 of Section 4.3, is similar to $\widehat{A}$. To save space, we have omitted those results.

We also compared the performance of LOVE with other off-the-shelf algorithms for overlapping clustering, and tested LOVE for nonoverlapping clustering. We included these results in Sections C.2 and C.3 of the Supplementary Material.

6. Application. To benchmark LOVE, we used a publicly available RNA-seq dataset of 285 blood platelet samples from patients with different malignant tumors (Best et al. (2015)). We extracted a small subset of 500 Ensembl genes to test the method. The goal of the benchmarking was to test whether (i) clusters corresponded to biological knowledge, specifically Gene Ontology (GO) functional annotation of the genes (Ashburner et al. (2000)), (ii) overlapping clusters corresponded to pleiotropic gene function. LOVE produced twelve overlapping clusters (Table 3) which aligned well with a priori expectation. Table 3 lists the number of pure genes and the total number of genes in twelve overlapping clusters. Figure 2 shows that each cluster overlaps with the other and also gives us a clear picture on how two clusters possibly overlap. For example, 18 genes belong to both cluster 3 and cluster 11, whereas cluster 2 and cluster 3 have only one common gene. The genes with the same GO biological process, molecular function or cellular component terms tended to be assigned to the same cluster. For example, ENSG00000273906 and ENSG00000273328 are both RNA genes. They were both assigned to the same cluster (cluster 6, Figure 2). However, they were also assigned to other clusters, suggesting they have pleiotropic functions. This suggests that the latent variables used for clustering are likely to have biological significance and can potentially be used for functional discovery for genes with underexplored functions.

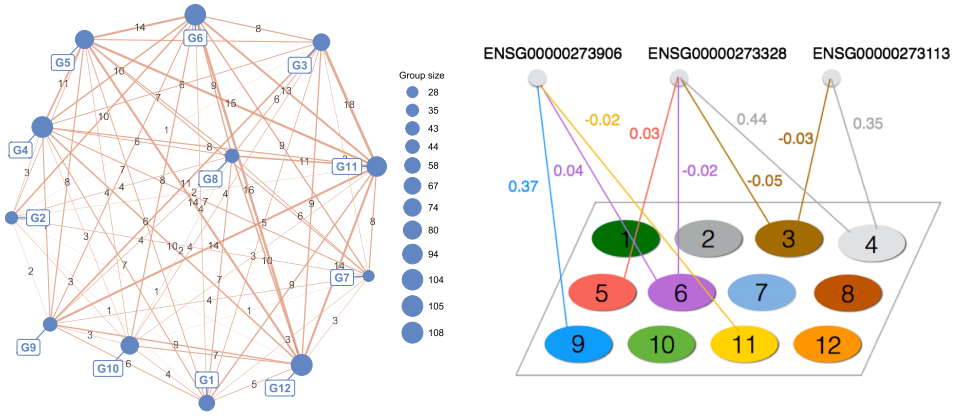

FIG. 2. *Left panel*: *Number of genes overlapped in different groups. The nodes represent* 12 *groups with the same labels and sizes as those in Table* 3. *The number shown on the edge between two nodes represents the number of genes shared by the two groups, which corresponds to the width of that edge. Right panel*: *Illustration of three genes ENSG*00000273906, *ENSG*00000273328 *and ENSG*00000273113 *and their allocation matrix relative to* 12 *groups. For instance, the* j*th gene ENSG*00000273906 *belongs to groups* 6, 9 *and* 11 *with* $\widehat{A}_{j6} = 0.04$, $\widehat{A}_{j9} = 0.37$, $\widehat{A}_{j11} = -0.02$.

We found 308 genes with zero expression across all samples. None of them were assigned to any of the 12 estimated clusters, as desired. Indeed, our model not only allows for the existence of pure noise variables $X_j = E_j$, but variables with structural zero values as well, as $\Gamma_{jj} = \mathrm{Var}(E_j) = 0$ is permitted. Formally, we place them in the pure noise cluster G_0, for further scientific scrutiny.

Acknowledgments. We thank the referees for their many insightful and helpful suggestions. We are grateful to Jishnu Das for help with the interpretation of our data analysis results.

The first author was supported in part by NSF Grant DMS-1407600.

The second and fourth authors were supported in part by NSF Grant DMS-1712709.

The third author was supported in part by NSF Grant DMS-1854637.

SUPPLEMENTARY MATERIAL

Supplement to "Adaptive estimation in structured factor models with applications to overlapping clustering" (DOI: 10.1214/19-AOS1877SUPP; .pdf). The supplementary document includes the proofs and additional numerical results.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1990662

ANDERSON, T. W. and AMEMIYA, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.* **16** 759–771. MR0947576 https://doi.org/10.1214/aos/1176350834

ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. V* 111–150. Univ. California Press, Berkeley and Los Angeles. MR0084943

ARORA, S., GE, R., HALPERN, Y., MIMNO, D. M., MOITRA, A., SONTAG, D., WU, Y. and ZHU, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *ICML* (2) 280–288.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25** 25–29.

BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313 https://doi.org/10.1214/11-AOS966

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259 https://doi.org/10.1111/1468-0262.00273

BEKKER, P. A. and TEN BERGE, J. M. F. (1997). Generic global indentification in factor analysis. *Linear Algebra Appl.* **264** 255–263. MR1465870 https://doi.org/10.1016/S0024-3795(96)00363-1

BELLONI, A., ROSENBAUM, M. and TSYBAKOV, A. B. (2017). Linear and conic programming estimators in high dimensional errors-in-variables models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 939–956. MR3641415 https://doi.org/10.1111/rssb.12196

BEST, M. G., SOL, N., KOOI, I., TANNOUS, J., WESTERMAN, B. A., RUSTENBURG, F., SCHELLEN, P., VERSCHUEREN, H., POST, E. et al. (2015). RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* **28** 666–676.

BEZDEK, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York. With a foreword by L. A. Zadeh, Advanced Applications in Pattern Recognition. MR0631231

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429 https://doi.org/10.1093/biomet/asr013

BIEN, J., BUNEA, F. and XIAO, L. (2016). Convex banding of the covariance matrix. *J. Amer. Statist. Assoc.* **111** 834–845. MR3538709 https://doi.org/10.1080/01621459.2015.1058265

BING, X., BUNEA, F. and WEGKAMP, M. H. (2018). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. ArXiv E-prints arXiv:1805.06837.

BING, X. and WEGKAMP, M. H. (2018). Adaptive estimation of the rank of the coefficient matrix in high dimensional multivariate response regression models. arXiv:1704.02381.

BING, X., BUNEA, F., NING, Y. and WEGKAMP, M. (2020). Supplement to "Adaptive estimation in structured factor models with applications to overlapping clustering." https://doi.org/10.1214/19-AOS1877SUPP.

BITTORF, V., RECHT, B., RE, C. and TROPP, J. A. (2012). Factoring nonnegative matrices with linear programs. arXiv:1206.1270.

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. A Wiley-Interscience Publication. MR0996025 https://doi.org/10.1002/9781118619179

BUNEA, F., GIRAUD, C. and LUO, X. (2016). Minimax optimal variable clustering in G-models via cord. ArXiv Preprint arXiv:1508.01939.

BUNEA, F., SHE, Y. and WEGKAMP, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355 https://doi.org/10.1214/11-AOS876

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

BUNEA, F., GIRAUD, C., LUO, X., ROYER, M. and VERZELEN, N. (2018). Model assisted variable clustering: Minimax-optimal recovery and algorithms. ArXiv E-prints arXiv:1508.01939. *Ann. Statist.* To appear.

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973 https://doi.org/10.1198/jasa.2011.tm10155

CAI, T. T., LIU, W. and ZHOU, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44** 455–488. MR3476606 https://doi.org/10.1214/13-AOS1171

CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000 https://doi.org/10.1145/1970392.1970395

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722 https://doi.org/10.1198/016214508000000869

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067 https://doi.org/10.1214/11-AOS949

CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. MR2817479 https://doi.org/10.1137/090761793

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.

CRADDOCK, R. C., JBABDI, S., YAN, C.-G., VOGELSTEIN, J. T., CASTELLANOS, F. X., DI MARTINO, A., KELLY, C., HEBERLEIN, K., COLCOMBE, S. et al. (2013). Imaging human connectomes at the macroscale. *Nat. Methods* **10** 524–539.

DONOHO, D. and STODDEN, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems* 16 (S. Thrun, L. K. Saul and P. B. Schölkopf, eds.) 1141–1148. MIT Press.

EVERITT, B. S. (1984). *An Introduction to Latent Variable Models*. *Monographs on Statistics and Applied Probability*. CRC Press, London; distributed by Methuen, Inc., New York. MR0769300 https://doi.org/10.1007/978-94-009-5564-6

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva. MR3091653 https://doi.org/10.1111/rssb.12016

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GAUTIER, E. and TSYBAKOV, A. B. (2011). High-dimensional instrumental variables regression and confidence sets. ArXiv Preprint arXiv:1105.2454v4.

GEWEKE, J. and ZHOU, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Rev. Financ. Stud.* **9** 557–587.

HSU, D., KAKADE, S. M. and ZHANG, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inform. Theory* **57** 7221–7234. MR2883652 https://doi.org/10.1109/TIT.2011.2158250

IZENMAN, A. J. (2008). *Modern Multivariate Statistical Techniques*: *Regression*, *Classification*, *and Manifold Learning*. *Springer Texts in Statistics*. Springer, New York. MR2445017 https://doi.org/10.1007/978-0-387-78189-1

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

KOOPMANS, T. C. and REIERSØL, O. (1950). The identification of structural characteristics. *Ann. Math. Stat.* **21** 165–181. MR0039967 https://doi.org/10.1214/aoms/1177729837

KRISHNAPURAM, R., JOSHI, A., NASRAOUI, O. and YI, L. (2001). Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE Trans. Fuzzy Syst.* **9** 595–607.

LAWLEY, D. N. and MAXWELL, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. American Elsevier Publishing Co., Inc., New York. MR0343471

LEDERMANN, W. (1937). On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika* **2** 85–93.

MCDONALD, R. P. (1999). *Test Theory*: *A Unified Treatment*. Taylor and Francis.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann*. *Statist*. **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

RUBIN, D. B. and THAYER, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** 69–76. MR0668505 https://doi.org/10.1007/BF02293851

SHAPIRO, A. (1982). Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika* **47** 187–199. MR0667012 https://doi.org/10.1007/BF02296274

SHAPIRO, A. (1985). Identifiability of factor analysis: Some results and open problems. *Linear Algebra Appl*. **70** 1–7. MR0808527 https://doi.org/10.1016/0024-3795(85)90038-2

WEGKAMP, M. and ZHAO, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* **22** 1184–1226. MR3449812 https://doi.org/10.3150/14-BEJ690

WIWIE, C., BAUMBACH, J. and RÖTTGER, R. (2015). Comparing the performance of biomedical clustering methods. *Nat*. *Methods* **12** 1033–1038.

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824 https://doi.org/10.1093/biomet/asm018