Bernoulli **29**(2), 2023, 1368–1391
https://doi.org/10.3150/22-BEJ1502

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

XIN BING[1,a], FLORENTINA BUNEA[2,b] and MARTEN WEGKAMP[2,3,c]

[1]*Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada,* [a]*xin.bing@utoronto.ca*
[2]*Department of Statistics and Data Science, Cornell University, Ithaca, New York, USA,* [b]*fb238@cornell.edu*
[3]*Department of Mathematics, Cornell University, Ithaca, New York, USA,* [c]*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 $\mathcal{H}$ of its row index set H. Both H and $\mathcal{H}$ 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.

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 [31–34,37–39], 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

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

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 Σ_Z is the covariance matrix of Z and Σ_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, [1,16–18,22–25,27], 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 [2–10,12,20,23,24,28,29].

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 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), 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 $\operatorname{rank}(A) = K$,

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

and Z is standardized such that

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

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

$$H := \big\{ i \in \{1,\dots,p\} : \ A_{i\bullet} \mathbin{/\!\!/} A_{j\bullet} \ \text{ for some } j \in \{1,\dots,p\} \setminus \{i\} \big\}$$

of A is non-empty and $\operatorname{rank}(A_{H\bullet}) = K$.

The partition of H consists of disjoint sets $H_1,\dots,H_G$, and all indices $i,j \in H_k$ correspond to parallel rows $A_{i\bullet}$ and $A_{j\bullet}$, for each $k \in [G] := \{1,2,\dots,G\}$. The assumption $H \neq \varnothing$ and $\operatorname{rank}(A_{H\bullet}) = K$ 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 $\mathcal{H} = \{H_1,\dots,H_G\}$, and K from the correlation matrix $R = [\operatorname{Corr}(X_i,X_j)]_{i,j\in[p]}$. 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 $R_{i\bullet\setminus\{i,j\}}$ and $R_{j\bullet\setminus\{i,j\}}$, each defined by leaving out, respectively, the ith and jth entries from the rows $R_{i\bullet}$ and $R_{j\bullet}$, are parallel in $\mathbb{R}^{p-2}$ if and only if $A_{i\bullet}$ and $A_{j\bullet}$ are parallel. This realization, combined with the fact that two non-zero vectors v,w are parallel if and only if

$$\min_{\|(\alpha,\beta)\|_r=1} \|\alpha v + \beta w\|_q = 0$$

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

$$S_{q,r}(i,j) = (p-2)^{-1/q} \min_{\|(\alpha,\beta)\|_r=1} \big\|\alpha R_{i\bullet\setminus\{i,j\}} + \beta R_{j\bullet\setminus\{i,j\}}\big\|_q, \ \text{for any } i,j \in [p], \tag{1.2}$$

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 $\mathcal{H}$,

$$S_{q,r}(i,j) = 0 \quad \Longleftrightarrow \quad i,j \in H_a \quad \text{for some } a \in [G] \tag{1.3}$$

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 $\mathbf{e}_k \in \mathbb{R}^K$, and the p rows can be partitioned in K groups of replicates of $\mathbf{e}_k$, 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

$$\|\Sigma_{i\bullet\setminus\{i,j\}} - \Sigma_{j\bullet\setminus\{i,j\}}\|_\infty$$

of the differences between rows $\Sigma_{i\bullet\setminus\{i,j\}}$ and $\Sigma_{j\bullet\setminus\{i,j\}}$ of the covariance matrix $\Sigma = \mathrm{Cov}(X)$, 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 (α, β) 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 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 $\widehat{S}_{q,r}$ of the criterion $S_{q,r}$ in (1.2), by simply replacing R by the empirical correlation matrix $\widehat{R}$. From the tight, in probability, bound

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

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 $\widehat{H}$ of all pairs (i,j) with $\widehat{S}_{q,r}(i,j) \le 2\delta_n$. In this paper, we derive the order of magnitude of δ_n under the assumption that X is sub-Gaussian.

In Corollary 7 of Section 2.3 we show that $\widehat{H}$ and its partition $\widehat{\mathcal{H}}$ consistently estimates H and $\mathcal{H}$, respectively. After estimating H and $\mathcal{H}$, Section 2.2 devises the estimation of K, exploiting the fact that $\Sigma - \Sigma_E = A\Sigma_Z A^\top$ 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 $A_{I\bullet}$ (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 $i \ne j$ such that $A_{ik} \ne 0$, $A_{jk} \ne 0$ 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 $\mathcal{I} = \{I_1, \ldots, I_K\}$ 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 Σ_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 Σ_E are known [8].

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 [36,41], and is routinely used in educational and psychological testing, where the latent variables are viewed as aptitudes or psychological states [5,13,42]. 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 $1 \le k \le K$. 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 [19] and network analysis (see, for instance, [30]), 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 [6,9], 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 [8] 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 representive variables is given in Lemma 9 of Section 3.1 based on successively maximizing certain Schur complements of the low-rank matrix $\mathrm{Cov}(AZ) = A\Sigma_Z A^\top$. 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 $\max_{i \in I_k} \|A_{i\bullet}\|_1$, 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 $\widehat{H}$ and $\widehat{K}$ under Assumption 2′. We show, in Corollary 11 of Section 3.3, that $\widehat{H}$ 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 $\widehat{K}$ 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 $\widehat{H}$. 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 [8], 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 $\widehat{I}$ of the set I, which is one of the focus points of this work. Once an estimator $\widehat{I}$ is found, [8] proposed the following strategy: estimate A by concatenating estimators $\widehat{A}_{\widehat{I}\bullet}$, $\widehat{A}_{\widehat{J}\bullet}$ of the sub-matrices $A_{I\bullet}$ and $A_{J\bullet}$, where $J = [p] \setminus I$, $\widehat{J} = [p] \setminus \widehat{I}$, and for any matrix M and row index set S, we denote by $M_{S\bullet}$ the sub-matrix formed from M by retaining S rows. The authors showed that the resulting $\widehat{A}$ 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 $\widehat{A}_{\widehat{I}\bullet}$ is a matrix consisting in canonical basis vectors. While we adopt the same strategy as in [8] for the overall estimation of A, we complement it by providing, in Section 3.4.1, an estimator $\widehat{A}_{\widehat{I}\bullet}$ 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. Appendix: Proofs and simulations

All proofs are collected in the Appendix A [11]. 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 [15,35,40,43], the operator norm concentration inequalities of the sample correlation matrix is relatively less explored. [21] studied the asymptotic behaviour of the limiting spectral distribution of the sample correlation matrix when $p/n \to (0, \infty)$. [26,44] prove a similar result for Kendall's tau sample correlation matrix.

Appendix B contains all 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.

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 ℓ_q-norm as $\|v\|_q$ for $0 \le q \le \infty$. We denote the k-th canonical unit vector in $\mathbb{R}^d$ by $\mathbf{e}_k$, 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, $M_{S_1 \bullet}$ ($M_{\bullet S_2}$) stands for the whole rows (columns) of M in S_1 (S_2). We use $\|M\|_{\mathrm{op}}$, $\|M\|_{\mathrm{F}}$ 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 $i \overset{\mathrm{S}}{\sim} j$ means that there exists $k \in [K]$ such that $i, j \in S_k$. Its complement $i \overset{\mathrm{S}}{\not\sim} j$ means i and j do not simultaneously belong to any S_k. Let $\Sigma := \mathrm{Cov}(X)$ be the covariance matrix of the random vector $X \in \mathbb{R}^p$ with diagonal matrix $D_\Sigma = \mathrm{diag}(\Sigma_{11}, \ldots, \Sigma_{pp})$ 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

$$\Sigma \;=\; A \Sigma_Z A^\top + \Sigma_E,$$

and consequently

$$R \;=\; D_\Sigma^{-1/2} \Sigma D_\Sigma^{-1/2} \;:=\; B \Sigma_Z B^\top + \Gamma, \tag{2.1}$$

where

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

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.1.1. We assume $\min_{i \in [p]} \|A_{i\bullet}\|_2 > 0$. Otherwise, we remove all zero rows of A in the pre-screening step described in Appendix B.6. The notation $i \overset{\mathrm{H}}{\sim} j$ means that both $i, j \in H_k$ for some $k \in [G]$. For any $i, j \in [p]$, let $R_{i, \setminus \{i,j\}} \in \mathbb{R}^{p-2}$ denote the ith row of R, with the entries in the ith and jth columns removed.

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

$$i \overset{\mathrm{H}}{\sim} j \quad \Longleftrightarrow \quad A_{i\bullet} \mathbin{/\!\!/} A_{j\bullet} \quad \Longleftrightarrow \quad R_{i,\setminus\{i,j\}} \mathbin{/\!\!/} R_{j,\setminus\{i,j\}}. \tag{2.2}$$

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.

To make use of Proposition 1, first recall that for any two non-zero vectors v and w, we have

$$v \mathbin{/\!\!/} w \quad \Longleftrightarrow \quad \min_{\|(a,b)\|_r=1} \|av + bw\|_q = 0$$

for any integers $0 \le q \le \infty$ and $0 < r \le \infty$, 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:

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

for any $i \neq j$, $0 < r \le \infty$ and $0 \le q \le \infty$. 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) $i \overset{\mathrm{H}}{\sim} j \iff S_{q,r}(i,j) = 0$ *for any* $0 < r \le \infty$ *and* $0 \le q \le \infty$.
(2) *For any fixed* $i, j \in [p]$, $S_{q,r}(i,j)$ *defined in (2.3) satisfies*
  (i) $S_{q,t}(i,j) \ge S_{q,r}(i,j)$ *for all* $0 < r \le t \le \infty$;
  (ii) $S_{t,r}(i,j) \ge S_{q,r}(i,j)$ *for all* $t \ge q$.

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 $i \overset{\mathrm{H}}{\not\sim} j$. Part *(2)(i)* of Proposition 2 immediately suggests taking $r = \infty$ and considering the class of score functions

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

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

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

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

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

[1] The partition $\mathcal{H}$ is unique up to a group permutation.

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

$$V_{ii}^{(ij)} := [R_{i,\setminus\{i,j\}}]^\top R_{i,\setminus\{i,j\}}, \quad V_{ij}^{(ij)} := [R_{i,\setminus\{i,j\}}]^\top R_{j,\setminus\{i,j\}} \qquad \forall i,j \in [p].$$

Proposition 3. *The score function $S_2(i,j)$ defined in (2.5) satisfies*

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

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

$$i \overset{\mathrm{H}}{\sim} j \qquad \Longleftrightarrow \qquad S_2(i,j) = 0. \tag{2.6}$$

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 $K = \mathrm{rank}(M_{HH})$, 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 $G = |\mathcal{H}|$ are unique. Moreover, if additionally*

$$\|R_{i,\setminus\{i,j\}}\|_q > 0, \qquad \forall\, i \overset{\mathrm{H}}{\sim} j, \tag{2.7}$$

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 $i \overset{\mathrm{H}}{\sim} j$. 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 $AZ = \widetilde{A}\widetilde{Z}$, where $\widetilde{A} = AQD$ has the same index set H as A, and the covariance matrix $\mathrm{Cov}(\widetilde{Z})$ of $\widetilde{Z} := D^{-1}Q^{-1}Z$ is positive definite and satisfies $\mathrm{diag}[\mathrm{Cov}(\widetilde{Z})] = \mathbf{1}$, as $\mathrm{Cov}(Z)$. 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 $\boldsymbol{X}$. We write the sample covariance matrix as

$$\widehat{\Sigma} = \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}$$

and denote the sample correlation matrix by $\widehat{R}$, with entries

$$\widehat{R}_{ij} = \widehat{\Sigma}_{ij}/\sqrt{\widehat{\Sigma}_{ii}\widehat{\Sigma}_{jj}}, \qquad \forall\, i,j \in [p].$$

Our estimation procedure is the sample level analogue of Theorem 4 of Section 2.1 above: we first estimate the parallel row index set H and its partition $\mathcal{H}$, then we estimate K. The statistical guarantees of these estimates are provided in Section 2.3.

Recall from part (2)(i) of Proposition 2 that we use $S_q(i,j) := S_{q,\infty}(i,j)$ 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

$$\widehat{S}_q(i,j) = (p-2)^{-1/q} \min_{\|(a,b)\|_\infty = 1} \left\| a\widehat{R}_{i,\setminus\{i,j\}} + b\widehat{R}_{j,\setminus\{i,j\}} \right\|_q \tag{2.8}$$

for each $i,j \in [p]$ and $i \neq j$. In particular, Proposition 3 implies that $\widehat{S}_2(i,j)$ has the closed form

$$\widehat{S}_2(i,j) = \left[\frac{\widehat{V}_{ii}^{(ij)} \wedge \widehat{V}_{jj}^{(ij)}}{p-2} \left(1 - \frac{[\widehat{V}_{ij}^{(ij)}]^2}{\widehat{V}_{ii}^{(ij)}\widehat{V}_{jj}^{(ij)}} \right) \right]^{1/2} \tag{2.9}$$

with $\widehat{V}_{ij}^{(ij)} = [\widehat{R}_{i,\setminus\{i,j\}}]^\top \widehat{R}_{j,\setminus\{i,j\}}$ for all $i,j \in [p]$.

Remark 1. For any pair (i,j) and any $q \geq 1$, the criterion $\widehat{S}_q(i,j)$ in (2.8) can be computed by solving two convex optimization problems because $\widehat{S}_q(i,j)$ is equal to

$$(p-2)^{-1/q} \min \left\{ \min_{|a|\leq 1} \|a\widehat{R}_{i,\setminus\{i,j\}} + \widehat{R}_{j,\setminus\{i,j\}}\|_q, \min_{|b|\leq 1} \|\widehat{R}_{i,\setminus\{i,j\}} + b\widehat{R}_{j,\setminus\{i,j\}}\|_q \right\}.$$

In particular, for $q = \infty$, computation of $\widehat{S}_\infty(i,j)$ 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 $\widehat{S}_q(i,j)$ below the threshold level 2δ. It returns not only the estimated index set $\widehat{H}$, but also its partition $\widehat{\mathcal{H}} := \{\widehat{H}_1, \ldots, \widehat{H}_{\widehat{G}}\}$. Algorithm 1 requires one single tuning parameter δ, with an explicit rate stated in Section 2.3. A fully data-driven criterion of selecting δ 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 δ.

Lemma 5. *Let $\widehat{H}(\delta)$ be the estimated set of parallel rows from Algorithm 1. Then*

$$|\widehat{H}(\delta)| \;\leq\; |\widehat{H}(\delta')| \qquad \forall\, \delta \leq \delta'.$$

Proof. For a given $\delta > 0$, suppose $i \in \widehat{H}(\delta)$. Then, from Algorithm 1, there exists some $j \neq i$ such that $\widehat{S}_q(i,j) \leq 2\delta$. This implies $\widehat{S}_q(i,j) \leq 2\delta'$ for any $\delta' \geq \delta$. Hence, $i \in \widehat{H}(\delta')$, as desired. □

Since $\widehat{G}$ 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 $\widehat{\mathcal{H}} = \{\widehat{H}_1, \ldots, \widehat{H}_{\widehat{G}}\}$ 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 $\ell_k \in H_k$ for each $1 \leq k \leq G$.

Algorithm 1 Estimate the parallel row index set H by $\widehat{H}$ and its partition $\mathcal{H}$ by $\widehat{\mathcal{H}}$

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

procedure PARALLEL($\widehat{R}$, δ)
  $\widehat{\mathcal{H}} \leftarrow \varnothing$.
  for $i = 1, \ldots, p-1$ **do**
    for $j = i+1, \ldots, p$ **do**
      Compute $\widehat{S}_q(i,j)$ by solving (2.8)
      if $\widehat{S}_q(i,j) \leq 2\delta$ **then**
        $\widehat{\mathcal{H}} \leftarrow$ MERGE($\{i,j\}$, $\widehat{\mathcal{H}}$)
  return $\widehat{\mathcal{H}}$, $\widehat{G} = |\widehat{\mathcal{H}}|$ and $\widehat{H} = \cup_k \widehat{\mathcal{H}}_k$

function MERGE(S, $\widehat{\mathcal{H}}$)
  Add = True
  for all $g \in \widehat{\mathcal{H}}$ **do** ▷ $\widehat{\mathcal{H}}$ is a collection of sets
    if $g \cap S \neq \varnothing$ **then**
      $g \leftarrow g \cup S$ ▷ Replace $g \in \widehat{\mathcal{H}}$ by $g \cup S$
      Add = False
  if Add **then**
    $\widehat{\mathcal{H}} = \widehat{\mathcal{H}} \cup \{S\}$ ▷ add S in $\widehat{\mathcal{H}}$
  return $\widehat{\mathcal{H}}$

- For each $k \in [\widehat{G}]$, we select one representative variable index from $\widehat{H}_k$ as

$$\widehat{\ell}_k := \arg\max_{i\in\widehat{H}_k} \big\|\widehat{R}_{i,\setminus\{i\}}\big\|_q, \tag{2.10}$$

  and create the set of representative indices

$$\widehat{L} := \big\{\widehat{\ell}_1, \ldots, \widehat{\ell}_{\widehat{G}}\big\}. \tag{2.11}$$

- Next, motivated by (A.2) and (A.3) in the proof of Theorem 4, we propose to estimate the submatrix $M_{\widehat{L}\widehat{L}}$ of $M := B\Sigma_Z B^\top$ by

$$\widehat{M}_{ij} = \widehat{R}_{ij}, \qquad \forall\, i,j \in \widehat{H},\ i \neq j, \tag{2.12}$$

  and

$$\widehat{M}_{ii} = |\widehat{R}_{i\widehat{j}}|\,\frac{\|\widehat{R}_{i,\setminus\{i,\widehat{j}\}}\|_q}{\|\widehat{R}_{\widehat{j},\setminus\{i,\widehat{j}\}}\|_q}, \qquad \forall i \in \widehat{H}_k,\ k \in [\widehat{G}], \tag{2.13}$$

$$\widehat{j} = \arg\min_{\ell\in\widehat{H}_k\setminus\{i\}} \widehat{S}_q(i,\ell).$$

  Instead of choosing $\widehat{j}$ for each $i \in \widehat{H}_k$ as above, we could alternatively estimate $\widehat{M}_{ii}$ via (2.13) by averaging over $j \in \widehat{H}_k \setminus \{i\}$. Our numerical experiments indicate that these two procedures have similar performance.

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

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

for some tuning parameter $\mu > 0$.

2.3. Statistical guarantees for $\widehat{H}$, $\widehat{\mathcal{H}}$ and $\widehat{K}$

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 γ–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 γ is called the sub-Gaussian constant. In this work, we treat γ as some absolute constant and write $c = c(\gamma)$ and $C = C(\gamma)$ for numerical constants depending on γ only.

Assumption 3. There exists a constant γ such that $\Sigma^{-1/2}X$ is γ–sub-Gaussian.[2]

The only tuning parameter in Algorithm 1 is δ with theoretical order given by

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

for some constant $c(\gamma)$ depending on γ only. δ_n is a key quantity that controls the deviation $\widehat{R}$ 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

$$\mathcal{E} := \left\{\max_{1\leq i,j\leq p} |\widehat{R}_{ij} - R_{ij}| \leq \delta_n\right\} \tag{2.16}$$

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 $\widehat{S}_q(i,j) - S_q(i,j)$ 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 $\widehat{S}_q$ used in Algorithm 1. We found in our simulations that $\widehat{S}_q$ 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 $1 \leq q \leq \infty$,*

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

If, in addition, model (1.1) and Assumption 1 hold, one has

$$\begin{aligned} \widehat{S}_q(i,j) \ &\leq \ 2\delta_n, && \text{for all} \quad i \overset{\mathrm{H}}{\sim} j; \\ \widehat{S}_q(i,j) \ &\geq \ \max\left\{0,\, S_q(i,j) - 2\delta_n\right\}, && \text{for all} \quad i,j \in [p] \end{aligned}$$

for any $1 \leq q \leq \infty$.

[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 γ_Z and γ_E, respectively, then $\Sigma^{-1/2}X$ is γ–sub-Gaussian with $\gamma = \max\{\gamma_Z, \gamma_E\}$.

Algorithm 1 with $\delta = \delta_n$ selects those indices $i, j \in \widehat{H}$ for which $\widehat{S}_q(i,j) \le 2\delta_n$. On the event $\mathcal{E}$, Theorem 6 tells us that

$$H \subseteq \widehat{H} \subseteq \left\{ i \in [p] : \ S_q(i,j) \le 4\delta_n \text{ for some } j \ne i \right\}. \tag{2.17}$$

Hence, with probability at least $1 - 4/(p \vee n)$, $\widehat{H}$ 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 $\widehat{\mathcal{H}}$, however, may not include all the groups that make up the partition $\mathcal{H}$. 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

$$S_q(i,j) > 4\delta_n, \quad \text{for all } i \overset{\mathrm{H}}{\not\sim} j \tag{2.18}$$

is a sufficient condition for consistent estimation of both H and $\mathcal{H}$, as summarized in the following corollary.

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

The following proposition states the explicit rate of the tuning parameter μ under which we provide theoretical guarantees for $\widehat{K}$ as an estimator of K, for any $q \ge 1$. For any $L = \{\ell_1, \ldots, \ell_G\}$ with $\ell_k \in H_k$ for all $k \in [G]$, define

$$\underline{c}(L) := \lambda_K(B_{L\bullet}\Sigma_Z B_{L\bullet}^\top), \qquad \overline{c}(L) := \lambda_1(B_{L\bullet}\Sigma_Z B_{L\bullet}^\top). \tag{2.19}$$

Theorem 8. *Under model (1.1), Assumptions 1 & 3 and condition (2.18), suppose there exists some constant $c > 0$ such that*

$$\|R_{i,\setminus\{i,j\}}\|_q \ge c\,(p-2)^{1/q}, \qquad \forall\, i \overset{\mathrm{H}}{\sim} j. \tag{2.20}$$

For $\mu = C(\sqrt{\widehat{G}\delta_n^2} + \widehat{G}\delta_n^2)$ in (2.14), and for some $C > 0$ depending on $\overline{c}(\widehat{L})$, we have

$$\mathbb{P}\{\widehat{K} \le K\} \ge 1 - c'(n \vee p)^{-c''}.$$

If, additionally, $\max_L [\overline{c}(L)/\underline{c}(L)] G\delta_n^2 \le c_0$ for some sufficiently small constant $c_0 > 0$, we further have

$$\mathbb{P}\{\widehat{K} = K\} \ge 1 - c'(n \vee p)^{-c''}.$$

The proof of Theorem 8 is deferred to Appendix A.1.6. We first prove that $\widehat{K} \le K$ always holds on the event $\mathcal{E}$ with the specified choice of μ. Proving consistency $\widehat{K} = K$ requires μ to be sufficiently small so that $\mu < \underline{c}(\widehat{L})$, which is guaranteed by $\max_L [\overline{c}(L)/\underline{c}(L)] G\delta_n^2 = o(1)$. When both $\underline{c}(L)$ and $\overline{c}(L)$ are bounded away from 0 and ∞, consistency only requires $G\delta_n^2 = o(1)$, that is, we allow $G = |\mathcal{H}|$ to grow but no faster than $O(n/\log(n \vee p))$. The data-driven choice of the leading constant of μ 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 $\min_{i \overset{\mathrm{H}}{\sim} j} \max_{\ell \ne i, \ell \ne j} |R_{i\ell}| \ge c$ for $q = \infty$. 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

$$J_1 := \left\{ j \in J : A_{j\bullet} \mathbin{/\!\!/} A_{\ell\bullet} \text{ for some } \ell \in J \setminus \{j\} \right\}, \tag{3.1}$$

and its partition $\mathcal{J}_1 := \{J_1^1, \ldots, J_1^N\}$. 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 $J_1 = \varnothing$ and $J_1 \neq \varnothing$, by simply comparing G with K.

If the only parallel rows in A correspond to pure variables ($J_1 = \varnothing$ or, equivalently, $G = K$), then I is identifiable as an immediate consequence of Theorem 4. If $J_1 \neq \varnothing$ (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 $J_1 \neq \varnothing$. 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 $\xi_k := \max_{i \in I_k} \|A_{i\bullet}\|_1$ be the largest loading in group I_k, in absolute value. We have

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

We note, trivially, that the assumption is automatically met when $J_1 = \varnothing$. 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 $\|A_{i\bullet}\|_1 = \xi$ for all $i \in I$ and $\|A_{j\bullet}\|_1 \leq \xi$ for all $j \in J_1$, which reduces it to the condition employed in [8] 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* $S_1 = \varnothing$ *satisfying* $i_a \in I_{\pi(a)}$ *for all* $1 \leq a \leq k-1$ *and some permutation* $\pi : [K] \to [K]$. *Then one has*

$$i_k := \arg\max_{j \in H} \Theta_{jj|S_k} \in I_{\pi(k)}, \qquad \textit{for all } 1 \leq k \leq K, \tag{3.2}$$

where $\Theta_{jj|S_k} = \Theta_{jj} - \Theta_{jS_k}^\top \Theta_{S_k S_k}^{-1} \Theta_{S_k j}$ *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 $W_H := X_H - E_H = A_{H\bullet} Z$ for $H \subseteq [p]$ and observe that Θ is the degenerate (rank K) covariance matrix of $W \in \mathbb{R}^p$. To add intuition, if Z has a multivariate normal distribution, then $\Theta_{H^c H^c|H} = \mathrm{Cov}(W_{H^c}|W_H)$. In this case, display (3.2) in Lemma 9 becomes

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

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

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

we see that maximizing the above conditional variance retains more information (for goal (i)), while avoiding linear dependence by reducing $\mathrm{Corr}(W_{i_1}, W_{i_2})$ (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 Θ 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 $\mathcal{I}$ 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 $\mathcal{I}$ 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 $\text{rank}(A) = K$ 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 Σ_E are known. From this perspective, when I and Σ_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:

$$\frac{\sum_{a \neq k} |A_{ja}|}{\|A_{j\bullet}\|_1} = 1 - \frac{|A_{jk}|}{\|A_{j\bullet}\|_1} \leq \varepsilon^2. \tag{3.3}$$

As we can see, when ε is small, the jth 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 kth factor. We show in Appendix A.5.1 that, for this i_k and j, the score in (2.5) satisfies $S_2(i_k, j) \leq 2\varepsilon$. By slight abuse of notation, we write $I_k = \{i_k, j\}$ and provided that $S_2(\ell, \ell') > 2\epsilon$ for all $\ell \overset{\downarrow}{\not\sim} \ell'$ 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 $\delta \geq 2\varepsilon$.

2. Discussion of Assumption 4. This assumption is only active when $J_1 \neq \varnothing$. If $J_1 \neq \varnothing$ 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 $\widehat{\mathcal{H}} = \{\widehat{H}_1, \ldots, \widehat{H}_{\widehat{G}}\}$ and $\widehat{K}$. If $\widehat{K} = \widehat{G}$, no further action is taken; if $\widehat{K} < \widehat{G}$, we add the pruning step stated below, based on the sample analogue of Lemma 9.

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

- Estimate $\Gamma_{\widehat{H}\widehat{H}}$ by $\widehat{\Gamma}_{ij} = 0$ for all $i \neq j$ and

$$\widehat{\Gamma}_{ii} = 1 - \widehat{M}_{ii}, \qquad \forall i \in \widehat{H} \tag{3.4}$$

  where $\widehat{M}_{ii}$ is obtained from (2.13) with $q = 2$.

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

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

procedure PRUNING($\widehat{\Sigma}$, $\widehat{R}$, $\widehat{\mathcal{H}}$, r)
  Compute $\widehat{\Theta}_{\widehat{H}\widehat{H}}$ from (3.5)
  Set $S = \varnothing$
  for $k = 1, \ldots, r$ **do**
    Compute i_k from (3.7) and add $i_k \in S$
  return $\widehat{\mathcal{I}}$ obtained from (3.8)

- Set $\Theta := A\Sigma_Z A^\top$ and, in view of (2.1), estimate $\Theta_{\widehat{H}\widehat{H}}$ by

$$\widehat{\Theta}_{\widehat{H}\widehat{H}} = \left[\widehat{\Sigma} - \widehat{\Sigma}_E\right]_{\widehat{H}\widehat{H}} = \left[\widehat{\Sigma} - \widehat{D}_{\widehat{\Sigma}}^{1/2}\widehat{\Gamma}\widehat{D}_{\widehat{\Sigma}}^{1/2}\right]_{\widehat{H}\widehat{H}} \tag{3.5}$$

  with $\widehat{D}_{\widehat{\Sigma}} = \mathrm{diag}(\widehat{\Sigma}_{11}, \ldots, \widehat{\Sigma}_{pp})$.
- For any set $S \subseteq \widehat{H}$ and $S^c = \widehat{H} \setminus S$, write the Schur complement of $\widehat{\Theta}_{SS}$ of $\widehat{\Theta}_{\widehat{H}\widehat{H}}$ as

$$\widehat{\Theta}_{S^cS^c|S} = \widehat{\Theta}_{S^cS^c} - \widehat{\Theta}_{S^cS}\widehat{\Theta}_{SS}^{-}\widehat{\Theta}_{SS^c} \tag{3.6}$$

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

$$i_k = \arg\max_{j\in\widehat{H}} \widehat{\Theta}_{jj|S_k}. \tag{3.7}$$

  When there are ties, arbitrarily pick one of the maximizers.
- The final estimate of $\mathcal{I}$ is defined as

$$\widehat{\mathcal{I}} = \left\{\{\widehat{H}_k\}_{1\le k\le r} : \text{there exists } a \in [r] \text{ such that } i_a \in \widehat{H}_k\right\}. \tag{3.8}$$

We note that Algorithm 2 can take any $1 \le r \le K$ 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 $\widehat{K}$ 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 $\widehat{\mathcal{I}}$ in the next section.

3.3. Statistical guarantees for the estimated pure variables index set

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

As shown in (2.17) of Section 2.3, on the event $\mathcal{E}$ in (2.16), $\widehat{H}$ 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 $\widehat{H}$ and its partition $\widehat{\mathcal{H}}$ 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:

$$S_q(i,j) > 4\delta_n \quad \text{for all } i \overset{1}{\not\sim} j,\ i \in I. \tag{3.9}$$

Note that we still allow

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

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 J in the new set

$$\bar{J}_1 := \left\{ j \in J :\ S_q(j,\ell) \le 4\delta_n \ \text{ for some } \ell \in J \setminus \{j\} \right\}, \tag{3.10}$$

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 $\widehat{H}$ and $\widehat{\mathcal{H}}$, in Corollary 11, and for $\widehat{K}$ 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 $\widehat{G}$ and $\widehat{\mathcal{H}} = \{\widehat{H}_1, \ldots, \widehat{H}_{\widehat{G}}\}$ of Algorithm 1, applied with $\delta = \delta_n$, satisfy*

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

Proof. The result is a direct consequence of Theorem 6 and (3.9) in conjunction with the fact that $\widehat{\mathcal{H}}$ 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 $B\Sigma_Z B^\top$, that (3.9) holds under the following mild separation condition between pure and non-pure variables,

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

We proceed to revisit and analyze $\widehat{K}$, constructed in Section 2.2, with $q = 2$. Theorem 12 below states the explicit rate of the tuning parameter μ 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

$$c_{b,I} := \min_{j \in I \cup \bar{J}_1} \|B_{j\bullet}\|_2^2, \quad \bar{c}_z := \lambda_1(\Sigma_Z), \quad c_z := \lambda_K(\Sigma_Z), \tag{3.11}$$

and

$$c_r := \min_{i \ne j} \frac{1}{p-2} \lambda_K \left(B_{\setminus\{i,j\}\bullet} \Sigma_Z B^\top_{\setminus\{i,j\}\bullet} \right). \tag{3.12}$$

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* $0 < c \le C < \infty$ *such that*

$$\min(c_{b,I}, c_z, c_r) > c, \qquad \bar{c}_z \le C \tag{3.13}$$

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*

$$\lim_{n\to\infty} \mathbb{P}\{\widehat{K} \le K\} = 1.$$

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

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

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

We first prove that $\widehat{K} \le K$ 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 $\{\widehat{G} = \widehat{K}\}$ implies the event $\{\widehat{G} = K\}$. In this case, our procedure stops and the output of Algorithm B.1 is $\widehat{\mathcal{H}}$ from Algorithm 1. On the other hand, if $\widehat{K} < \widehat{G}$, as explained in Section 3.2, Algorithm B.1 uses the pruning step from Algorithm 2 to estimate I and its partition.

Proving consistency $\widehat{K} = K$ requires μ 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 $1 \le r \le K$, and, in particular, the output of Algorithm B.1 if r is set to $\widehat{K}$. Recall that ξ_k is defined in Assumption 4.

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

$$\lim_{n\to\infty} K\delta_n^2 = 0 \tag{3.14}$$

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

$$\max_{1\le k\le K} \xi_k \le C_0 \min_{1\le k\le K} \xi_k, \tag{3.15}$$

$$\max_{j\in\bar{J}_1} \left(\sum_{k=1}^{K} \frac{|A_{jk}|}{\xi_k}\right)^2 < 1 - \varepsilon. \tag{3.16}$$

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

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

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

As already mentioned, the statement of Theorem 13 is formulated in terms of r groups, for any $1 \le r \le K$. 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 $1 \le r \le K$, 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 $\widehat{\Theta}_{jj|S_k}$ and its population-level counterpart, uniformly for all $j \in \widehat{H}$, all S_k selected from (3.7) and all $1 \le k \le r$. The difficulty is further elevated by the fact that $\bar{J}_1 \neq \varnothing$. 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 $\widehat{M}_{\widehat{H}\widehat{H}} - M_{\widehat{H}\widehat{H}}$, the operator norm of $\widehat{M}_{S_k S_k} - M_{S_k S_k}$ and the quadratic form $\widehat{M}_{iS_k}^\top \widehat{M}_{S_k S_k}^{-1} \widehat{M}_{S_k i}$, 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 $K\delta_n^2 \le c$, 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 $1 - c'(p \vee n)^{-c''}$ 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, $K \log(p \vee n) = o(n)$. Condition $|\bar{J}_1|\delta_n = o(1)$ in Theorem 12 allows the size of the index set corresponding to near-parallel rows in A to grow, but slower than $\sqrt{n/\log(n \vee p)}$.

Conditions (3.15) and (3.16) are only needed when $\bar{J}_1 \neq \varnothing$. 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 $\max_k \max_{i \in I_k} \Sigma_{ii} \le C \min_k \max_{i \in I_k} \Sigma_{ii}$ coupled with (3.13). Finally, condition (3.13) is a mild regularity condition on the matrix B and Σ_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 [8], we first estimate $B_{I\bullet} := D_\Sigma^{-1/2} A_{I\bullet}$ and $B_{J\bullet} := D_\Sigma^{-1/2} A_{J\bullet}$ by the estimators $B_{\widehat{I}\bullet}$ and $B_{\widehat{J}\bullet}$ presented below. Once this is done, A is easily estimated by $\widehat{A}$ obtained by concatenating the sub-matrices

$$\widehat{A}_{\widehat{I}\bullet} = \widehat{D}_{\widehat{\Sigma}}^{1/2} \widehat{B}_{\widehat{I}\bullet}, \quad \widehat{A}_{\widehat{J}\bullet} = \widehat{D}_{\widehat{\Sigma}}^{1/2} \widehat{B}_{\widehat{J}\bullet}, \qquad \text{with} \quad \widehat{D}_{\widehat{\Sigma}} = \operatorname{diag}(\widehat{\Sigma}_{11}, \cdots, \widehat{\Sigma}_{pp}). \tag{3.17}$$

3.4.1. Estimation of $B_{I\bullet}$, Σ_Z and $A_{I\bullet}$.

Since Assumption 2′ and $\operatorname{diag}(\Sigma_Z) = \mathbf{1}$ imply that $M_{ii} = B_{ik}^2$ for any $i \in I_k$, and given the estimated partition of the pure variables $\widehat{\mathcal{I}} = \{\widehat{I}_1, \ldots, \widehat{I}_{\widehat{K}}\}$, we propose to estimate $B_{I\bullet}$ by

$$\widehat{B}_{ik}^2 = \widehat{M}_{ii}, \quad \forall i \in \widehat{I}_k,\ k \in [\widehat{K}] \tag{3.18}$$

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

$$\operatorname{sgn}(\widehat{B}_{i_k k}) = 1, \qquad \operatorname{sgn}(\widehat{B}_{jk}) = \operatorname{sgn}(\widehat{R}_{i_k j}), \qquad \forall j \in \widehat{I}_k \setminus \{i_k\}. \tag{3.19}$$

We estimate the diagonal of Σ_Z by $\mathrm{diag}(\widehat{\Sigma}_Z) = \mathbf{1}$ and, following (A.21) and (A.22), we estimate the off-diagonal elements of Σ_Z by the corresponding entries of

$$\widehat{B}^+_{\widehat{I}\bullet}\left(\widehat{R}_{\widehat{I}\widehat{I}} - \widehat{\Gamma}_{\widehat{I}\widehat{I}}\right)\left[\widehat{B}^+_{\widehat{I}\bullet}\right]^\top \tag{3.20}$$

with $\widehat{B}^+_{\widehat{I}\bullet} = \left[\widehat{B}^\top_{\widehat{I}\bullet}\widehat{B}_{\widehat{I}\bullet}\right]^{-1}\widehat{B}^\top_{\widehat{I}\bullet}$ being the left inverse of $\widehat{B}_{\widehat{I}\bullet}$, and $\widehat{\Gamma}_{\widehat{I}\widehat{I}}$ estimated from (3.4).

We provide theoretical guarantees for the estimated loadings of the pure variable sub-matrix $A_{I\bullet}$, and for the estimates of the covariance matrix Σ_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 $|\bar{J}_1|\delta_n = o(1)$, with δ_n given in (2.15). With probability tending to one, as $n \to \infty$, we have $\widehat{I}_k = I_{\pi(k)}$ for $k \in [K]$ and for some permutation $\pi : [K] \to [K]$, and*

$$\min_{P\in\mathcal{P}_K}\max_{i\in I_k}\|\widehat{B}_{i\bullet} - PB_{i\bullet}\|_\infty \lesssim \delta_n,$$

$$\min_{P\in\mathcal{P}_K}\|\widehat{\Sigma}_Z - P^\top\Sigma_Z P\|_\infty \lesssim \delta_n,$$

$$\min_{P\in\mathcal{P}_K}\max_{i\in I_k}\frac{\|\widehat{A}_{i\bullet} - PA_{i\bullet}\|_\infty}{\sqrt{\Sigma_{ii}}} \lesssim \delta_n.$$

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 $A_{J\cdot}$.

Our proposed estimation of $A_{J\cdot}$ is identical to that in [8], 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

$$\begin{bmatrix} R_{II} & R_{IJ} \\ R_{JI} & R_{JJ} \end{bmatrix} = \begin{bmatrix} B_{I\bullet}\Sigma_Z B^\top_{I\bullet} & B_{I\bullet}\Sigma_Z B^\top_{J\bullet} \\ B_{J\bullet}\Sigma_Z B^\top_{I\bullet} & B_{J\bullet}\Sigma_Z B^\top_{J\bullet} \end{bmatrix} + \begin{bmatrix} \Gamma_{II} & \\ & \Gamma_{JJ} \end{bmatrix}.$$

In particular, from $R_{IJ} = B_{I\bullet}\Sigma_Z B^\top_{J\bullet}$, the submatrix $B_{J\bullet}$ is solved via

$$B^\top_{J\bullet} = \Sigma_Z^{-1}[B^\top_{I\bullet}B_{I\bullet}]^{-1}B^\top_{I\bullet}R_{IJ}.$$

Hence, after estimating Σ_Z and $B_{\widehat{I}\bullet}$, using the estimate $\widehat{R}$, we obtain a plug–in estimate for $B_{\widehat{J}\bullet}$ with $\widehat{J} = [p] \setminus \widehat{I}$. Alternatively, we can regress $\widehat{R}_{\widehat{I}\widehat{J}}$ on $\widehat{B}_{\widehat{I}\bullet}\widehat{\Sigma}_Z$, or $[\widehat{B}^\top_{\widehat{I}\bullet}\widehat{B}_{\widehat{I}\bullet}]^{-1}\widehat{B}^\top_{\widehat{I}\bullet}\widehat{R}_{\widehat{I}\widehat{J}}$ on $\widehat{\Sigma}_Z$. This approach has been adopted in [8] (and [9] 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 [8], for an overview of existing approaches. We also refer to Appendix C.1 in [8], 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′, when I and K are known.

As an illustration, suppose we take the approach in (3.13) – (3.14) of [8] to estimate each row $B_{j\bullet}$ of $B_{J\bullet}$, and estimate A via (3.17). More precisely, we estimate $B_{j\bullet}$ for each $j \in J$ by

$$\widehat{B}_{j\bullet} = \arg\min_{\beta} \|\beta\|_1 \qquad \text{subject to} \quad \|\widehat{\Sigma}_Z\beta - [\widehat{B}_{I\bullet}^\top \widehat{B}_{I\bullet}]^{-1}\widehat{B}_{I\bullet}^\top \widehat{R}_{Ij}\|_\infty \le C_0\delta_n \tag{3.21}$$

for some constant $C_0 > 0$. Here $\widehat{\Sigma}_Z$ is constructed via (3.20) above with $\operatorname{diag}(\widehat{\Sigma}_Z) = \mathbf{1}$. The following theorem provides the convergence rate of $\|\widehat{A} - AP\|_{\infty,q}$ for all $q \ge 1$, in the setting of bounded $\|\cdot\|_{\infty,1}$-norms of A and Σ_Z^{-1} as in [8]. Its proof can be found in Appendix A.2.6. Let $s = \max_i \|A_{i\bullet}\|_0$.

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

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

According to Theorem 6 of [8], the rate in Theorem 15 is minimax optimal, up to a multiplicative factor of $\sqrt{\log(p \vee n)}$. Note that $C'' = 1$ in [8] and we refer to Remark 4 of [8] for a detailed discussion on $\|\Sigma_Z^{-1}\|_{\infty,1}$.

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.

Funding

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: 10.3150/22-BEJ1502SUPP; .pdf). The supplementary document includes the proofs and simulation results.

References

[1] Agarwal, A., Negahban, S. and Wainwright, M.J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.* **40** 1171–1197. MR2985947 https://doi.org/10.1214/12-AOS1000

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

[3] Anandkumar, A., Ge, R., Hsu, D., Kakade, S.M. and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. MR3270750

[4] Anandkumar, A., Hsu, D. and Kakade, S.M. A method of moments for mixture models and hidden Markov models. In *Proceedings of Machine Learning Research. JMLR Workshop and Conference Proceedings* **23** 33.1–33.34. Edinburgh, Scotland.

[5] Anderson, T.W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. V* 111–150. Berkeley-Los Angeles, CA: Univ. California Press. MR0084943

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

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

[8] Bing, X., Bunea, F., Ning, Y. and Wegkamp, M. (2020). Adaptive estimation in structured factor models with applications to overlapping clustering. *Ann. Statist.* **48** 2055–2081. MR4134786 https://doi.org/10.1214/19-AOS1877

[9] Bing, X., Bunea, F. and Wegkamp, M. (2020). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli* **26** 1765–1796. MR4091091 https://doi.org/10.3150/19-BEJ1166

[10] Bing, X., Bunea, F. and Wegkamp, M. (2020). Optimal estimation of sparse topic models. *J. Mach. Learn. Res.* **21** Paper No. 177. MR4209463

[11] Bing, X., Bunea, F. and Wegkamp, M. (2023). Supplement to "Detecting approximate replicate components of a high-dimensional random vector with latent structure." https://doi.org/10.3150/22-BEJ1502SUPP

[12] Bing, X., Bunea, F. and Wegkamp, M. (2022). Inference in latent factor regression with clusterable features. *Bernoulli* **28** 997–1020. MR4388927 https://doi.org/10.3150/21-bej1374

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

[14] Bunea, F., Giraud, C., Luo, X., Royer, M. and Verzelen, N. (2020). Model assisted variable clustering: Minimax-optimal recovery and algorithms. *Ann. Statist.* **48** 111–137. MR4065155 https://doi.org/10.1214/18-AOS1794

[15] Bunea, F. and Xiao, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli* **21** 1200–1230. MR3338661 https://doi.org/10.3150/14-BEJ602

[16] Candès, E.J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11. MR2811000 https://doi.org/10.1145/1970392.1970395

[17] Chandrasekaran, V., Sanghavi, S., Parrilo, P.A. and Willsky, A.S. (2009). Sparse and low-rank matrix decompositions. *IFAC Proc. Vol.* **41** 493–1498.

[18] Donoho, D., Gavish, M. and Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.* **46** 1742–1778. MR3819116 https://doi.org/10.1214/17-AOS1601

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

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

[21] El Karoui, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Probab.* **19** 2362–2405. MR2588248 https://doi.org/10.1214/08-AAP548

[22] Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991 https://doi.org/10.1016/j.jeconom.2008.09.017

[23] Fan, J., Liao, Y. and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410 https://doi.org/10.1214/11-AOS944

[24] Fan, J., Liu, H. and Wang, W. (2018). Large covariance estimation through elliptical factor models. *Ann. Statist.* **46** 1383–1414. MR3819104 https://doi.org/10.1214/17-AOS1588
[25] Fan, J., Wang, W. and Zhong, Y. (2019). Robust covariance estimation for approximate factor models. *J. Econometrics* **208** 5–22. MR3906959 https://doi.org/10.1016/j.jeconom.2018.09.003
[26] Han, F. and Liu, H. (2017). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli* **23** 23–57. MR3556765 https://doi.org/10.3150/15-BEJ702
[27] Hsu, D., Kakade, S.M. and Zhang, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inf. Theory* **57** 7221–7234. MR2883652 https://doi.org/10.1109/TIT.2011.2158250
[28] Hyvärinen, A., Karhunen, J. and Oja, E. (2004). *Independent Component Analysis* **46**. New York: Wiley.
[29] Jaffe, A., Weiss, R., Nadler, B., Carmi, S. and Kluger, Y. (2018). Learning binary latent variable models: A tensor eigenpair approach. In *International Conference on Machine Learning* 2196–2205.
[30] Jin, J., Ke, Z.T. and Luo, S. (2017). Estimating network memberships by simplex vertex hunting. Preprint. Available at arXiv:1708.07852.
[31] Jöreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32** 443–482. MR0221659 https://doi.org/10.1007/BF02289658
[32] Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34** 183–202.
[33] Jöreskog, K.G. (1970). A general method for analysis of covariance structures. *Biometrika* **57** 239–251. MR0269024 https://doi.org/10.2307/2334833
[34] Joreskog, K.G. (1977). Factor analysis by least squares and maximum likelihood methods. In *Statistical Methods for Digital Computers III* (A.R.K. Enslein and H.S. Wilf, eds.) 125–153. Wiley.
[35] Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730
[36] Koopmans, T.C. and Reiersøl, O. (1950). The identification of structural characteristics. *Ann. Math. Stat.* **21** 165–181. MR0039967 https://doi.org/10.1214/aoms/1177729837
[37] Lawley, D.N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc. R. Soc. Edinb.* **60** 64–82. MR0002754
[38] Lawley, D.N. (1942). Further investigations in factor estimation. *Proc. Roy. Soc. Edinburgh Sect. A* **61** 176–185. MR0005579
[39] Lawley, D.N. (1943). The application of the maximum likelihood method to factor analysis. *British Journal of Psychology* **33** 172–175.
[40] Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058. MR3217437 https://doi.org/10.3150/12-BEJ487
[41] McDonald, R.P. (1999). *Test Theory: A Unified Treatment*. London: Taylor and Francis.
[42] Thurstone, L. (1931). Multiple factor analysis. *Psychological Review* **38** 406–427.
[43] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge: Cambridge Univ. Press. MR2963170
[44] Wegkamp, M. and Zhao, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* **22** 1184–1226. MR3449812 https://doi.org/10.3150/14-BEJ690
[45] Yalcin, I. and Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statist. Sci.* **16** 275–294. MR1874155 https://doi.org/10.1214/ss/1009213729

Received September 2021 and revised February 2022