Detecting approximate replicate components of a high-dimensional random vector with latent structure

XIN BING\textsuperscript{1,*} FLORENTINA BUNEA\textsuperscript{1,**} and MARTEN WEGKAMP\textsuperscript{1,2}

\textsuperscript{1}Department of Statistics and Data Science, Cornell University, Ithaca, New York, USA.
E-mail: *xb43@cornell.edu; **fb238@cornell.edu

\textsuperscript{2}Department of Mathematics, Cornell University, Ithaca, New York, USA.
E-mail: mhw73@cornell.edu

High-dimensional feature vectors are likely to contain sets of measurements that are approximate replicates of one another. In complex applications, or automated data collection, these feature sets are not known a priori, and need to be determined.

This work proposes a class of latent factor models on the observed, high-dimensional, random vector $X \in \mathbb{R}^p$, for defining, identifying and estimating the index set of its approximately replicate components. The model class is parametrized by a $p \times K$ loading matrix $A$ that contains a hidden sub-matrix whose rows can be partitioned into groups of parallel vectors. Under this model class, a set of approximate replicate components of $X$ corresponds to a set of parallel rows in $A$; these entries of $X$ are, up to scale and additive error, the same linear combination of the $K$ latent factors; the value of $K$ is itself unknown.

The problem of finding approximate replicates in $X$ reduces to identifying, and estimating, the location of the hidden sub-matrix within $A$, and of the partition $H$ of its row index set $H$. Both $H$ and $H$ can be fully characterized in terms of a new family of criteria based on the correlation matrix of $X$, and their identifiability, as well as that of the unknown latent dimension $K$, are obtained as consequences. The constructive nature of the identifiability arguments enables computationally efficient procedures, with consistency guarantees.

Furthermore, when the loading matrix $A$ has a particular sparse structure, provided by the errors-in-variable parametrization, the difficulty of the problem is elevated. The task becomes that of separating out groups of parallel rows that are proportional to canonical basis vectors from other, possibly dense, parallel rows in $A$. This is met under a scale assumption, via a principled way of selecting the target row indices, guided by the successive maximization of Schur complements of appropriate covariance matrices. The resulting procedure is an enhanced version of that developed for recovering general parallel rows in $A$. It is also computationally efficient, consistent. It has immediate applications to latent space overlapping clustering and the estimation of loading matrices that satisfy a canonical parametrization.

Primary 62H12, 62J07.

Keywords: high-dimensional statistics, identification, latent factor model, matrix factorization, replicate measurements, pure variables, overlapping clustering.
1. Introduction

Latent factor models are simple, ubiquitous, tools for describing data generating mechanisms that yield random vectors \( X \in \mathbb{R}^p \) with possibly very correlated entries, and subsequently approximately low-rank covariance matrix. The history of factor analysis can be traced back to the 1940s [36, 37, 38, 30, 31, 32, 33], with foundational work established by [5], and a wealth of recent works motivated by applications to economy and finance, educational testing and psychology, forecasting, biology, to give a limited number of examples.

A factor model assumes the existence of an integer \( 1 \leq K \leq p \) and of a random vector \( Z \in \mathbb{R}^K \) of unobservable, latent, factors, such that the observed \( X \in \mathbb{R}^p \) has the representation

\[
X = AZ + E,
\]

for some real-valued loading matrix \( A \in \mathbb{R}^{p \times K} \) and random noise \( E \in \mathbb{R}^p \) uncorrelated with \( Z \). The corresponding covariance matrix of \( X \) has the expression \( \Sigma = A \Sigma_Z A^\top + \Sigma_E \), where \( \Sigma_Z \) is the covariance matrix of \( Z \) and \( \Sigma_E \) that of \( E \). A large amount of literature has been and continues to be devoted to the estimation of approximately low-rank covariance matrices corresponding to factor models, for instance, [16, 26, 15, 1, 21, 22, 23, 24, 19], to name a few.

A related, but different, line of research is devoted to the estimation of the loading matrix \( A \) itself, in identifiable factor models that place structure on \( A \) [5, 27, 18, 7, 22, 3, 2, 6, 4, 23, 28, 8, 9, 10, 11].

We treat a problem of an intermediate nature in this work: we also model the dependency between the components of a high-dimensional random vector \( X \in \mathbb{R}^p \) via latent factors, as in the covariance estimation literature, and place structure on \( A \), as in the literature devoted to loading matrix estimation, but our focus is different. We study factor models relative to matrices \( A \) that are allowed to have groups of parallel rows, a structure that in general is not sufficient for identifying \( A \). The indices of rows that are respectively parallel form a partition \( \mathcal{H} \) of the collection \( \mathcal{H} \) of all indices of parallel rows in \( A \). The components \( X_j \) of \( X \), with \( j \) in a group of \( \mathcal{H} \) are, up to scale and additive error, the same linear combination of the background latent factors. This parametrization of \( A \) thus provides a way of modeling those components of \( X \) that are very highly dependent, in that they are “approximate replicates” of each other, while allowing for general dependency between the other entries in \( X \), albeit modeled via a factor model.

Very high-dimensional random vectors do typically have entries that are approximately redundant, and we give only a few motivating examples from biology. For instance, in human systems immunology, in addition to measuring serological features (cytokines, chemokines, antibody titers etc) that are highly correlated with each other, one often measures the same feature under slightly different technical conditions (e.g. same titer measured at different serum dilutions, same cytokines measured using different technical platforms). In genetic perturbation experiments, one could quantify the effect of the exact same genetic perturbation using different reporter assays. Although some of these redundancies may be obvious based on the experimental design (technical redundancies),
Detecting approximate replicate components

others are known to be induced by latent, underlying biological mechanisms, but it is unknown which of the collected measurements reflect them. We address the latter problem in this work.

We study factor models on high-dimensional vectors $X \in \mathbb{R}^p$ that contain approximate replicate components, in unknown positions. The focus of our work is in determining their locations, a problem that reduces to that of identifying and estimating the location of a hidden sub-matrix of $A$, with unknown row index set $H$, and of unknown partition $\mathcal{H}$. Since the entirety of a matrix $A$ with such structure is typically not identifiable, we also study an instance of it, provided by an added sparsity constraint, under which both a hidden sparse sub-matrix of $A$, and $A$ itself are identifiable.

The following section provides a detailed summary of our approach and results.

1.1. Our contributions

To state our results, we will assume that $X$ follows a factor model (1.1) with $A \in \mathbb{R}^{p \times K}$ and $\text{rank}(A) = K$,

$$\mathbb{E}[E] = 0 \quad \text{and} \quad \Sigma_E = \mathbb{E}[EE^\top] = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2) > 0$$

and $Z$ is standardized such that

$$\mathbb{E}[Z] = 0 \quad \text{and} \quad \text{diag}(\Sigma_Z) = 1 \quad \text{with} \quad \Sigma_Z := \mathbb{E}[ZZ^\top] \quad \text{and} \quad \text{rank}(\Sigma_Z) = K.$$

We make the following assumption, which we will show later identifies $K$. It is based on the intuition that for $p \gg K$, the matrix $A$ is expected to contain many parallel rows.

**Assumption 1.** The index set of parallel rows

$$H := \{i \in \{1, \ldots, p\} : A_i \perp A_j, \text{ for some } j \in \{1, \ldots, p\} \setminus \{i\}$$

of $A$ is non-empty and $\text{rank}(A_H) = K$.

The partition of $H$ consists of disjoint sets $H_1, \ldots, H_G$, and all indices $i, j \in H_k$ correspond to parallel rows $A_i$ and $A_j$, for each $k \in [G] := \{1, 2, \ldots, G\}$. The assumption $H \neq \emptyset$ and $\text{rank}(A_{H_k}) = K$ implies that $G \geq K$, and $|H_k| \geq 2$ for each $k \in [G]$. While we are not aware of a study of factor models under Assumption 1, we mention that it is a generalization of the errors-in-variables parametrization, see for instance [45], which we discuss in more detail below.

With the parametrization of $A$ provided by Assumption 1, we define approximate replicate measurements relative to the parallel rows in $A$ as the groups of variables $X_j$, with $j \in H_k$ and $k \in [G]$; they are, up to scale, the same linear combination of the $K$ background factors, up to the additive measurement error term $E_j$. The problem of detecting approximate replicate features $X_j$ reduces to that of recovering the parallel rows of $A$, and their partition.

In the context of this problem and modelling assumptions, we give below the organization of the paper, section by section, and summarize our contributions.
1.1.1. Section 2: Identification and recovery of approximately replicate features

The results of our Section 2 can be summarized as follows.

1. A new score function for an if and only if characterization of the partition of parallel rows of a loading matrix. We show in Section 2.1 that Assumption 1 is sufficient for the unique, and constructive, determination of $H$, its partition $H = \{H_1, \ldots, H_G\}$, and $K$ from the correlation matrix $R = [\text{Corr}(X_i, X_j)]_{i,j \in [p]}$. In Proposition 1, we prove the non-trivial fact that the parallelism between rows of $A$ is preserved by appropriately modified rows in $R$. To be precise, we show that the vectors $R_{i \setminus \{i,j\}}$ and $R_{j \setminus \{i,j\}}$, each defined by leaving out, respectively, the $i$th and $j$th entries from the rows $R_i$ and $R_j$, are parallel in $R_{p-2}$ if and only if $A_{i \setminus \{i,j\}}$ and $A_{j \setminus \{i,j\}}$ are parallel. This realization, combined with the fact that two non-zero vectors $v, w$ are parallel if and only if

\[
\min_{\|v\|, \|w\| = 1} \|\alpha v + \beta w\|_q = 0
\]

for any $0 < r \leq \infty$ and $0 \leq q \leq \infty$, naturally leads to defining the class of criteria

\[
S_{q,r}(i,j) = (p-2)^{-1/q} \min_{\|\alpha, \beta\|_r = 1} \|\alpha R_{i \setminus \{i,j\}} + \beta R_{j \setminus \{i,j\}}\|_q,
\]

for any $i, j \in [p]$, (1.2) for identifying parallel vectors in $A$ via the model-independent correlation matrix $R$.

In Proposition 2, we establish the following characterization of both $H$ and its partition $\mathcal{H}$, \(S_{q,r}(i,j) = 0 \iff i, j \in H_a \quad \text{for some} \quad a \in [G] \) (1.3) for any $0 < r \leq \infty$ and $0 \leq q \leq \infty$. The new criteria (1.2) are indexed by two parameters $(q, r)$. Proposition 2 shows that $r = \infty$ is optimal. In practice, we prefer $(q, r) = (2, \infty)$ since $S_{2,\infty}$ can be written in closed form, as proved in Proposition 3. In Theorem 4, we prove that Assumption 1 identifies not only $H$ and $\mathcal{H}$, but the dimension $K$ as well.

To the best of our knowledge, the only criterion similar in spirit to our proposed (1.2) is that in [14], introduced in the much more restricted setting of a factor model in which the matrix $A$ has only 0/1 entries, each row is a canonical basis vector $e_k \in \mathbb{R}^K$, and the $p$ rows can be partitioned in $K$ groups of replicates of $e_k$, for each $k \in [K]$. Thus, in our notation, $H = [p]$ and $G = K$. In this mathematically simpler setting, it suffices to compute the supremum-norm

\[
\|\Sigma_{i \setminus \{i,j\}} - \Sigma_{j \setminus \{i,j\}}\|_\infty
\]

of the differences between rows $\Sigma_{i \setminus \{i,j\}}$ and $\Sigma_{j \setminus \{i,j\}}$ of the covariance matrix $\Sigma = \text{Cov}(X)$, for all pairs $i, j \in [p]$. When $A$ has general real-valued entries, this criterion no longer discriminates between general parallel rows, which we prove is made possible by an additional minimization over $(\alpha, \beta)$ on the set \{(\alpha, \beta) \in \mathbb{R}^2 : \|(\alpha, \beta)\|_r = 1\}. Furthermore, our proposed class of criteria is based on the scale invariant correlation matrix $R$.

2. A new method for estimating the approximate replicate index set, its partition, and the latent dimension. As the identifiability proofs of $H$ and $K$ are constructive, they lead
Detecting approximate replicate components

to a practical estimation procedure, stated in Section 2.2, that is easy to implement, even for large $p$. Section 2.2 first introduces the empirical counterpart $\hat{S}_{q,r}$ of the criterion $S_{q,r}$ in (1.2), by simply replacing $R$ by the empirical correlation matrix $\hat{R}$. From the tight, in probability, bound

$$\max_{i,j} |\hat{S}_{q,r}(i,j) - S_{q,r}(i,j)| \leq 2\delta_n$$

in Theorem 6 of Section 2.3, with $\delta_n = O(\sqrt{\log(p \vee n)/n})$, in conjunction with the characterization of $H$ in (1.3), we estimate $H$ by the set $\hat{H}$ of all pairs $(i,j)$ with $\hat{S}_{q,r}(i,j) \leq 2\delta_n$. In this paper, we derive the order of magnitude of $\delta_n$ under the assumption that $X$ is sub-Gaussian.

In Corollary 7 of Section 2.3 we show that $\hat{H}$ and its partition $\hat{H}$ consistently estimates $H$ and $H$, respectively. After estimating $H$ and $H$, Section 2.2 devises the estimation of $K$, exploiting the fact that $\Sigma - \Sigma_E = AS_ZA^\top$ has rank $K$. Theorem 8 in Section 2.3 shows that this procedure is consistent under mild regularity conditions.

1.1.2. Section 3: Identification and recovery under a canonical parametrization

While in Section 2 we studied the recovery of generic approximate replicates, in this section we shift focus to replicates generated in a particular way, and motivate our interest in this problem below.

1. Pure variables with arbitrary loadings. A particular instance of Assumption 1 is

**Assumption 2.** There exists a subset $I \subseteq H$ such that $A_I$ (up to row permutation) contains at least two $K \times K$ diagonal matrices with non-zero diagonal entries,

to which we refer in the sequel as a canonical parametrization. This assumption can be re-stated, equivalently as

**Assumption 2'.** For any $k \in [K]$, there exist at least two $i, j \in H$ with $i \neq j$ such that $A_{ik} \neq 0$, $A_{jk} \neq 0$ and $A_{ik'} = A_{jk'} = 0$ for all $k' \in [K] \setminus \{k\}$.

This version of the assumption will be referred to as the pure variable parametrization. The collection of the set of indices with existence postulated by Assumption 2' is the set $I$, defined in Assumption 2. We let $I = \{I_1, \ldots, I_K\}$ denote its partition.

One arrives at this sparse parametrization of $A$ from both mathematical and applied statistics perspectives. It has been long understood, see for instance [5, 7], that a particular version of Assumption 2 determines $A$ uniquely. When $I$ and $\Sigma_E$ are known, it is sufficient to assume the existence of only one diagonal sub-matrix in $A$, whereas the existence of a duplicate is a sufficient identifiability condition, when neither $I$ nor $\Sigma_E$ are known [9].

From an applied perspective, the interest in this parametrization can be most easily seen from its equivalent formulation, Assumption 2'. It is popular in the social sciences literature [35, 40], and is routinely used in educational and psychological testing, where
the latent variables are viewed as aptitudes or psychological states [41, 5, 12]. The components of $X$ are test results, with some tests specifically designed to measure only a single aptitude $Z_k$, for each aptitude, whereas others test mixtures of aptitudes.

This brings into focus a particular type of the replicate measurements considered here, sparse replicates, which satisfy $X_j = A_{jk} Z_k + E_j$, for all $j \in I_k$, and each $1 \leq k \leq K$. By experimental design, the index sets $I_k$ and the dimension $K$ are known in the classical applications mentioned above, and in this case Assumption $2'$ is known as the errors-in-variables parametrization, see for instance the review paper [45].

In modern applications, when $p$ is very large, as in genetic applications, or when many features are automatically collected, neither $I$ nor $K$ are known in advance, and are not identifiable without further modeling assumptions. We show in Section 3 that Assumption 2, or equivalently, Assumption $2'$, is sufficient for this task. We call the latter the pure-variable parametrization by following earlier terminology used in conjunction with models similar in spirit to ours, such as non-negative matrix factorization [17] and network analysis (see, for instance, [29]), but which are otherwise mathematically different.

A noteworthy aspect of Assumption $2'$ is that the loadings $A_{ik}$ and $A_{jk}$ of two pure variables $X_i$ and $X_j$, connected to the same latent factor $Z_k$, are allowed to be different. This is in line with assumptions made in topic models [6, 10], but has not yet been extended, to the best of our knowledge, to latent factor models for arbitrary random vector $X \in \mathbb{R}^p$, corresponding to $Z \in \mathbb{R}^K$ and a matrix $A$ with arbitrary real values. The work of [9] uses a restricted version of Assumption $2'$, motivated by biological applications, in which $|A_{ik}| = |A_{jk}| = 1$. Their entire estimation procedure of $I$ and $A$ is crucially tailored to a model in which all pure variables, in all groups, have the same loading, by convention taken to be equal to one, and it cannot be generalized to the model under the parametrization considered here. The procedure we propose, and therefore its analysis, are new, and entirely different from existing work.

2. Identifying the pure variable index set when the loading matrix has additional, non-pure, parallel, rows. Allowing for different loadings of the pure variables brings challenges in establishing the identifiability of $I$ and therefore of $A$, especially when we allow for the existence of other parallel, but with arbitrary entries, rows in $A$. In Section 3.1, we formally establish the identifiability of $I$ and of its partition $\mathcal{I} = \{I_1, \ldots, I_K\}$ under Assumption $2'$, and an additional assumption that we will discuss shortly.

Based on the observation that $I \subseteq H$, the first step towards identifying $I$ finds the set of parallel rows $H$ and its partition $\mathcal{H}$. When there are no parallel rows of $A$ corresponding to non-pure variables, $H$ and $\mathcal{H}$ reduce to $I$ and $\mathcal{I}$, respectively, and the results of Section 2 apply directly. However, if there exist non-pure variables corresponding to rows in $A$ that are parallel, it turns out that the set $I$ is not identifiable: the index set $J_1$ corresponding to these non-pure variables is also included in $H$, and $H = I \cup J_1$.

Separating $I$ from $J_1$ reduces to selecting $K$ distinct indices from $H$, and proving that they correspond to pure variables. One has liberty in selecting these indices. We opted for selecting those that correspond to variables that contain as much information as possible. A systematic way of selecting $K$ representative variables is given in Lemma 9 of Section 3.1 based on successively maximizing certain Schur complements of the low-rank matrix.
\[
\text{Cov}(AZ) = A \Sigma_Z A^\top.
\]
These quantities are equivalent with conditional variances, when \( Z \) follows a Gaussian distribution. Whereas this selection process may be of interest in its own right, we guarantee that its output is indeed a set of pure variables under the additional Assumption 4, stated in Section 3.1. It is a scaling assumption, that compares the loading \(|A_{jk}|\), for each \( j \in J_1 \) and \( k \in [K] \), with \( \max_{i \in I_k} \|A_{ip}\|_1 \), the largest loading of pure variables in \( I_k \).

We formally prove in Theorem 10 of Section 3.1 that the pure variable index set \( I \), its partition, as well as the assignment matrix \( A \), are indeed identifiable under Assumptions 2' & 4. We give a constructive proof that determines these quantities uniquely from the correlation matrix \( R \).

3. Estimating the pure variable set and its partition, with guarantees. The estimation of \( I \) follows the steps of the identifiability proofs. As the first step estimates \( H \) and \( K \) by the procedure in Section 2.2, we revisit theoretical guarantees of \( \hat{H} \) and \( \hat{K} \) under Assumption 2'. We show, in Corollary 11 of Section 3.3, that \( \hat{H} \) estimates \( I \), while possibly including a few indices \( i, j \) belonging to non-pure variables that are near-parallel in the sense that \( S_q(i,j) < 4\delta_n \).

Theorem 12 in Section 3.3 further shows the consistency of \( \hat{K} \) under mild assumptions, in particular on the size of the indices corresponding to the non-pure variables that are near-parallel.

After consistently estimating \( K \), Section 3.2 gives a principled way for sifting the pure variables from other variables with indices in the set \( \hat{H} \). This pruning step is a sample adaptation of the constructive method given, at the population level, in Lemma 9 of Section 3.1. Theorem 13 of Section 3.3 shows that this procedure consistently yields the pure variable index set, under certain regularity conditions. Its technical proof involves comparing estimated Schur complements of appropriate sub-matrices of \( A \Sigma_Z A^\top \). These quantities are not easy to handle, especially under the additional layer of complexity induced by the existence of near parallel variables, and a delicate uniform control between these matrices and their empirical counterparts is required. Nevertheless, under a simple set of conditions, we prove that our proposed procedure consistently finds the partition of the pure variable index set.

An application of the new procedure is provided by latent overlapping clustering, using the rationale in [9], but adding the flexibility provided by Assumption 2'.

A second application is to the estimation of a loading matrix \( A \) which satisfies Assumption 2'. The most difficult step in estimating \( A \) is the construction of an estimator \( \hat{I} \) of the set \( I \), which is one of the focus points of this work. Once an estimator \( \hat{I} \) is found, [9] proposed the following strategy: estimate \( A \) by concatenating estimators \( \hat{A}_{I_1} \), \( \hat{A}_{J_q} \), of the sub-matrices \( A_{I_1} \) and \( A_{J_q} \), where \( J = [p] \setminus I \), \( J = [p] \setminus \hat{I} \), and for any matrix \( M \) and row index set \( S \), we denote by \( M_S \), the sub-matrix formed from \( M \) by retaining \( S \) rows. The authors showed that the resulting \( \hat{A} \) is minimax-rate optimal and adaptive, but worked under the assumption that all pure variables loadings in Assumption 2' are equal to 1, and therefore \( \hat{A}_{I_1} \) is a matrix consisting in canonical basis vectors. While we adopt the same strategy as in [9] for the overall estimation of \( A \), we complement it by providing, in Section 3.4.1, an estimator \( \hat{A}_{I_1} \) tailored to Assumption 2'. Under fairly
mild regularity conditions, we establish its consistency in Theorem 14, and show that the resulting estimator of $A$ continues to be minimax-rate adaptive in Theorem 15.

1.1.3. Section 4 & Appendix: Simulations and proofs

In Section 4 we verify our theoretical claims via extensive simulations. Appendix B contains supplementary simulation results and practical considerations associated with the implementation of our procedure, including the selection of tuning parameters and a pre-screening procedure that detects variables with weak signal.

Finally, all proofs are collected in the Appendix. Perhaps of independent interest, as a byproduct of our proof, we establish in Appendix A.3 deviation inequalities in operator norm for the empirical sample correlation matrix based on $n$ independent sub-Gaussian random vectors in $\mathbb{R}^p$, with $p$ allowed to exceed $n$. While similar deviation inequalities for the sample covariance matrix have been well understood [42, 39, 13, 34], the operator norm concentration inequalities of the sample correlation matrix is relatively less explored. [20] studied the asymptotic behaviour of the limiting spectral distribution of the sample correlation matrix when $p/n \to (0, \infty)$. [43, 25] prove a similar result for Kendall’s tau sample correlation matrix.

1.2. Notation

For any positive integer $d$, we write $[d] := \{1, \ldots, d\}$. For two real numbers $a$ and $b$, we write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

For any vector $v \in \mathbb{R}^d$, we write its $\ell_q$-norm as $\|v\|_q$ for $0 \leq q \leq \infty$. We denote the $k$-th canonical unit vector in $\mathbb{R}^d$ by $e_k$, that has zero coordinates except for the $k$-th coordinate, which equals 1. Two vectors $u, v \in \mathbb{R}^d$ are parallel, and we write $v \parallel u$, if and only if $|\sin(\angle(u, v))| = 0$. For a subset $S \subset [d]$, we define $v_S$ as the subvector of $v$ with corresponding indices in $S$.

Let $M \in \mathbb{R}^{d_1 \times d_2}$ be any matrix. For any set $S_1 \subseteq [d_1]$ and $S_2 \subseteq [d_2]$, we use $M_{S_1, S_2}$ to denote the submatrix of $M$ with corresponding rows $S_1$ and columns $S_2$. We also write $M_{S_1, S_2}$ by removing the comma when there is no confusion. In particular, $M_{S_1, S_2}$ stands for the whole rows (columns) of $M$ in $S_1$ ($S_2$). We use $\|M\|_{\text{op}}, \|M\|_F$ and $\|M\|_{\infty}$ to denote the operator norm, the Frobenius norm and elementwise sup-norm, respectively. For any positive semi-definite matrix $M$, we denote by $\lambda_k(M)$ its $k$-th eigenvalue with non-increasing order.

Let $S = S_1 \cup \cdots \cup S_K$ be the collection of $K$ sets of indices with $S \subseteq [p]$. For any $i, j \in [p]$, the notation $i \overset{S}{\sim} j$ means that there exists $k \in [K]$ such that $i, j \in S_k$. Its complement $i \overset{\text{not } S}{\sim} j$ means $i$ and $j$ do not simultaneously belong to any $S_k$. Let $\Sigma := \text{Cov}(X)$ be the covariance matrix of the random vector $X \in \mathbb{R}^p$ with diagonal matrix $D_\Sigma = \text{diag}(\Sigma_{11}, \ldots, \Sigma_{pp})$ and we denote the correlation matrix of $X$ by $R = D_\Sigma^{-1/2}\Sigma D_\Sigma^{-1/2}$.
2. Identification and recovery of approximately replicate features

In this section we show that under Assumption 1 stated in the Introduction, both $H$ and its partition $\mathcal{H}$ introduced in Section 1.1, as well as the latent dimension $K$, can be uniquely determined from the scale invariant correlation matrix $R$. Our identifiability proofs are constructive, and are the basis of the estimation procedures described in Section 2.2, and further analyzed in Section 2.3.

2.1. Identification of the parallel row index set and latent dimension of $A$

We begin by noting that model (1.1) implies the decomposition

$$\Sigma = A\Sigma Z^\top + \Sigma E,$$

and consequently

$$R = D_{\Sigma}^{-1/2}\Sigma D_{\Sigma}^{-1/2} := B\Sigma Z B^\top + \Gamma, \quad (2.1)$$

where

$$B := D_{\Sigma}^{-1/2}A \quad \text{and} \quad \Gamma := D_{\Sigma}^{-1/2}\Sigma E D_{\Sigma}^{-1/2}.$$

Since the matrix $B$ has the same support as $A$, both matrices $A$ and $B$ share the same index set $H$ of parallel rows, and its partition $\mathcal{H}$.

The following proposition provides an if and only if characterization of both $H$ and $\mathcal{H}$. Its proof is deferred to Appendix A.1. We assume $\min_{i \in [p]} \|A_i\|_2 > 0$. Otherwise, we remove all zero rows of $A$ in the pre-screening step described in Appendix B.5. The notation $i \overset{H}{\sim} j$ means that both $i, j \in H_k$ for some $k \in [G]$. For any $i, j \in [p]$, let $R_{i,\{i,j\}} \in \mathbb{R}^{p-2}$ denote the $i$th row of $R$, with the entries in the $i$th and $j$th columns removed.

**Proposition 1.** Under model (1.1) and Assumption 1,

$$i \overset{H}{\sim} j \iff A_i \parallel A_j \iff R_{i,\{i,j\}} \parallel R_{j,\{i,j\}}. \quad (2.2)$$

Furthermore, $H$ and $\mathcal{H}$ are uniquely defined.\(^1\)

The first equivalence in (2.2) trivially follows from the definition of $H$. The second equivalence in (2.2) is key, especially with a view towards estimation, as it shows that the connection between $H$ and $A$ can be transferred over to the correlation matrix $R$, a model-free estimable quantity.

\(^1\)The partition $\mathcal{H}$ is unique up to a group permutation.
To make use of Proposition 1, first recall that for any two non-zero vectors \( v \) and \( w \), we have

\[
v \parallel w \iff \min_{\| (a,b) \|_r = 1} \| av + bw \|_q = 0
\]

for any integers \( 0 \leq q \leq \infty \) and \( 0 < r \leq \infty \), with \( (a,b) \in \mathbb{R}^2 \). This observation, in conjunction with Proposition 1, suggests the usage of the following general score function for determining, constructively, the index set \( H \):

\[
S_{q,r}(i,j) = (p-2)^{-1/q} \min_{\| (a,b) \|_r = 1} \| aR_{i\setminus \{i,j\}} + bR_{j\setminus \{i,j\}} \|_q \tag{2.3}
\]

for any \( i \neq j \), \( 0 < r \leq \infty \) and \( 0 \leq q \leq \infty \). The factor \((p-2)^{-1/q} \) serves as a normalizing constant. The following proposition justifies its usage, and offers guidelines on the practical choices of \( r \) and \( q \). Its proof can be found in Appendix A.1.2.

**Proposition 2** (A general score function for finding parallel rows). Under model (1.1) and Assumption 1, we have:

1. \( i \sim j \iff S_{q,r}(i,j) = 0 \) for any \( 0 < r \leq \infty \) and \( 0 \leq q \leq \infty \).
2. For any fixed \( i,j \in [p] \), \( S_{q,r}(i,j) \) defined in (2.3) satisfies
   
   (i) \( S_{q,t}(i,j) \geq S_{q,r}(i,j) \) for all \( 0 < r \leq t \leq \infty \);
   
   (ii) \( S_{t,r}(i,j) \geq S_{q,r}(i,j) \) for all \( t \geq q \).

In view of part (1) of Proposition 2, one should select \((q,r)\) such that \( S_{q,r}(i,j) \) is as large as possible whenever \( i \not\sim j \). Part (2)(i) of Proposition 2 immediately suggests taking \( r = \infty \) and considering the class of score functions

\[
S_q(i,j) := S_{q,\infty}(i,j) = (p-2)^{-1/q} \min_{\| (a,b) \|_\infty = 1} \| aR_{i\setminus \{i,j\}} + bR_{j\setminus \{i,j\}} \|_q \tag{2.4}
\]

In conjunction with part (2)(ii) of Proposition 2, the most ideal score function is

\[
S_\infty(i,j) := S_{\infty,\infty}(i,j) = \min_{\| (a,b) \|_\infty = 1} \| aR_{i\setminus \{i,j\}} + bR_{j\setminus \{i,j\}} \|_\infty \quad \forall i,j \in [p],
\]

Corresponding to \( q = \infty \). While this score function could in principle be computed via linear programming, it is expensive for very large \( p \) since \( S_{\infty,\infty}(i,j) \) needs to be computed for each pair \((i,j)\). A compromise is to choose \( q = 2 \), especially because the score function

\[
S_2(i,j) := S_{2,\infty}(i,j) = \frac{1}{\sqrt{p-2}} \min_{\| (a,b) \|_\infty = 1} \| aR_{i\setminus \{i,j\}} + bR_{j\setminus \{i,j\}} \|_2 \tag{2.5}
\]

has a closed form expression, given in Proposition 3 below and proved in Appendix A.1.3. To simplify our notation, and recalling that \( R_{i\setminus \{i,j\}}, R_{j\setminus \{i,j\}} \in \mathbb{R}^{p-2} \), we define

\[
V_{ii}^{(ij)} := [R_{i\setminus \{i,j\}}]_i^T R_{i\setminus \{i,j\}}, \quad V_{ij}^{(ij)} := [R_{i\setminus \{i,j\}}]_j^T R_{j\setminus \{i,j\}} \quad \forall i,j \in [p].
\]
Proposition 3. The score function $S_2(i, j)$ defined in (2.5) satisfies

$$[S_2(i, j)]^2 = \frac{V_{ii}^{(ij)} \wedge V_{jj}^{(ij)}}{p - 2} \left(1 - \frac{[V_{ij}^{(ij)}]^2}{V_{ii}^{(ij)} V_{jj}^{(ij)}}\right) \quad \forall \ i, j \in [p].$$

In particular, under model (1.1) and Assumption 1, we have:

$$i \overset{H}{\sim} j \iff S_2(i, j) = 0. \quad (2.6)$$

We showed in Proposition 1 that $H$ and its partition $\mathcal{H}$ are identifiable under Assumption 1, via an if and only if characterization of $H$, and further provided, in Proposition 3, a constructive if and only if characterization of $H$. The latter is used in the proof of Theorem 4 below to show that the latent dimension $K = \text{rank}(M_{HH})$, with $M := A\Sigma_Z A^\top$, can be uniquely determined, by showing that the matrix $M_{HH}$ itself is identifiable. Theorem 4 summarizes these identifiability results and its proof is given in Appendix A.1.4.

Theorem 4 (Partial identifiability). Under model (1.1) and Assumption 1, the set $H$, its partition $\mathcal{H}$ and $G = |H|$ are unique. Moreover, if additionally

$$\|R_{i,\setminus\{i,j\}}\|_q > 0, \quad \forall \ i \overset{H}{\sim} j, \quad (2.7)$$

holds, for any $q \geq 0$, then $K$ is uniquely determined.

Condition (2.7) ensures the uniqueness of $M_{HH}$. It holds if there exists at least one $\ell \in [p] \setminus \{i, j\}$ such that $R_{i,\ell} \neq 0$ for all $i \overset{H}{\sim} j$. Equivalently, under Assumption 1, (2.7) holds if each column of $A\Sigma_Z^{1/2}$ contains at least three non-zero entries. The latter is known to be necessary for identifying $M$, see Theorem 5.6 of [5].

Assumption 1 and (2.7) are sufficient for identifying $K$, but not the entire matrix $A$. To see this, note that for any invertible matrix $Q \in \mathbb{R}^{K \times K}$, there exists some diagonal (scaling) matrix $D \in \mathbb{R}^{K \times K}$ such that $AZ = \tilde{A}Z$, where $\tilde{A} = AQD$ has the same index set $H$ as $A$, and the covariance matrix $	ext{Cov}(\tilde{Z})$ of $\tilde{Z} := D^{-1}Q^{-1}Z$ is positive definite and satisfies $\text{diag}[\text{Cov}(\tilde{Z})] = 1$, as $	ext{Cov}(Z)$. We therefore term partial identifiability the results of Theorem 4. We revisit them in Section 3, where we introduce assumptions under which not only $K$, but also $A$ of model (1.1) can be identified.

2.2. Estimation of the parallel row index set $H$, its partition $\mathcal{H}$ and latent dimension $K$

Suppose we have access to $n$ i.i.d. copies of $X \in \mathbb{R}^p$, collected in a $n \times p$ data matrix $X$. We write the sample covariance matrix as

$$\hat{\Sigma} = \frac{1}{n} X^\top X.$$
Lemma 5. Let \( \delta \) be driven criterion of selecting one single tuning parameter \( \delta \). The estimated index set \( \hat{S} \) is stated in Appendix B.1, relying on the following lemma that we use \( \hat{S}_q(i, j) := S_{q, \infty}(i, j) \) as a generic score for finding \( H \), with any \( q \geq 1 \). We propose to estimate \( S_q(i, j) \) by solving the optimization problem
\[
\hat{S}_q(i, j) = (p - 2)^{-1/q} \min_{\|(a, b)\|_\infty = 1} \left\| a\hat{R}_{i, \{j\}} + b\hat{R}_{j, \{i\}} \right\|_q \tag{2.8}
\]
for each \( i, j \in [p] \) and \( i \neq j \). In particular, Proposition 3 implies that \( \hat{S}_2(i, j) \) has the closed form
\[
\hat{S}_2(i, j) = \left( \frac{\hat{V}_{ij}^{(ij)} \wedge \hat{V}_{ji}^{(ij)}}{p - 2} \left( 1 - \frac{[\hat{V}_{ij}^{(ij)}]^2}{\hat{V}_{ii}^{(ij)} \hat{V}_{jj}^{(ij)}} \right) \right)^{1/2} \tag{2.9}
\]
with \( \hat{V}_{ij}^{(ij)} = [\hat{R}_{i, \{j\}}]^T \hat{R}_{j, \{i\}} \) for all \( i, j \in [p] \).

Remark 1. For any pair \( (i, j) \) and any \( q \geq 1 \), the criterion \( \hat{S}_q(i, j) \) in (2.8) can be computed by solving two convex optimization problems because \( \hat{S}_q(i, j) \) is equal to
\[
(p - 2)^{-1/q} \min \left\{ \min_{|a| \leq 1} \|a\hat{R}_{i, \{j\}} + \hat{R}_{j, \{i\}}\|_q, \min_{|b| \leq 1} \|\hat{R}_{i, \{j\}} + b\hat{R}_{j, \{i\}}\|_q \right\}.
\]
In particular, for \( q = \infty \), computation of \( \hat{S}_\infty(i, j) \) requires solving two linear programs, while for \( q = 2 \), we have the closed form expression in (2.9).

Algorithm 1 gives the procedure of estimating parallel row index set, which reduces to finding all pairs \( (i, j) \) with \( \hat{S}_q(i, j) \) below the threshold level \( 2\delta \). It returns not only the estimated index set \( \hat{H} \), but also its partition \( \hat{H} := \{\hat{H}_1, \ldots, \hat{H}_{\hat{K}}\} \). Algorithm 1 requires one single tuning parameter \( \delta \), with an explicit rate stated in Section 2.3. A fully data-driven criterion of selecting \( \delta \) is stated in Appendix B.1, relying on the following lemma that shows that the number of estimated parallel rows of Algorithm 1 increases in \( \delta \).

**Lemma 5.** Let \( \hat{H}(\delta) \) be the estimated set of parallel rows from Algorithm 1. Then
\[
|\hat{H}(\delta)| \leq |\hat{H}(\delta')| \quad \forall \delta \leq \delta'.
\]

**Proof.** For a given \( \delta > 0 \), suppose \( i \in \hat{H}(\delta) \). Then, from Algorithm 1, there exists some \( j \neq i \) such that \( \hat{S}_q(i, j) \leq 2\delta \). This implies \( \hat{S}_q(i, j) \leq 2\delta' \) for any \( \delta' \geq \delta \). Hence, \( i \in \hat{H}(\delta') \), as desired. \( \square \)
Algorithm 1 Estimate the parallel row index set $H$ by $\hat{H}$ and its partition $\mathcal{H}$ by $\hat{\mathcal{H}}$

Require: Matrix $\hat{R} \in \mathbb{R}^{p \times p}$, a positive integer $q \geq 1$, a tuning parameter $\delta > 0$.

\begin{algorithm}
\begin{algorithmic}[1]
\Procedure{Parallel}{$\hat{R}, \delta$}
\State $\hat{H} \leftarrow \emptyset$.
\For{$i = 1, \ldots, p - 1$}
\For{$j = i + 1, \ldots, p$}
\State Compute $\hat{S}_q(i, j)$ by solving (2.8)
\If{$\hat{S}_q(i, j) \leq 2\delta$}
\State $\hat{H} \leftarrow \text{Merge}((i, j), \hat{H})$
\EndIf
\EndFor
\EndFor
\State \textbf{return} $\hat{H}, \hat{G} = |\hat{H}|$ and $\hat{H} = \cup_k \hat{H}_k$
\EndProcedure
\Function{Merge}{$S, \hat{H}$}
\State Add $\leftarrow$ True
\ForAll{$g \in \hat{H}$} \Comment{$\hat{H}$ is a collection of sets}
\If{$g \cap S \neq \emptyset$}
\State $g \leftarrow g \cup S$ \Comment{Replace $g \in \hat{H}$ by $g \cup S$}
\State Add $\leftarrow$ False
\EndIf
\EndFor
\If{Add}
\State $\hat{H} \leftarrow \hat{H} \cup \{S\}$ \Comment{add $S$ in $\hat{H}$}
\EndIf
\State \textbf{return} $\hat{H}$
\EndFunction
\end{algorithmic}
\end{algorithm}

Since $\hat{G}$ estimates $G$ and $G$ is typically larger than $K$, unless there are exactly $K$ sets of parallel rows in $A$, we propose the following procedure for estimating $K$ by using the output $\hat{H} = \{\hat{H}_1, \ldots, \hat{H}_\hat{G}\}$ of Algorithm 1. It relies on the observation that $[B \Sigma Z B^\top]_{LL}$ has rank $K$ where $L = \{\ell_1, \ldots, \ell_G\}$ with $\ell_k \in \hat{H}_k$ for each $1 \leq k \leq G$.

- For each $k \in [\hat{G}]$, we select one representative variable index from $\hat{H}_k$ as
  \begin{equation}
  \hat{\ell}_k := \arg \max_{i \in \hat{H}_k} \|\hat{R}_{i, \{i\}}\|_q, \quad (2.10)
  \end{equation}
  and create the set of representative indices
  \begin{equation}
  \hat{L} := \{\hat{\ell}_1, \ldots, \hat{\ell}_{\hat{G}}\}. \quad (2.11)
  \end{equation}

- Next, motivated by (A.2) and (A.3) in the proof of Theorem 4, we propose to estimate the submatrix $M_{LL}$ of $M := B \Sigma Z B^\top$ by
  \begin{equation}
  \hat{M}_{ij} = \hat{R}_{ij}, \quad \forall i, j \in \hat{H}, \ i \neq j, \quad (2.12)
  \end{equation}
  and
  \begin{equation}
  \hat{M}_{ii} = |\hat{R}_{i, \{i\}}| \frac{\|\hat{R}_{i, \{i\}}\|_q}{\|\hat{R}_{j, \{i\}}\|_q}, \quad \forall i \in \hat{H}_k, \ k \in [\hat{G}], \quad (2.13)
  \end{equation}
  \begin{equation}
  \hat{j} = \arg \min_{\ell \in \hat{L}_k \setminus \{i\}} \hat{S}_q(i, \ell).
  \end{equation}
Instead of choosing $\hat{j}$ for each $i \in \hat{H}_k$ as above, we could alternatively estimate $\hat{M}_{i,i}$ via (2.13) by averaging over $j \in \hat{H}_k \setminus \{i\}$. Our numerical experiments indicate that these two procedures have similar performance.

- Finally, we determine the approximate rank $\hat{K}$ of the matrix $\hat{M}_{\hat{L}\hat{L}}$ from (2.12) – (2.13) by

$$\hat{K} := \max \left\{ k \in [\hat{G}] : \lambda_k(\hat{M}_{\hat{L}\hat{L}}) \geq \mu \right\}$$  \hspace{1cm} (2.14)

for some tuning parameter $\mu > 0$.

### 2.3. Statistical guarantees for $\hat{H}$, $\hat{H}$ and $\hat{K}$

We will assume that the feature $X$ is a sub-Gaussian random vector. Recall that a centered random vector $X \in \mathbb{R}^d$ is $\gamma$–sub-Gaussian if $\mathbb{E}[\exp(u^\top X)] \leq \exp(\|u\|_2^2 \gamma^2/2)$ for any fixed $u \in \mathbb{R}^d$. The quantity $\gamma$ is called the sub-Gaussian constant. In this work, we treat $\gamma$ as some absolute constant and write $c = c(\gamma)$ and $C = C(\gamma)$ for numerical constants depending on $\gamma$ only.

**Assumption 3.** There exists a constant $\gamma$ such that $\Sigma^{-1/2}X$ is $\gamma$–sub-Gaussian.\(^2\)

The only tuning parameter in Algorithm 1 is $\delta$ with theoretical order given by

$$\delta_n := c(\gamma) \sqrt{\log(p \vee n)/n}$$ \hspace{1cm} (2.15)

for some constant $c(\gamma)$ depending on $\gamma$ only. $\delta_n$ is a key quantity that controls the deviation $\hat{R}$ from $R$. Indeed, under Assumption 3 and $\log p \leq n$, Lemma A.14 in Appendix A.5 shows that, with probability $1 - 4/(p \vee n)$, the event

$$\mathcal{E} := \left\{ \max_{1 \leq i,j \leq p} |\hat{R}_{ij} - R_{ij}| \leq \delta_n \right\}$$ \hspace{1cm} (2.16)

holds. Throughout the rest of the paper, we make the blanket assumption that $\log p \leq n$.

The following theorem provides the uniform deviation bounds for $\hat{S}_q(i,j) - S_q(i,j)$ over all $i,j \in [p]$ for any $1 \leq q \leq \infty$. Its proof is deferred to Appendix A.1.5. At this point, it is useful to discuss the value of $q$ in the criterion $\hat{S}_q$ used in Algorithm 1. We found in our simulations that $\hat{S}_q$ with $q = 2$ performs well in terms of statistical accuracy and computational speed, the latter due to its closed form. While we present our conditions and statements in terms of a general, fixed $q \geq 1$, our preferred choice is $q = 2$.

**Theorem 6.** On the event $\mathcal{E}$, one has, for any $1 \leq q \leq \infty$,

$$\max_{1 \leq i,j \leq p} \left| \hat{S}_q(i,j) - S_q(i,j) \right| \leq 2\delta_n.$$

\(^2\)Under model (1.1), if there exist constants $\gamma_Z, \gamma_E > 0$, such that $\Sigma_Z^{-1/2}Z$ and $\Sigma_E^{-1/2}E$ are sub-Gaussian random vectors with sub-Gaussian constants $\gamma_Z$ and $\gamma_E$, respectively, then $\Sigma^{-1/2}X$ is $\gamma$–sub-Gaussian with $\gamma = \max\{\gamma_Z, \gamma_E\}$.
If, in addition, model (1.1) and Assumption 1 hold, one has
\[
\hat{S}_q(i, j) \leq 2\delta_n, \quad \text{for all } i \not\sim j;
\]
\[
\hat{S}_q(i, j) \geq \max \{0, S_q(i, j) - 2\delta_n\}, \quad \text{for all } i, j \in [p]
\]
for any \(1 \leq q \leq \infty\).

Algorithm 1 with \(\delta = \delta_n\) selects those indices \(i, j \in \hat{H}\) for which \(\hat{S}_q(i, j) \leq 2\delta_n\). On the event \(E\), Theorem 6 tells us that
\[
H \subseteq \hat{H} \subseteq \{i \in [p]: S_q(i, j) \leq 4\delta_n \text{ for some } j \neq i\}.
\]
Hence, with probability at least \(1 - 4/(p \lor n)\), \(\hat{H}\) includes all parallel rows in \(H\) and may mistakenly include near-parallel rows corresponding to \(S_q(i, j) < 4\delta_n\). Note that this holds without imposing any signal strength condition.

On the other hand, the partition \(\hat{H}\), however, may not include all the groups that make up the partition \(H\). For instance, there may be two distinct groups of parallel rows with weak signals that get merged, while some subgroups of parallel rows may be enlarged by a few near-parallel rows. Nevertheless, Theorem 6 immediately implies that
\[
S_q(i, j) > 4\delta_n, \quad \text{for all } i \not\sim j
\]
is a sufficient condition for consistent estimation of both \(H\) and \(\hat{H}\), as summarized in the following corollary.

**Corollary 7.** Under model (1.1) and Assumption 1, assume that (2.18) holds. Then, on the event \(E\), the output \(\hat{G}\) and \(\hat{H} = \{\hat{H}_1, \ldots, \hat{H}_q\}\) from Algorithm 1 with \(\delta = \delta_n\) satisfy: \(\hat{G} = G\), \(\hat{H} = H\) and \(\hat{H}_k = H_{\pi(k)}\) for all \(k \in [G]\), for some permutation \(\pi : [G] \to [G]\).

The following proposition states the explicit rate of the tuning parameter \(\mu\) under which we provide theoretical guarantees for \(\hat{K}\) as an estimator of \(K\), for any \(q \geq 1\). For any \(L = \{\ell_1, \ldots, \ell_q\}\) with \(\ell_k \in H_k\) for all \(k \in [G]\), define
\[
\varphi(L) := \lambda_k(B_{L_k}.\Sigma_Z B_{L_k}^\top), \quad \varphi(L) := \lambda_1(B_{L_1}.\Sigma_Z B_{L_1}^\top).
\]

**Theorem 8.** Under model (1.1), Assumptions 1 \& 3 and condition (2.18), suppose there exists some constant \(c > 0\) such that
\[
\|R_{i,\setminus(i, j)}\|_q \geq c(p - 2)^{1/q}, \quad \forall i \sim j.
\]
For \(\mu = C(\sqrt{\hat{G}\delta_n^2 + \hat{G}\delta_n^2})\) in (2.14), and for some \(C > 0\) depending on \(\varphi(\hat{L})\), we have
\[
P\{\hat{K} \leq K\} \geq 1 - c'(n \lor p)^{-c''}.
\]
If, additionally, \(\max_k[\varphi(L)/\varphi(L)]\hat{G}\delta_n^2 \leq c_0\) for some sufficiently small constant \(c_0 > 0\), we further have
\[
P\{\hat{K} = K\} \geq 1 - c'(n \lor p)^{-c''}.
\]
The proof of Theorem 8 is deferred to Appendix A.1.6. We first prove that \( \hat{K} \leq K \) always holds on the event \( \mathcal{E} \) with the specified choice of \( \mu \). Proving consistency \( \hat{K} = K \) requires \( \mu \) to be sufficiently small so that \( \mu < c(\hat{L}) \), which is guaranteed by

\[
\max_L [\tau(L)/\zeta(L)] G \delta_n^2 = o(1).
\]

When both \( c(L) \) and \( \zeta(L) \) are bounded away from 0 and \( \infty \), consistency only requires

\[
G \delta_n^2 = o(1),
\]

that is, we allow \( G \) to grow but no faster than \( O(n/\log(n \vee p)) \). The data-driven choice of the leading constant of \( \mu \) is discussed in Appendix B.1. Condition (2.20) is a mild regularity condition and can be viewed as the sample analogue of (2.7). For instance, it requires

\[
\min_{i \neq j} \max_{\ell \neq i, \ell \neq j} |R_{i\ell}| \geq c \quad \text{for} \quad q = \infty.
\]

Sufficient conditions of (2.20) for \( q = 2 \) are provided in Theorem 12 of Section 3.3.

3. Identifiability and recovery under canonical parametrization

As argued in Section 2.1, although Assumption 1 is sufficient for identifying \( H \), its partition \( \mathcal{H} \) and the dimension of the model, it only determines \( A \) up to rotations. Considering the parametrization provided by Assumption 2' is a first step towards the unique determination of \( A \). If \( I \) is known, one can use the results of Section 3.1 below to identify \( A \). However, as \( I \) is also unknown, Assumption 2' is not sufficient for identifying \( I \), hence neither for \( A \). In Section 3.1 we provide another assumption that, combined with Assumption 2', is provably sufficient for determining \( I \), and therefore \( A \), uniquely. As the identifiability proofs are constructive, they naturally lead to the procedure of estimating the pure variable index set, as stated in Section 3.2. Section 3.3 provides the statistical guarantees for its output. Finally, the estimation of the loading matrix \( A \) and its statistical guarantees are provided in Section 3.4.

We begin by introducing the notation employed in the following sections. For model (1.1) satisfying Assumption 2', we let \( J := [p] \setminus I \) be the index set corresponding to non-pure variables. We introduce the index set corresponding to parallel, but non-pure, rows of \( A \) as

\[
J_1 := \{ j \in J : A_{j, \ell} \parallel A_{j, \ell} \text{ for some } \ell \in J \setminus \{ j \} \},
\]

and its partition \( J_1 := \{ J_1^1, \ldots, J_1^N \} \). With this notation, the index set \( H \) of all parallel rows of \( A \) and its partition, decompose as

\[
H = I \cup J_1 \quad \text{and} \quad \mathcal{H} = \{ H_1, \ldots, H_G \} = \{ I_1, \ldots, I_K, J_1^1, \ldots, J_1^N \}.
\]

The total number of groups in \( \mathcal{H} \) is \( G = K + N \).

3.1. Model identifiability under a canonical parametrization

Recall that \( H \) and \( \mathcal{H} \), as well as \( G = |\mathcal{H}| \) and \( K \), are identifiable under Assumption 1, hence under Assumption 2'. To begin discussing when \( I \) is also identifiable, we distinguish between two cases \( J_1 = \emptyset \) and \( J_1 \neq \emptyset \), by simply comparing \( G \) with \( K \).
If the only parallel rows in $A$ correspond to pure variables ($J_1 = \emptyset$ or, equivalently, $G = K$), then $I$ is identifiable as an immediate consequence of Theorem 4. If $J_1 \neq \emptyset$ (or equivalently, $G > K$), the pure variable set $I$ cannot be distinguished from $J_1$ (see, Appendix C for an example), unless further structure is imposed on the model.

Our general identifiability result of $I$ and its partition, stated in Theorem 10 below, allows for $J_1 \neq \emptyset$. The rationale behind its proof is the following:

**Step 0.** Show that $H$, $\mathcal{H}$ and $K$ can be uniquely determined, as in Theorem 4. With $|\mathcal{H}| = G$, if $G = K$, appeal to Theorem 4 to identify $I$ and its partition.

**Step 1.** If $G > K$, provide a statistically meaningful criterion for selecting $K$ representative indices from $H$. There is freedom in the choice of such a criterion, and we build one such that: (i) each representative contains as much information as possible; (ii) representatives of different groups are as uncorrelated as possible.

**Step 2.** Provide further conditions on $A$ under which the thus selected indices correspond to distinct pure variable indices. Then, use Proposition 2 to reconstruct the entire set $I$, and its partition, by the aid of the score function $S_2$ (or any $S_q$).

For **Step 1**, we propose to select indices of variables that maximize, successively, Schur complements of appropriately defined matrices, with general form given by (3.2). For the second step, we make Assumption 4, which is sufficient for proving the following fact: if the index $i_k$ is given by (3.2), then $X_{i_k}$ is indeed a pure variable, and can be taken as the representative of its group. This is formally stated in Lemma 9 and explained in its subsequent remark. With a view towards estimation, we note that this selection criterion will be constructively used in Section 3.2.

**Assumption 4.** Let $\xi_k := \max_{i \in I_k} \|A_i, \|_1$ be the largest loading in group $I_k$, in absolute value. We have

$$\sum_{k=1}^{K} \frac{|A_{jk}|}{\xi_k} \leq 1, \quad \forall j \in J_1.$$  

We note, trivially, that the assumption is automatically met when $J_1 = \emptyset$. Assumption 4 imposes a scaling constraint between $I$ and $J_1$, without placing any restriction on the rows of $A$ corresponding to $J \setminus J_1$. A sufficient condition for Assumption 4 is $\|A_i, \|_1 = \xi$ for all $i \in I$ and $\|A_j, \|_1 \leq \xi$ for all $j \in J_1$, which reduces it to the condition employed in [9] with $J_1 = J$ and $\xi = 1$, and thus generalizes that work.

Under Assumption 4, the lemma below states a systematic way of finding $K$ group representatives by successively maximizing certain Schur complements of $\Theta := A\Sigma_Z A^\top$. Its proof can be found in Appendix A.2.1. A toy example is provided in Appendix C.

**Lemma 9.** Assume model (1.1) and Assumptions 2′ & 4. For any $1 \leq k \leq K$, let $S_k = \{i_1, \ldots, i_{k-1}\}$ with $S_1 = \emptyset$ satisfying $i_a \in I_{\pi(a)}$ for all $1 \leq a \leq k - 1$ and some permutation $\pi : [K] \rightarrow [K]$. Then one has

$$i_k := \arg \max_{j \in H} \Theta_{jj} S_k \in I_{\pi(k)}, \quad \text{for all } 1 \leq k \leq K, \quad (3.2)$$
where $\Theta_{j|S_k} = \Theta_{jj} - \Theta_{jS_k} \Theta_{S_kS_k}^{-1} \Theta_{S_kj}$ and $H = I \cup J_1$.

**Remark 2.** The procedure in Lemma 9 is based on the following rationale which achieves the previous two goals (i) and (ii). Let $W_H := X_H - E_H = A_H Z$ for $H \subseteq [p]$ and observe that $\Theta$ is the degenerate (rank $K$) covariance matrix of $W \in \mathbb{R}^p$. To add intuition, if $Z$ has a multivariate normal distribution, then $\Theta_{H-H^c|H} = \text{Cov}(W_H|W_{H^c})$. In this case, display (3.2) in Lemma 9 becomes

$$i_k = \arg \max_{j \in H} \text{Var}(W_j|W_{S_k}), \quad \text{for all } 1 \leq k \leq K,$$

and the procedure returns the $K$ largest conditional variances $\text{Var}(W_j|W_{i_1}, \ldots, W_{i_{k-1}})$. Suppose we have already selected $W_{i_1}$ and we are considering the selection of a new index, $i_2$. Since

$$\text{Var}(W_j|W_{i_1}) = \Theta_{jj}[1 - \text{Corr}(W_j, W_{i_1})],$$

we see that maximizing the above conditional variance retains more information (for goal (i)), while avoiding linear dependence by reducing $\text{Corr}(W_{i_1}, W_{i_2})$ (for goal (ii)).

While one can always select $K$ variables, from a given collection, in this manner, it is Assumption 4 that ensures that their indices do indeed correspond to pure variables in this parametrization of the model. The de-noising step implicit in Lemma 9 is crucial for this procedure. It is made possible by the determination of the superset $H = I \cup J_1$, which in turn enables the identifiability of $\Theta$ and of its various functionals employed above, as shown in the proof of Theorem 4. In Section 3.2, these arguments will be used constructively for estimation purposes.

Based on the procedure in Lemma 9, both $I$ and $I$ are identifiable, and so is the entire loading matrix $A$. We summarize these results in the theorem below. Its proof is constructive and is deferred to Appendix A.2.2.

**Theorem 10.** Under model (1.1), Assumptions 2′ & 4 and condition (2.7), $I$ is identifiable and its partition $I$ is identifiable up to a group permutation. Furthermore, the matrix $A$ is identifiable up to a $K \times K$ signed permutation matrix.

**Remark 3** (Discussion of Assumptions 2′ and 4).

1. Discussion of Assumption 2′: quasi-pure variables. As discussed in detail in the Introduction, one way to parametrize a factor model (1.1) in which rank($A$) = $K$ is via the errors-in-variables parametrization, which refers to loading matrices $A$ that contain a $K \times K$ diagonal sub-matrix. In the terminology of this paper, this requires the existence of one pure variable per latent factor, and it fixes $A$ uniquely, when $I$ and $\Sigma_E$ are known. From this perspective, when $I$ and $\Sigma_E$ are not known, Assumption 2′ requires the existence of only one additional pure variable per latent factor.

To preserve full generality, it is perhaps more realistic to assume the existence of one extra quasi-pure variable, provided that it also leads to the identifiability results proved above. We argue below that this is possible, although it may lead to unnecessarily heavy
techniqualities that would obscure the main message of this work. We therefore content ourselves to explaining how such an assumption can be used, without pursuing it fully throughout the paper.

Suppose that there exists only one pure variable $i_k$ for some group $k \in [K]$, that is, $|I_k| = 1$. Assume also that there exists some quasi-pure variable $j$ for this group $k$, in the sense that for that index $j$ we have:

$$\sum_{a \neq k} |A_{ja}| \parallel A_{ja} \parallel_1 = 1 - \frac{|A_{jk}| \parallel A_{j} \parallel_1}{\epsilon^2} \leq \epsilon^2. \quad (3.3)$$

As we can see, when $\epsilon$ is small, the $j$th variable is close to the pure variable $i_k$, as the majority of the weights in its corresponding row of $A$ are placed on the $k$th factor. We show in Appendix A.5.1 that, for this $i_k$ and $j$, the score in (2.5) satisfies $S_2(i_k, j) \leq 2\epsilon$. By slight abuse of notation, we write $I_k = \{i_k, j\}$ and provided that $S_2(\ell, \ell') > 2\epsilon$ for all $\ell \not\sim \ell'$ with $S_2(\ell, \ell')$ defined in (2.5), both $I$ and its partition can be recovered uniquely by applying Algorithm 1 to the population correlation matrix $R$, with $q = 2$ and $\delta \geq 2\epsilon$.

2. Discussion of Assumption 4. This assumption is only active when $J_1 \neq \emptyset$. If $J_1 \neq \emptyset$ and Assumption 2' holds, but Assumption 4 does not hold, then the representative selection of Lemma 9, as well as group reconstruction, can still be performed in an identical manner. However, one cannot guarantee that all groups consist of only pure variables. Nevertheless, their representatives will continue to have properties (i) and (ii) in Step 1, by construction, and therefore still be statistically meaningful.

3.2. Estimation of the pure variables index set

The estimation procedure follows, broadly, the Steps 0 – 2 employed in the proof of Theorem 10 of the previous section. We first use Algorithm 1 and the procedure of estimating $K$ described in Section 2.2 to obtain estimates $\hat{H} = \{\hat{H}_1, \ldots, \hat{H}_G\}$ and $\hat{K}$. If $\hat{K} = \hat{G}$, no further action is taken; if $\hat{K} < \hat{G}$, we add the pruning step stated below, based on the sample analogue of Lemma 9.

For any input $r < \hat{G}$ (for instance, $r = \hat{K}$ when $\hat{K} < \hat{G}$), the pruning step consists of the following steps to estimate $\hat{I}$, collected in Algorithm 2.

- Estimate $\Gamma_{\hat{H}\hat{H}}$ by $\hat{\Gamma}_{ij} = 0$ for all $i \neq j$ and
  $$\hat{\Gamma}_{ii} = 1 - \hat{M}_{ii}, \quad \forall i \in \hat{H} \quad (3.4)$$
  where $\hat{M}_{ii}$ is obtained from (2.13) with $q = 2$.
- Set $\Theta := A\Sigma_Z A^\top$ and, in view of (2.1), estimate $\Theta_{\hat{H}\hat{H}}$ by
  $$\hat{\Theta}_{\hat{H}\hat{H}} = \left[\hat{\Sigma} - \hat{\Sigma}_E\right]_{\hat{H}\hat{H}} = \left[\hat{\Sigma} - \hat{D}_\Sigma^{1/2}\hat{D}_\Sigma^{1/2}\right]_{\hat{H}\hat{H}} \quad (3.5)$$
  with $\hat{D}_\Sigma = \text{diag}(\hat{\Sigma}_{11}, \ldots, \hat{\Sigma}_{pp})$. 

For any set $S \subseteq \hat{H}$ and $S^c = \hat{H} \setminus S$, write the Schur complement of $\hat{\Theta}_{SS}$ of $\hat{\Theta}_{H\hat{H}}$ as

$$
\hat{\Theta}_{S^c S^c | S} = \hat{\Theta}_{S^c S^c} - \hat{\Theta}_{S^c S} \hat{\Theta}_{SS} \hat{\Theta}_{SS^c}
$$

(3.6)

with $M^-$ denoting the Moore-Penrose pseudo-inverse of $M$. Set $S_1 = \emptyset$, and let $S_k = \{i_1, \ldots, i_{k-1}\}$ for each $2 \leq k \leq r$, and define

$$
i_k = \arg \max_{j \in \hat{H}} \hat{\Theta}_{jj | S_k}.
$$

(3.7)

When there are ties, arbitrarily pick one of the maximizers.

The final estimate of $I$ is defined as

$$
\hat{I} = \left\{ \{\hat{H}_k\}_{1 \leq k \leq r} : \text{there exists } a \in [r] \text{ such that } i_a \in \hat{H}_k \right\}.
$$

(3.8)

Algorithm 2 Prune the parallel row index set obtained from Algorithm 1

Require: $\hat{\Sigma}, \hat{R} \in \mathbb{R}^{p \times p}$, the partition $\hat{H}$ with $\hat{G} = |\hat{H}|$, the integer $1 \leq r < \hat{G}$.

1: \textbf{procedure} Pruning$(\hat{\Sigma}, \hat{R}, \hat{H}, r)$
2: \hspace{1em} Compute $\hat{\Theta}_{\hat{H}\hat{H}}$ from (3.5)
3: \hspace{1em} Set $S = \emptyset$
4: \hspace{1em} for $k = 1, \ldots, r$ do
5: \hspace{2em} Compute $i_k$ from (3.7) and add $i_k \in S$
6: \hspace{1em} return $\hat{I}$ obtained from (3.8)

We note that Algorithm 2 can take any $1 \leq r \leq K$ as input, including a random $r$. This adds flexibility to the procedure, should one want to use a value of $r$ different from the value of $\hat{K}$ defined in (2.14), for instance if one uses a different estimator of $K$, or one is interested in a pre-specified value of $r$. We analyze $\hat{I}$ in the next section.

For the reader’s convenience, we summarize our whole procedure for estimating the pure variable index set and its partition, Pure Variable Selection (PVS), in Algorithm 3.

Algorithm 3 Pure Variable Selection (PVS)

Require: $\hat{\Sigma}, \hat{R} \in \mathbb{R}^{p \times p}$, a positive integer $q \geq 1$, two positive tuning parameters $\delta$ and $\mu$.

1: Obtain $\hat{H}$ and $\hat{G} = |\hat{H}|$ from Parallel$(\hat{R}, \delta)$ in Algorithm 1
2: Compute $\hat{L}$ from (2.10) – (2.11) and compute $\hat{M}_{\hat{L}\hat{L}}$ from (2.12) – (2.13)
3: Estimate $\hat{K}$ from (2.14)
4: if $\hat{K} < \hat{G}$ then
5: \hspace{1em} Obtain $\hat{I}$ from Pruning$(\hat{\Sigma}, \hat{R}, \hat{H}, \hat{K})$
6: else
7: \hspace{1em} Set $\hat{I} = \hat{H}$
8: return $\hat{I}$
3.3. Statistical guarantees for the estimated pure variable index set

We provide statistical guarantees for the estimated pure variable index set obtained via Algorithm 3. As Algorithm 1 is used to estimate $\hat{H}, \hat{H}$ and $K$ first, we start by revisiting the statistical guarantees for $\hat{H}, \hat{H}$ and $\hat{K}$ under the canonical parametrization of $A$. The theoretical properties of $\hat{I}$ are given in Theorem 13.

As shown in (2.17) of Section 2.3, on the event $\mathcal{E}$ in (2.16), $\hat{H}$ from Algorithm 1 includes $H = I \cup J_1$ with $J_1$ defined in (3.1), but may mistakenly include other variable indices, for instance, pairs satisfying $S_q(i,j) < 4\delta_n$. Under Assumption 1, if the signal condition (2.18) holds, we further showed that both $\hat{H}$ and its partition $\hat{\mathcal{H}}$ are consistent.

Under the canonical parametrization provided by Assumption 2', the set of indices involved in condition (2.18) can be reduced, and the following weaker condition suffices for consistent recovery:

$$S_q(i,j) > 4\delta_n \quad \text{for all } i \neq j, i \in I.$$  \hfill (3.9)

Note that we still allow

$$S_q(i,j) \leq 4\delta_n \quad \text{for some } i \neq j, i,j \in J,$$

by recalling that $J = [p] \setminus I$. We show, in Lemma A.11 of Appendix A.4, that non-pure variables with indices satisfying the above display have near-parallel corresponding rows in $A$. By collecting these near-parallel rows with indices in $\bar{J}$ in the new set

$$\bar{J}_1 := \{j \in J : S_q(j,\ell) \leq 4\delta_n \text{ for some } \ell \in J \setminus \{j\}\},$$ \hfill (3.10)

we have $J_1 \subseteq \bar{J}_1$. In fact, as the quantity $4\delta_n$ in (3.10) originates from the estimation error of the score function $S_q$, the set $\bar{J}_1$ can be viewed as the sample analogue of $J_1$.

We begin by presenting the analogue of the results of Section 2.3, by giving recovery guarantees for $\hat{H}$ and $\hat{\mathcal{H}}$, in Corollary 11, and for $\hat{K}$ in Theorem 12, under model (1.1) satisfying the canonical parametrization given by Assumption 2'. Write $\lfloor x \rfloor$ to denote the largest integer that is no greater than $x$.

**Corollary 11.** Under model (1.1), Assumption 2' and (3.9), on the event $\mathcal{E}$, the outputs $\hat{\mathcal{G}}$ and $\hat{\mathcal{H}} = \{\hat{H}_1,\ldots,\hat{H}_{\hat{\mathcal{G}}}\}$ of Algorithm 1, applied with $\delta = \delta_n$, satisfy

(1) $K \leq \hat{\mathcal{G}} \leq K + \lfloor |\bar{J}_1|/2 \rfloor$;
(2) $I \cup J_1 \subseteq \hat{\mathcal{H}} \subseteq (I \cup \bar{J}_1)$;
(3) $I_k = \hat{H}_{\hat{\pi}(k)}$ for all $1 \leq k \leq K$ for some permutation $\hat{\pi} : [\hat{\mathcal{G}}] \to [\hat{\mathcal{G}}]$.

**Proof.** The result is a direct consequence of Theorem 6 and (3.9) in conjunction with the fact that $\hat{\mathcal{H}}$ contains at most $\lfloor |\bar{J}_1|/2 \rfloor$ groups that only consist of variables in $\bar{J}_1$. 


Since (3.9) is weaker than (2.18) as $I \subseteq H$, Corollary 11 is a stronger result than Corollary 7. Appendix A.4 provides insight into condition (3.9) in terms of its induced restrictions on the model parameters. Summarizing the discussion therein, we show, under mild regularity conditions on $B\Sigma ZB^\top$, that (3.9) holds under the following mild separation condition between pure and non-pure variables,

$$\min_{i\in I,j\notin I} \sin(\angle(A_i, A_j)) \gtrsim \delta_n.$$

We proceed to revisit and analyze $\hat{K}$, constructed in Section 2.2, with $q = 2$. Theorem 12 below states the explicit rate of the tuning parameter $\mu$ required for its estimation. The theorem offers the same guarantees as Theorem 8, but they are established under slightly different conditions, that reflect the usage of the canonical parametrization given by Assumption 2', and of the weaker condition (3.9) enabled by this parametrization. Define

$$c_{b,I} := \min_{j\in I} \|B_j\|_2^2, \quad \bar{c}_z := \lambda_1(\Sigma Z), \quad c_z := \lambda_K(\Sigma Z),$$

and

$$c_r := \min_{i\neq j} \frac{1}{p-2} \lambda_K\left(B_{\{i,j}\}, \Sigma Z B_{\{i,j\}}^\top\right).$$

(3.11)

(3.12)

**Theorem 12.** Under model (1.1) and Assumptions 2' & 3, assume (3.9) and $\log p = o(n)$. In addition, suppose there exist absolute constants $0 < c \leq C < \infty$ such that

$$\min(c_{b,I}, c_z, c_r) > c, \quad \bar{c}_z \leq C$$

(3.13)

For $\mu = C'(\sqrt{K\delta_n^2} + K\delta_n^2 + |\bar{J}_1|\delta_n)$ in (2.14), and for a large enough constant $C' > 0$, we have

$$\lim_{n\to\infty} P\{\hat{K} \leq K\} = 1.$$

If, in addition, $\max\{K\delta_n^2, |\bar{J}_1|\delta_n\} = o(1)$ as $n \to \infty$, we further have

$$\lim_{n\to\infty} P\{\hat{K} = K\} = 1.$$

The proof of Theorem 12 is deferred to Appendix A.2.3. It relies on a careful analysis of $\|\hat{M}_{LL} - M_{LL}\|_{op}$, performed when allowing $\bar{J}_1 \neq \emptyset$.

We first prove that $\hat{K} \leq K$ always holds on the event $\mathcal{E}$ in (2.16). In conjunction with part (1) of Corollary 11, this ensures that, with high probability, the event $\{\hat{G} = \hat{K}\}$ implies the event $\{\hat{G} = K\}$. In this case, our procedure stops and the output of Algorithm 3 is $\hat{H}$ from Algorithm 1. On the other hand, if $\hat{K} < \hat{G}$, as explained in Section 3.2, Algorithm 3 uses the pruning step from Algorithm 2 to estimate $I$ and its partition.

Proving consistency $\hat{K} = K$ requires $\mu$ to be sufficiently small so that $\mu < c_{b,I}c_z$, which is guaranteed by $\max\{K\delta_n^2, |\bar{J}_1|\delta_n\} = o(1)$. See Remark 4 below for a discussion of this condition and of (3.13).
The following theorem gives theoretical guarantees for the output of Algorithm 2 with any $1 \leq r \leq K$, and, in particular, the output of Algorithm 3 if $r$ is set to $\hat{K}$. Recall that $\xi_k$ is defined in Assumption 4.

**Theorem 13.** Under model (1.1) and Assumptions 2′ & 3, assume (3.9) and (3.13). Suppose that

$$\lim_{n \to \infty} K\delta_n^2 = 0$$

(3.14)

and, if $\bar{J}_1 \neq \emptyset$, assume that there exist absolute constants $0 < C_0 < \infty$ and $0 < \varepsilon < 1$ such that

$$\max_{1 \leq k \leq K} \xi_k \leq C_0 \min_{1 \leq k \leq K} \xi_k,$$

(3.15)

$$\max_{j \in \bar{J}_1} \left( \sum_{k=1}^{K} |A_{jk}| \xi_k \right)^2 < 1 - \varepsilon.$$  

(3.16)

Then, there exists some permutation $\pi : [K] \to [K]$ such that the output $\hat{I}$ of Algorithm 2 by using any $1 \leq r \leq K$ satisfies

$$\lim_{n \to \infty} \mathbb{P} \left\{ \hat{I}_a = I_{\pi(a)} \text{ for all } 1 \leq a \leq r \right\} = 1.$$

In particular, the claim is valid for $r = \hat{K}$ defined in (2.14). In this case, if additionally $|\bar{J}_1|\delta_n = o(1)$, then with probability tending to one as $n \to \infty$, $\hat{K} = K$ and the output $\hat{I}$ from Algorithm 3 satisfies $\hat{I}_a = I_{\pi(a)}$, for all $1 \leq a \leq K$.

As already mentioned, the statement of Theorem 13 is formulated in terms of $r$ groups, for any $1 \leq r \leq K$. In this way, in case we use some estimator that under-estimates $K$, Theorem 13 still ensures that a subset of the groups consisting in pure variable indices (corresponding to the largest conditional variances) can be consistently estimated. When $K$ is consistently estimated, consistent estimation of the entire pure variable index set, and of its partition, is guaranteed.

Allowing for a general statement in Theorem 13, which is valid for a general $1 \leq r \leq K$, brings technical difficulty to its proof, which is stated in Appendix A.2.4. One of the main challenges is to control the difference between the estimated Schur complement $\hat{\Theta}_{jj|S_k}$ and its population-level counterpart, uniformly for all $j \in \hat{H}$, all $S_k$ selected from (3.7) and all $1 \leq k \leq r$. The difficulty is further elevated by the fact that $\bar{J}_1 \neq \emptyset$. We provide this uniform control in Lemma A.6 of Appendix A.2.4 and its proof further relies on the uniform controls of the sup-norm of $\hat{M}_{\hat{H},\hat{H}} - M_{\hat{H},\hat{H}}$, the operator norm of $\hat{M}_{\hat{S}_k}$, and the quadratic form $\hat{M}_{\hat{S}_k}^{1/2} \hat{M}_{\hat{S}_k}^{-1/2} \hat{M}_{\hat{S}_k}$, collected in Lemmas A.4, A.5 and A.7 of Appendices A.2.3 and A.2.4.
Remark 4 (On the conditions of Theorems 12 & 13). The proofs for both results are non-asymptotic, in the sense that their statements continue to hold when \( K \delta_n^2 \leq c \), for some sufficiently small constant \( c \), instead of the assumed \( K \delta_n^2 = o(1) \) in (3.14). Then, they would hold on an event with probability \( 1 - c'(p \lor n)^{-c''} \) for some constants \( c',c'' > 0 \). To avoid the delicate interplay between different constants, we opted for the current asymptotic formulation.

Condition (3.14) requires \( K \) not to grow too fast relative to \( n \), \( K \log(p \lor n) = o(n) \). Condition \( |J_1| \delta_n = o(1) \) in Theorem 12 allows the size of the index set corresponding to near-parallel rows in \( A \) to grow, but slower than \( \sqrt{n}/\log(n \lor p) \).

Conditions (3.15) and (3.16) are only needed when \( J_1 \neq \emptyset \). Condition (3.16) is the analogue of Assumption 4 and allows us to distinguish, at the sample level, pure variables from non-pure variables corresponding to rows in \( A \) that are close to parallel (in the sense \( S_q(i,j) < 4\delta_n \)). Condition (3.15) allows us to eventually separate two pure groups, \( I_a \) and \( I_b \), with \( a \neq b \), from each other. It prevents the pure variable variances from being very different. For instance, (3.15) holds if \( \max_k \max_{i \in I_k} \Sigma_{ii} \leq C \min_k \max_{i \in I_k} \Sigma_{ii} \) coupled with (3.13). Finally, condition (3.13) is a mild regularity condition on the matrix \( B \) and \( \Sigma_Z \), which is needed in our rather technically involved proofs. More discussion of this condition can be found in Appendix A.4.

3.4. Application: Estimation of \( A \) under a canonical parametrization

As announced in the Introduction, an immediate application of estimating the index set of pure variables \( I \) is to the estimation of a loading matrix \( A \) that satisfies Assumption 2'. Following [9], we first estimate \( B_{\tilde{I}_k} := D_{\Sigma}^{-1/2} A_{\tilde{I}_k} \), and \( B_{\tilde{J}_j} := D_{\Sigma}^{-1/2} A_{\tilde{J}_j} \), by the estimators \( \hat{B}_{\tilde{I}_k} \) and \( \hat{B}_{\tilde{J}_j} \), presented below. Once this is done, \( A \) is easily estimated by \( \hat{A} \) obtained by concatenating the sub-matrices

\[
\hat{A}_{\tilde{I}_k} = \hat{D}_{\Sigma}^{1/2} \hat{B}_{\tilde{I}_k}, \quad \hat{A}_{\tilde{J}_j} = \hat{D}_{\Sigma}^{1/2} \hat{B}_{\tilde{J}_j}, \quad \text{with} \quad \hat{D}_{\Sigma} = \text{diag}(\hat{\Sigma}_{11}, \ldots, \hat{\Sigma}_{pp}). \quad (3.17)
\]

3.4.1. Estimation of \( B_{\tilde{I}_k}, \Sigma_Z \) and \( A_{\tilde{I}_k} \)

Since Assumption 2' and \( \text{diag}(\Sigma_Z) = 1 \) imply that \( M_{ii} = B_{ii}^2 \) for any \( i \in I_k \), and given the estimated partition of the pure variables \( \tilde{I} = \{ \tilde{I}_1, \ldots, \tilde{I}_{\tilde{K}} \} \), we propose to estimate \( B_{\tilde{I}_k} \) by

\[
\hat{B}_{ik}^{\tilde{I}_k} = \hat{M}_{ik}, \quad \forall i \in \hat{I}_k, \quad k \in \{ \tilde{K} \} \quad (3.18)
\]

with \( \hat{M}_{ik} \) defined in (2.13). Since we can only identify \( B_{\tilde{I}_k} \), up to a signed permutation matrix, we use the convention that \( i_k \) is the first element in \( \hat{I}_k \). Using the rationale in (A.20), in the proof of Theorem 13, we set

\[
\text{sgn}(\hat{B}_{ik}) = 1, \quad \text{sgn}(\hat{B}_{jk}) = \text{sgn}(\hat{R}_{ij}), \quad \forall j \in \hat{I}_k \setminus \{ i_k \}. \quad (3.19)
\]
We estimate the diagonal of $\Sigma_Z$ by $\text{diag}(\hat{\Sigma}_Z) = 1$ and, following (A.21) and (A.22), we estimate the off-diagonal elements of $\Sigma_Z$ by the corresponding entries of
\[
\hat{B}_{i*}^+ \left( \hat{R}_{II} - \hat{\Gamma}_{II} \right) \begin{bmatrix} \hat{B}_{i*}^+ \end{bmatrix}^\top \tag{3.20}
\]
with $\hat{B}_{i*}^+ = \left[ \hat{B}_{i*} \hat{B}_{i*}^\top \right]^{-1} \hat{B}_{i*}^\top$ being the left inverse of $\hat{B}_{i*}$, and $\hat{\Gamma}_{II}$ estimated from (3.4).

We provide theoretical guarantees for the estimated loadings of the pure variable sub-matrix $A_1$, and for the estimates of the covariance matrix $\Sigma_Z$ of $Z$, obtained from the above procedure for any $q \geq 2$. The proofs are deferred to Appendix A.2.5.

**Theorem 14.** Under the conditions of Theorem 13, assume $|J_1| \delta_n = o(1)$, with $\delta_n$ given in (2.15). With probability tending to one, as $n \to \infty$, we have $\hat{I}_k = I_{\pi(k)}$ for $k \in [K]$ and for some permutation $\pi : [K] \to [K]$, and
\[
\min_{P \in \mathcal{P}_K} \max_{i \in I_k} \| \hat{B}_{i*} - PB_{i*}^\top \|_\infty \lesssim \delta_n,
\]
\[
\min_{P \in \mathcal{P}_K} \| \hat{\Sigma}_Z - P^\top \Sigma_Z P \|_\infty \lesssim \delta_n,
\]
\[
\min_{P \in \mathcal{P}_K} \max_{i \in I_k} \frac{\| \hat{A}_{i*} - PA_{i*}^\top \|_\infty}{\sqrt{\Sigma_{ii}}} \lesssim \delta_n.
\]
The minimum is taken over the set $\mathcal{P}_K$ of $K \times K$ signed permutation matrices.

Similar to Theorem 13, the results of Theorem 14 can be easily stated non-asymptotically, as explained in Remark 4 of the preceding section.

### 3.4.2. Estimation of $A_J$

Our proposed estimation of $A_J$ is identical to that in [9], and we include the main steps for completeness. Since the correlation matrix $R$ takes the form $B \Sigma_Z B^\top + \Gamma$, we can write
\[
\begin{bmatrix} R_{II} & R_{IJ} \\ R_{JI} & R_{JJ} \end{bmatrix} = \begin{bmatrix} B_{IJ}, \Sigma_Z B_{IJ}^\top \\ B_{JI}, \Sigma_Z B_{JI}^\top \end{bmatrix} + \begin{bmatrix} \Gamma_{II} \\ \Gamma_{JJ} \end{bmatrix}.
\]
In particular, from $R_{II} = B_{IJ}, \Sigma_Z B_{IJ}^\top$, the submatrix $B_{IJ}$, is solved via
\[
B_{IJ}^\top = \Sigma_Z^{-1}[B_{IJ} B_{IJ}]^{-1} B_{IJ}^\top, R_{IJ}.
\]
Hence, after estimating $\Sigma_Z$ and $B_{IJ}$, using the estimate $\hat{R}$, we obtain a plug-in estimate for $B_{IJ}$, with $\hat{J} = [p] \setminus \hat{I}$. Alternatively, we can regress $\hat{R}_{IJ}$ on $B_{IJ}, \hat{\Sigma}_Z$, or $[B_{IJ} B_{IJ}]^{-1} B_{IJ}^\top, \hat{R}_{IJ}$ on $\hat{\Sigma}_Z$. This approach has been adopted in [9] (and [10] in the context of topic models), and allows for incorporating sparsity restrictions on $B$ and hence on $A$, and ultimately can result in estimation at minimax-optimal rates.
Although there exists a large body of literature on loading matrix estimation, results accompanied by finite sample risk bounds are very scarce, and limited to the case when $I$ is known [5, 7]. A full discussion on the estimation of $A$ is beyond the scope of this paper, but we refer the reader to pages 2073–2075 in Section 4.4. of [9], for an overview of existing approaches. We also refer to Appendix C.1 in [9], for a detailed comparison with a pseudo-likelihood based method proposed by [7]. This method is, to the best of our knowledge, the only procedure accompanied by theoretical guarantees for estimating $A$ in a framework similar to ours, albeit with focus on asymptotic results, derived under the classical version of Assumption 2’.

As an illustration, suppose we take the approach in (3.13) – (3.14) of [9] to estimate each row $B_{j*}$ of $B_J$, and estimate $A$ via (3.17). More precisely, we estimate $B_{j*}$ for each $j \in J$ by

$$\hat{B}_{j*} = \arg \min_\beta \|\beta\|_1 \text{ subject to } \|\hat{\Sigma}_Z Z^\top \beta - \hat{B}_{j*} \hat{I} \|_\infty \leq C_0 \delta_n \quad (3.21)$$

for some constant $C_0 > 0$. Here $\hat{\Sigma}_Z$ is constructed via (3.20) above with $\text{diag}(\hat{\Sigma}_Z) = 1$.

The following theorem provides the convergence rate of $\|\hat{A} - AP\|_{\infty,q}$ for all $q \geq 1$, in the setting of bounded $\|\cdot\|_{\infty,1}$-norms of $A$ and $\Sigma^{-1}_Z$ as in [9]. Its proof can be found in Appendix A.2.6. Let $s = \max_i \|A_i\|_0$.

**Theorem 15.** Under conditions of Theorem 14, assume there exists positive finite constants $c, C, C', C''$ such that $c \leq \min_i \Sigma_{ii} \leq \max_i \Sigma_{ii} \leq C$, $\|\Sigma^{-1}_Z\|_{\infty,1} \leq C'$ and $\|A\|_{\infty,1} \leq C''$. Then, with probability tending to one, we have, for all $1 \leq q \leq \infty$,

$$\min_{P \in \mathcal{P}_K} \|\hat{A} - AP\|_{\infty,q} \lesssim s^{1/q} \delta_n.$$

According to Theorem 6 of [9], the rate in Theorem 15 is minimax optimal, up to a multiplicative factor of $\sqrt{\log(p \vee n)}$. Note that $C'' = 1$ in [9] and we refer to Remark 4 of [9] for a detailed discussion on $\|\Sigma^{-1}_Z\|_{\infty,1}$.

4. Simulations

In this section, we conduct extensive simulations to support our theoretical findings from two aspects: the recovery of pure variables and the estimation errors of $A_J$, $\Sigma_Z$ and $[\Sigma E]_{IJ}$. We start by describing the metrics used to quantify these aspects.

**Metrics** We first introduce the metrics that quantify the estimation performance of the pure variable index set and respective partition. Recall that, for any estimate $\hat{I}$ of $I$, we need to evaluate both its estimation quality, and also that of $\hat{I} = \cup_k \hat{I}_k$ as an estimate of $I$. Towards this end, the first two metrics that we consider are the true positive rate (TPR) and the true negative rate (TNR), defined as

$$\text{TPR} = \frac{|\hat{I} \cap I|}{|I|}, \quad \text{TNR} = \frac{|\hat{J} \cap J|}{|J|}.$$
Here, $I$ is the true set of pure variables, $J = [p] \setminus I$ and $\hat{J} := [p] \setminus \hat{I}$. TPR and TNR quantify the performance of the estimated set $\hat{I}$ but do not reflect how well we recover the partition of pure variables. The pairwise approach of \cite{44} measures how well two partitions are aligned with each other. This approach also allows $\hat{K} = |\hat{I}|$ to be different from $K$. For any pair $1 \leq j < k \leq d := |\hat{I} \cup I|$, define

$$TP_{jk} = \begin{cases} 1 & \text{if } j, k \in I_a \text{ and } j, k \in \hat{I}_b \text{ for some } 1 \leq a \leq K \text{ and } 1 \leq b \leq \hat{K} \end{cases},$$

$$TN_{jk} = \begin{cases} 1 & \text{if } j, k \notin I_a \text{ and } j, k \notin \hat{I}_b \text{ for any } 1 \leq a \leq K \text{ and } 1 \leq b \leq \hat{K} \end{cases},$$

$$FP_{jk} = \begin{cases} 1 & \text{if } j, k \notin I_a \text{ for any } 1 \leq a \leq K \text{ and } j, k \in \hat{I}_b \text{ for some } 1 \leq b \leq \hat{K} \end{cases},$$

$$FN_{jk} = \begin{cases} 1 & \text{if } j, k \in I_a \text{ for some } 1 \leq a \leq K \text{ and } j, k \notin \hat{I}_b \text{ for any } 1 \leq b \leq \hat{K} \end{cases}.$$

and we define

$$TP = \sum_{1 \leq j < k \leq d} TP_{jk}, \quad TN = \sum_{1 \leq j < k \leq d} TN_{jk},$$

$$FP = \sum_{1 \leq j < k \leq d} FP_{jk}, \quad FN = \sum_{1 \leq j < k \leq d} FN_{jk}.$$

We use sensitivity (SN) and specificity (SP) to evaluate the performance of $\hat{I}$, defined as

$$SP = \frac{TN}{TN + FP}, \quad SN = \frac{TP}{TP + FN}. \quad (4.1)$$

To measure the errors of estimating $A_I$ and $\Sigma_Z$ (and indirectly $|\Sigma_E|_I$), we use

$$\text{Err}(A) = \frac{1}{|I|} \left\| \hat{A}_I \hat{A}_I^\top - A_I A_I^\top \right\|_F, \quad \text{Err}(\Sigma_Z) = \frac{1}{K} \left\| \hat{\Sigma}_Z - \Sigma_Z \right\|_F.$$

Methods Throughout the simulation, we consider the PVS procedure given in Algorithm 3 with $q = 2^3$. The tuning parameters $\delta$ and $\mu$ are selected by the cross-validation approach outlined in Appendix B.1. We compare its performance with that of the LOVE algorithm proposed in \cite{9}, developed for factor models in which pure variables in the same partition group have equal factor loadings, an assumption relaxed in this work.

Data generating mechanism We first describe how we generate $\Sigma_Z$ and $\Sigma_E$. The number of latent factor $K$ is set to 10. The matrix $\Sigma_Z$ is generated as $\text{diag}(\Sigma_Z) = 1$ and $[\Sigma_Z]_{ab} = (-1)^{a+b} \rho_{Z}^{|a-b|}$ for any $a, b \in [K]$, where the parameter $\rho_Z \in [0, 1)$ controls the correlation of the latent factor $Z$. $\Sigma_E$ is a diagonal matrix with diagonal elements independently sampled from $\text{Unif}(1,3)$.

\footnote{The same procedure, with $q = \infty$, is not included as it has similar performance but is much more computationally demanding. The supplement contains more simulation results regarding the numerical comparison between $q = 2$ and $q = \infty$.}
We generate $A$ as follows. First, we generate a homogeneous matrix $\tilde{A}$ with the same support as $A$ and satisfies $|A_{i,k}| = e_k$ for all $i \in I_k$, $k \in [K]$ and $\|A_j\|_1 \leq 1$ for all $j \in J$. Specifically, 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 10 groups, each sign pattern is repeated twice. We construct $\tilde{A}$ by independently sampling each entry from $N(0,1)$ and normalize each row to have $\ell_1$-norm equal to 1. Finally, we create a heterogeneous $A$ via $A := D_A \tilde{A}$ using $D_A = \text{diag}(d_1, \ldots, d_p)$ with

$$d_j = \alpha \frac{p v_j^\eta}{\sum_{i=1}^p v_i^\eta} \quad \text{and} \quad v_1, \ldots, v_p \overset{i.i.d.}{\sim} \text{Unif}(1, 2).$$

The parameter $\alpha > 0$ controls the overall scale of $A$ as $p^{-1} \sum_{j=1}^p d_j = \alpha$. Hence the larger $\alpha$, the stronger signal we have. The parameter $\eta \geq 0$ controls the level of heterogeneity of the rows of $A$. The larger $\eta$, the more heterogeneous $A$ is. The case $\eta = 0$ corresponds to the homogeneous $A$, that is, $A = \alpha \tilde{A}$.

Finally, the data matrix $X$ is generated from $X = ZA^\top + E$ where the rows of $Z$ and $E$ are i.i.d. samples from $N(0, \Sigma_Z)$ and $N(0, \Sigma_E)$, respectively. For each setting in the simulation studies below, we repeat generating the data matrix 100 times and report the averaged metrics that we introduced earlier based on 100 repetitions.

### 4.1. Performance of estimating the pure variables

We compare PVS with LOVE in terms of TPR, TNR, SP and SN under various scenarios. To illustrate the dependency on different parameters, we vary $n$, $p$, $\alpha$, $\rho_Z$ and $\eta$ one at a time. We fix $n = 300$, $p = 500$, $\alpha = 2.5$, $\rho_Z = 0.3$ and $\eta = 1$, as a baseline setting, when they are not varied.

**Varying $n$ and $p$ one at a time.** To study the performance of PVS and LOVE under different combinations of $n$ and $p$, we first vary $p \in \{100, 300, 500, 700, 900\}$ with $n = 300$, and then vary $n \in \{100, 300, 500, 700, 900\}$ with $p = 500$. Figure 1 depicts the TPR, TNR, SP and SN metrics of PVS and LOVE in each setting.$^4$

PVS consistently recovers both the set and the partition of pure variables. LOVE fails to control TPR and SN. Its often lower TPR means that some pure variables are not selected. The undercount of pure variables further leads to a low SN. The low performance of LOVE is to be expected when $\eta = 1$ as it has theoretical guarantees only when the rows of $A$ are homogeneous. The performance of PVS improves further as $n$ increases, but it isn’t impacted much as $p$ gets larger, in line with Theorem 6 and Corollary 11.

**Varying $\alpha$ and $\rho_Z$ to change the signal strength.** In this section we study how the performance of PVS changes for various levels of the signal strength. Recall from Section

---

$^4$The invisible lines of LOVE overlap with those of PVS.
Detecting approximate replicate components

2.3 that $\min_i \| A_i \|_2$ and $\lambda_K(\Sigma_Z)$ jointly contribute to the signal strength. Further recall that $\alpha$ controls the overall scale of $A$ while $\rho_Z$ controls $\lambda_K(\Sigma_Z)$. We first set $\rho_Z = 0.3$ with $\alpha$ varying within $\{1.5, 2, 2.5, 3, 3.5\}$, then set $\alpha = 2.5$ and vary $\rho_Z \in \{0, 0.2, 0.4, 0.6, 0.8\}$.

From Figure 2, PVS has near-perfect performance as long as $\alpha > 2$ and $\rho_Z < 0.8$. Its performance gets affected when $\alpha$ is too small or the latent factor $Z$ is too correlated. In particular, PVS tends to miss a few pure variables while still has a perfect TNR and SP.

Varying $\eta$ to change the heterogeneity of $A$. Finally, we investigate the effect of heterogeneity of $A$ on the performance of PVS and LOVE. For this end, we vary $\eta \in \{0, 0.5, 1, 1.5, 2\}$ to change the heterogeneity of $A$ (recall that small $\eta$ means less heterogeneity). Figure 3 shows the SP, SN, TPR and TPN of both LOVE and PVS.

Starting from $\eta = 0$, that is, in the absence of heterogeneity, both LOVE and PVS perfectly recover the pure variables. As the heterogeneity increases, the performance of LOVE drops dramatically whereas PVS is only slightly affected. In the presence of strong heterogeneity, that is, $\eta \geq 1.5$, PVS starts to select less pure variables. The explanation is that some pure variables have too weak a signal to be captured by PVS when $\eta$ is large (recall that the total signal strength over all rows of $A$ is fixed for various of $\eta$).
4.2. Errors of estimating $A_I$ and $\Sigma_Z$

We evaluate PVS in terms of $\text{Err}(A)$ and $\text{Err}(\Sigma_Z)$ as the sample size $n$ changes. We set $p = 300$, $K = 10$, $\rho_Z = 0.3$, $\eta = 1$, $\alpha = 2.5$ and vary $n \in \{100, 300, 500, 700, 900\}$. For each setting, the averaged $\text{Err}(A)$ and $\text{Err}(\Sigma_Z)$ together with their standard deviations are collected in Table 1. As expected from Theorem 14, the errors for both metric decrease as $n$ increases.

| $n$     | $n = 100$ | $n = 300$ | $n = 500$ | $n = 700$ | $n = 900$ |
|---------|-----------|-----------|-----------|-----------|-----------|
| $\text{Err}(A)$ | 0.133 (0.048) | 0.046 (0.019) | 0.023 (0.009) | 0.017 (0.008) | 0.014 (0.007) |
| $\text{Err}(\Sigma_Z)$ | 0.012 (0.006) | 0.005 (0.005) | 0.003 (0.003) | 0.002 (0.003) | 0.004 (0.006) |

Finally, since $\text{Err}(\Sigma_Z)$ relies on $\hat{K} = K$, we remark that for almost all of the settings we have considered so far, Algorithm 3 consistently selects $K$ (we remark that Algorithm 1 consistently selects $K$ as well). To save space, we defer more simulation results regarding the selection of $K$ to the supplement, including some other settings where Algorithm 3 consistently selects $K$ but Algorithm 1 does not. We also defer to the supplement a procedure that screens out variables with weak signal, and its related simulation results.
Detecting approximate replicate components

Acknowledgements

We thank the referees and the AE for their many insightful and helpful suggestions. We are grateful to Boaz Nadler for stimulating our interest in this problem, and for suggesting a preliminary version of the score function used in this work. Bunea and Wegkamp were supported in part by NSF grants DMS-1712709 and DMS-2015195.

Supplementary Material

Supplement to “Detecting approximate replicate components of a high-dimensional random vector with latent structure”: (doi: TBA). The supplementary document includes the proofs and supplementary simulation results.

References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.*, 40(2):1171–1197, 04 2012.

[2] A. Anandkumar, D. P. Foster, D. J. Hsu, S. M. Kakade, and Y. kai Liu. A spectral algorithm for latent dirichlet allocation. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 917–925. Curran Associates, Inc., 2012.

[3] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. volume 23 of *Proceedings of Machine Learning Research*, pages 33.1–33.34, Edinburgh, Scotland, 25–27 Jun 2012. JMLR Workshop and Conference Proceedings.

[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(80):2773–2832, 2014.
[5] T. W. Anderson and H. Rubin. Statistical inference in factor analysis. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry, pages 111–150, Berkeley, Calif., 1956. University of California Press.

[6] S. Arora, R. Ge, Y. Halpern, D. M. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In ICML (2), pages 280–288, 2013.

[7] J. Bai and K. Li. Statistical analysis of factor models of high dimension. Ann. Statist., 40(1):436–465, 02 2012.

[8] X. Bing, F. Bunea, and M. Wegkamp. Inference in interpretable latent factor regression models. arXiv e-prints, pages arXiv–2019, 2019.

[9] X. Bing, F. Bunea, Y. Ning, and M. Wegkamp. Adaptive estimation in structured factor models with applications to overlapping clustering. Ann. Statist., 48(4):2055–2081, 08 2020.

[10] X. Bing, F. Bunea, and M. Wegkamp. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. Bernoulli, 26(3):1765–1796, 08 2020.

[11] X. Bing, F. Bunea, and M. Wegkamp. Optimal estimation of sparse topic models. Journal of Machine Learning Research, 21(177):1–45, 2020.

[12] K. A. Bollen. Structural Equations with Latent Variables. Wiley, 1989.

[13] F. Bunea and L. Xiao. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. Bernoulli, 21(2):1200–1230, 05 2015.

[14] F. Bunea, C. Giraud, X. Luo, M. Royer, and N. Verzelen. Model assisted variable clustering: Minimax-optimal recovery and algorithms. Ann. Statist., 48(1):111–137, 02 2020.

[15] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? J. ACM, 58(3), June 2011. ISSN 0004-5411.

[16] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Sparse and low-rank matrix decompositions. IFAC Proceedings Volumes, 42(10):1493 – 1498, 2009. ISSN 1474-6670. 15th IFAC Symposium on System Identification.

[17] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. Saul, and B. Schölkopf, editors, Advances in Neural Information Processing Systems, volume 16. MIT Press, 2004.

[18] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. K. Saul, and B. Schölkopf, editors, Advances in Neural Information Processing Systems 16, pages 1141–1148. MIT Press, 2004.

[19] D. Donoho, M. Gavish, and I. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. Ann. Statist., 46(4):1742–1778, 08 2018.

[20] N. El Karoui. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. Ann. Appl. Probab., 19(6):2362–2405, 12 2009.

[21] J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a
factor model. Journal of Econometrics, 147(1):186 – 197, 2008. ISSN 0304-4076.

Econometric modelling in finance and risk management: An overview.
[22] J. Fan, Y. Liao, and M. Mincheva. High-dimensional covariance matrix estimation in approximate factor models. Ann. Statist., 39(6):3320–3356, 12 2011.

[23] J. Fan, H. Liu, and W. Wang. Large covariance estimation through elliptical factor models. Ann. Statist., 46(4):1383–1414, 08 2018.

[24] J. Fan, W. Wang, and Y. Zhong. Robust covariance estimation for approximate factor models. Journal of Econometrics, 208(1):5 – 22, 2019. ISSN 0304-4076. Special Issue on Financial Engineering and Risk Management.

[25] F. Han and H. Liu. Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. Bernoulli, 23(1):23–57, 02 2017.

[26] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. IEEE Transactions on Information Theory, 57(11):7221–7234, 2011.

[27] A. Hyvärinen, J. Karhunen, and E. Oja. Independent component analysis, volume 46. John Wiley & Sons, 2004.

[28] A. Jaffe, R. Weiss, B. Nadler, S. Carmi, and Y. Kluger. Learning binary latent variable models: A tensor eigenpair approach. In International Conference on Machine Learning, pages 2196–2205, 2018.

[29] J. Jin, Z. T. Ke, and S. Luo. Estimating network memberships by simplex vertex hunting. arXiv preprint arXiv:1708.07852, 2017.

[30] K. G. Joreskog. Some contributions to maximum likelihood factor analysis. Psychometrika, 32:443–482, 1967.

[31] K. G. Joreskog. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34:183–202, 1969.

[32] K. G. Joreskog. A general method for analysis of covariance structure. Biometrika, 57:239–252, 1970.

[33] K. G. Joreskog. Factor analysis by least squares and maximum likelihood methods. In A. R. K. Enslein and H. S. Wilf, editors, Statistical Methods for Digital Computers III, pages 125–153. Wiley, 1977.

[34] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. Bernoulli, 23(1):110–133, 02 2017.

[35] T. C. Koopmans and O. Reiersol. The identification of structural characteristics. Annals of Mathematical Statistics, 21(2):165–181, 06 1950.

[36] D. N. Lawley. The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh, Section A, 60:64–82, 1940.

[37] D. N. Lawley. Further investigations in factor estimation. Proceedings of the Royal Society of Edinburgh, Section A, 61:176–185, 1941.

[38] D. N. Lawley. The application of the maximum likelihood method to factor analysis. British Journal of Psychology, 33:172–175, 1943.

[39] K. Lounici. High-dimensional covariance matrix estimation with missing observations. Bernoulli, 20(3):1029–1058, 08 2014.

[40] R. P. McDonald. Test theory: a unified treatment. Taylor and Francis, 1999.

[41] L. Thurstone. Multiple factor analysis. Psychological Review, 38(5):406–427, 1931.

[42] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices, page
210–268. Cambridge University Press, 2012.

[43] M. Wegkamp and Y. Zhao. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli*, 22(2):1184–1226, 05 2016.

[44] C. Wiwie, J. Baumbach, and R. Röttger. Comparing the performance of biomedical clustering methods. *Nature methods*, 12(11):1033–1038, 2015.

[45] I. Yalcin and Y. Amemiya. Nonlinear factor analysis as a statistical method. *Statist. Sci.*, 16(3):275–294, 08 2001.