Journal of Multivariate Analysis 199 (2024) 105244

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# Large factor model estimation by nuclear norm plus $\ell_1$ norm penalization

Matteo Farnè [*], Angela Montanari

*Dipartimento di Scienze Statistiche, Università di Bologna, Via delle Belle Arti 41, Bologna, Italy*

ARTICLE INFO

*Article history:*
Received 2 March 2023
Received in revised form 13 October 2023
Accepted 15 October 2023
Available online 19 October 2023

*AMS 2020 subject classifications:*
primary 62H25
secondary 65F55
15A18

*Keywords:*
Factor model
High dimension
Nuclear norm
Numerical instability
Sparsity

ABSTRACT

This paper provides a comprehensive estimation framework via nuclear norm plus $\ell_1$ norm penalization for high-dimensional approximate factor models with a sparse residual covariance. The underlying assumptions allow for non-pervasive latent eigenvalues and a prominent residual covariance pattern. In that context, existing approaches based on principal components may lead to misestimate the latent rank. On the contrary, the proposed optimization strategy recovers with high probability both the covariance matrix components and the latent rank and the residual sparsity pattern. Conditioning on the recovered low rank and sparse matrix varieties, we derive the finite sample covariance matrix estimators with the tightest error bound in minimax sense and we prove that the ensuing estimators of factor loadings and scores via Bartlett's and Thomson's methods have the same property. The asymptotic rates for those estimators of factor loadings and scores are also provided.

## 1. Introduction

The digital revolution has enormously enlarged the amount of available data for researchers and practitioners. Consequently, the need rises to develop methodologies able to summarize the content of high-dimensional datasets, in order to derive meaningful information from them.

The factor model is an effective tool to this end, as it detects the latent covariance structure behind a set of variables. We can define the factor model for any $p$-dimensional mean-centered random vector $\mathbf{x}$ as

$$\mathbf{x} = \mathbf{B}\mathbf{f} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{B}$ is a $p \times r$ matrix, $\mathbf{f}$ is a $r \times 1$ random vector with $\mathrm{E}[\mathbf{f}] = \mathbf{0}_r$ and $\mathrm{Var}[\mathbf{f}] = \mathbf{I}_r$, and $\boldsymbol{\epsilon}$ is a $p \times 1$ random vector with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}_p$ and $\mathrm{Var}[\boldsymbol{\epsilon}] = \mathbf{S}^*$, with $\mathbf{S}^*$ full rank $p \times p$ matrix.

Let us indicate by $\boldsymbol{\Sigma}^*$ the $p \times p$ covariance matrix of the random vector $\mathbf{x}$. Assuming that $\mathbf{f}$ and $\boldsymbol{\epsilon}$ are componentwise uncorrelated, the factor model (1) induces in $\boldsymbol{\Sigma}^*$ a low rank plus residual decomposition of the following type:

$$\boldsymbol{\Sigma}^* = \mathbf{L}^* + \mathbf{S}^* = \mathbf{B}\mathbf{B}^\top + \mathbf{S}^*, \tag{2}$$

where $\mathbf{L}^* = \mathbf{B}\mathbf{B}^\top = \mathbf{U}_L \boldsymbol{\Lambda}_L \mathbf{U}_L^\top$, with $\mathbf{U}_L$ $p \times r$ semi-orthogonal matrix and $\boldsymbol{\Lambda}_L$ $r \times r$ diagonal matrix. Representation (2) is invariant under orthogonal transforms, and it is therefore unidentifiable from the data without further constraints.

[*] Corresponding author.
*E-mail address:* matteo.farne@unibo.it (M. Farnè).

https://doi.org/10.1016/j.jmva.2023.105244

Suppose that we have a i.i.d. sample $\mathbf{x}_k$, $k \in \{1,\ldots,n\}$. The $p \times p$ sample covariance matrix is defined as $\boldsymbol{\Sigma}_n = n^{-1}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^\top$. Most of factor model estimation methods rely on $\boldsymbol{\Sigma}_n$ as an input, and make essentially use of two techniques: principal component analysis (see [28] for an overview) and maximum likelihood. As outlined in [7], however, a large dimension $p$ leads to some particular estimation problems for model (1), due to the limitations of $\boldsymbol{\Sigma}_n$ in high dimensions.

From a historical perspective, the classical inferential theory for factor models [2] prescribes that the dimension $p$ is fixed while the sample size $n$ tends to infinity. In particular, the strict condition $p < n$ is required to ensure consistency. As a consequence, the classical framework is clearly inappropriate if $p$ is large. When $p \geq n$, in fact, $\boldsymbol{\Sigma}_n$ becomes inconsistent and no longer distributed as non-singular Wishart.

At the same time, when the dimension $p$ and the sample size $n$ are finite, [3] shows that the use of the principal components of $\boldsymbol{\Sigma}_n$ to estimate $\mathbf{B}$ leads to factor loadings and scores estimates which are incoherent with model assumptions, because any estimate of $\mathbf{S}^*$ so derived will never be full rank. That is the reason why [17] proves that the principal components of $\boldsymbol{\Sigma}_n$ consistently identify $\mathbf{L}^*$ under model (1) as $p \to \infty$, provided that the $r$ eigenvalues of $\mathbf{L}^*$ diverge with $p$ and $\mathbf{S}^*$ is a non-diagonal matrix with vanishing eigenvalues as $p$ diverges.

Another relevant aspect concerns the ratio $p/n$. If $p/n \to 1^-$, where "$\to 1^-$" means "tends to 1 from the left", the bad conditioning properties of $\boldsymbol{\Sigma}_n$ inevitably affect the consistency of Principal Component Analysis (PCA) as a factor model estimation method. In fact, the sample eigenvalues follow the Marcenko–Pastur law [32], which crucially depends on the ratio $p/n$. In particular, if $p/n \to 1^-$, it is more likely to observe small sample eigenvalues, thus making $\boldsymbol{\Sigma}_n$ numerically unstable.

An overall inferential theory of PCA as a high-dimensional factor model estimation method has been developed in [4]. As also outlined in [4,17] shows that the pervasiveness of the eigenvalues of $\mathbf{L}^*$ as $p \to \infty$ is crucial for the recovery of the latent rank $r$, performed by the identification criteria of [6]. If that condition is violated, the latent rank $r$ may be underestimated by any PCA-based method, as one or more latent eigenvalues may be unrecovered, because the corresponding sample eigenvalues may not be large enough. In order to achieve consistency, PCA tolerates a non-diagonal residual covariance matrix $\mathbf{S}^*$ and residual heteroscedasticity, provided that $p$ and $n$ are both large and $\sqrt{n}/p$ tends to 0. On the contrary, if only $n$ is large, no non-diagonal residual covariance structure is admitted.

The authors in [22] propose to estimate the covariance matrix $\boldsymbol{\Sigma}^*$ in high dimensions under representation (2) by taking out the principal components of $\boldsymbol{\Sigma}_n$ and then thresholding their orthogonal complement, under the assumption that $\mathbf{S}^*$ has a bounded $\ell_1$ norm as $p$ diverges. The uniform parametric consistency of loadings, factor scores and commonalities obtained is established. That sparsity assumption on $\mathbf{S}^*$ also allows to make the estimation error of $\boldsymbol{\Sigma}_n$ vanish in relative terms as $p$ diverges.

The asymptotic distribution of factors and factor loadings estimated via PCA when both $p$ and $n$ are large is derived in [8]. A relevant merit of that paper is that factors and loadings are precisely identified without the need of any rotation. Under relatively weak factors in terms of explained variance proportion, [34] derives the (normal) asymptotic distribution of the Ordinary Least Squares (OLS) estimated coefficients in the regressions of the PC estimates of factors (loadings) on the true factors (true loadings). That distribution has good approximation properties even when both $p$ and $n$ are reasonably small.

Concerning maximum likelihood estimation, [2] shows that the exact maximum likelihood is consistent for loading estimation, even if it is still inconsistent as far as factor scores estimation is concerned. Nevertheless, factor scores can be consistently estimated by the conditional maximum likelihood, via Bartlett's [11] or Thomson's estimator [37].

The consistency of maximum likelihood (ML) to estimate a high-dimensional factor model has been studied in [5] (previous contributions on the topic also include [29,30]). Differently from the estimator of factor scores based on PCA, the one based on ML is consistent also for small $p$ and $n$, even if the estimator distribution is less complicated to derive when $p$ diverges. ML has a better asymptotic rate and is more efficient than PCA in the case of independent and heteroscedastic residuals. However, in presence of a non-diagonal residual covariance structure, the convergence rates and the optimality conditions of ML estimators become cumbersome. It is important to note that the relative magnitude of $p$ and $n$ is a crucial issue for both methods (ML and PCA) to provide consistent factor model estimates.

Given these premises, the interest arises to find an alternative estimation method to ML and PCA, as they both present some relevant drawbacks in high dimensions. First of all, the latent rank recovery fails if the latent eigenvalues are not spiked enough with respect to the dimension. Then, the sample covariance matrix is increasingly numerically unstable as the dimension increases, such that the need to regularize sample eigenvalues rises. In addition, a more effective sampling theory is needed with respect to the degree of spikiness of latent eigenvalues and the degree and pattern of residual sparsity. Ideally, latent rank recovery and numerical stability should be ensured for any finite values of $p$ and $n$.

In [9], it is proposed to use the nuclear norm heuristics in place of PCA. That work provides the asymptotic normality and parametric consistency of approximate factor model estimates as both $p$ and $n$ diverge. The proposed objective function is a least squares loss penalized by a nuclear norm plus $\ell_1$ norm heuristics, which is useful to detect covariance matrix decompositions of type (2) where $\mathbf{S}^*$ is element-wise sparse. In [23], the authors exploit the same heuristics to derive algebraically consistent covariance matrix estimates, that is, the latent rank and the residual sparsity pattern are recovered with high probability for finite values of $p$ and $n$. Such a feature is extremely important, as it allows to avoid the use of any identification criterion for the latent rank like the one described in [9].

Algebraic and parametric consistency are here defined analogously to [18].

**Definition 1.** A pair of symmetric matrices $(\mathbf{S}, \mathbf{L})$ with $\mathbf{S}, \mathbf{L} \in \mathbb{R}^{p\times p}$ is an algebraically consistent estimate of the low rank plus sparse decomposition (2) for the covariance matrix $\boldsymbol{\Sigma}^*$ if the following conditions hold: (a) the sign pattern of $\mathbf{S}$ is the same as that of $\mathbf{S}^*$: $\mathrm{sgn}(\mathbf{S}_{ij}) = \mathrm{sgn}(\mathbf{S}^*_{ij})$, for all $i, j \in \{1, \dots, p\}$ (we assume that $\mathrm{sgn}(0) = 0$); (b) the rank of $\mathbf{L}$ is the same as the rank of $\mathbf{L}^*$: $\mathrm{rk}(\mathbf{L}) = \mathrm{rk}(\mathbf{L}^*)$; (c) matrices $\mathbf{L}+\mathbf{S}$, $\mathbf{S}$, and $\mathbf{L}$ are such that $\mathbf{L}+\mathbf{S}$ and $\mathbf{S}$ are positive definite and $\mathbf{L}$ is positive semidefinite.

Parametric consistency holds if the pair of estimates $(\mathbf{S}, \mathbf{L})$ is close to the pair $(\mathbf{S}^*, \mathbf{L}^*)$ in some norm with probability approaching 1.

**Definition 2.** A pair of symmetric matrices $(\mathbf{S}, \mathbf{L})$ with $\mathbf{S}, \mathbf{L} \in \mathbb{R}^{p\times p}$ is a parametrically consistent estimate of the low rank plus sparse decomposition (2) for the covariance matrix $\boldsymbol{\Sigma}^*$ if the norm $g_\gamma = \max\left(\|\mathbf{S} - \mathbf{S}^*\|_\infty/\gamma, \|\mathbf{L} - \mathbf{L}^*\|_2/\|\mathbf{L}^*\|_2\right)$, where $\gamma \in \mathbb{R}^+$ and $\|.\|_\infty$ denotes the maximum norm, converges to 0 with probability approaching 1.

The results of [23] are obtained by allowing for intermediate degrees of spikiness for latent eigenvalues and intermediate degrees of sparsity for the residual component. In particular, their assumptions prescribe that the latent eigenvalues are spiked in the sense of Yu and Samworth ([22], p. 656), thus allowing for intermediately pervasive latent factors as $p$ diverges. What is more, the number of non-zeros in the residual component $\mathbf{S}^*$ is allowed to grow with $p$ (even if slower than the latent eigenvalues). The identifiability of the matrix components $\mathbf{L}^*$ and $\mathbf{S}^*$ is ensured by imposing that $\mathbf{L}^*$ and $\mathbf{S}^*$ are far enough from being sparse and low rank respectively.

The model setup of [23] broadens the one of [22], allowing for an intermediate degree of latent eigenvalues spikiness and residual sparsity. Unlike [9,22], the solution method in [23] recovers with high probability both the latent rank and the residual sparsity pattern, even when the latent eigenvalues diverge at a rate slower than $O(p)$ as $p \to \infty$, which may cause the rank selection criteria in [6] to fail (also see Lemma S.1 in the Supplement). As a consequence, establishing the optimality properties of the estimators of loadings and factor scores derived from the covariance matrix estimators of [23] is extremely important, as it allows to estimate high-dimensional factor models under a broader variety of data settings than existing competitors.

For these reasons, in this paper we study the estimators of loadings, factor scores and commonalities obtained under a model framework that builds over the one of [23]. The state of the art of high-dimensional factor model estimation is discussed in Section 2. In Section 3, the asymptotic consistency of factor model estimators based on least squares penalized by the nuclear norm plus $\ell_1$ norm heuristics is proved under appropriate regimes of intermediate latent eigenvalue spikiness and residual covariance sparsity. In Section 4, we present a finite sample version of the covariance matrix estimators obtained via that heuristics and we derive their optimal properties in terms of minimax error in spectral norm. In Section 5, we highlight that the ensuing Bartlett's and Thomson's estimators of factor scores provide the minimax error bound in Euclidean norm within the classes of algebraically consistent low rank and sparse component estimates given a finite sample. The conclusions follow in Section 6. A contains some essential technical details. A wide simulation study and a real data example proving the validity of our approach are provided in a Supplement, beyond some ancillary technical results.

## 2. Theoretical background

### *2.1. Notation*

Given a $p \times p$ symmetric positive semi-definite matrix $\mathbf{M}$, we denote by $\lambda_i(\mathbf{M})$, $i \in \{1, \dots, p\}$, the eigenvalues of $\mathbf{M}$ in decreasing order. To indicate that $\mathbf{M}$ is positive definite or semi-definite we use the notations: $\mathbf{M} \succ 0$ or $\mathbf{M} \succeq 0$, respectively. The expressions $\mathrm{diag}(\mathbf{M})$ and $\mathrm{off} - \mathrm{diag}(\mathbf{M})$ identify the diagonal elements and off-diagonal elements of $\mathbf{M}$, respectively. The minimum nonzero off-diagonal element of $\mathbf{M}$ in absolute value is denoted as $\|\mathbf{M}\|_{\min,\mathrm{off}} = \min_{\substack{1\le i,j\le p\\ i\ne j, \mathbf{M}_{ij}\ne 0}} |\mathbf{M}_{ij}|$. Then, we recall the following norm definitions: 1. element-wise: (a) $l_0$ norm: $\|\mathbf{M}\|_0 = \sum_{i=1}^p \sum_{j=1}^p \mathbb{1}(\mathbf{M}_{ij} \ne 0)$; (b) $l_1$ norm: $\|\mathbf{M}\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\mathbf{M}_{ij}|$; (c) Frobenius norm: $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^p \mathbf{M}_{ij}^2}$; (d) maximum norm: $\|\mathbf{M}\|_\infty = \max_{i\le p, j\le p} |\mathbf{M}_{ij}|$; 2. induced by vector: (a) $\|\mathbf{M}\|_{0,v} = \max_{i\le p} \sum_{j\le p} \mathbb{1}(\mathbf{M}_{ij} \ne 0)$, which is the maximum 'degree' of $\mathbf{M}$; (b) $\|\mathbf{M}\|_{1,v} = \max_{j\le p} \sum_{i\le p} |\mathbf{M}_{ij}|$; (c) spectral norm: $\|\mathbf{M}\|_2 = \lambda_1(\mathbf{M})$; 3. Schatten: (a) nuclear norm of $\mathbf{M}$, here defined as the sum of the eigenvalues of $\mathbf{M}$: $\|\mathbf{M}\|_* = \sum_{i=1}^p \lambda_i(\mathbf{M})$.

For any $t \ge 0$, we define: $\mathcal{T}_t^{(H)}$, the hard-thresholding operator with parameter $t$, such that the $p \times p$ matrix $\mathcal{T}_t^{(H)}(\mathbf{M})$ has $(i, j)$ element $\mathbf{M}_{ij}$ if $|\mathbf{M}_{ij}| \ge t$, 0 otherwise; $\mathcal{T}_t^{(S)}$, the soft-thresholding operator with parameter $t$, such that the $p \times p$ matrix $\mathcal{T}_t^{(S)}(\mathbf{M})$ has $(i, j)$ element $\mathrm{sgn}(\mathbf{M}_{ij}) \max(|\mathbf{M}_{ij}| - t, 0)$; $\mathcal{T}_t^{(SVT)}$, the singular value thresholding operator with parameter $t$, such that the $p \times p$ matrix $\mathcal{T}_t^{(SVT)}(\mathbf{M})$ is equal to $\mathbf{U}_M \mathcal{T}_t^{(S)}(\boldsymbol{\Lambda}_M)\mathbf{U}'_M$, where $\mathbf{U}_M \boldsymbol{\Lambda}_M \mathbf{U}'_M$ is the spectral decomposition of $\mathbf{M}$.

Given a $p$-dimensional vector $\mathbf{v}$, we denote by $\|\mathbf{v}\| = \sqrt{\sum_{i=1}^p \mathbf{v}_i^2}$ the Euclidean norm of $\mathbf{v}$, by $\|\mathbf{v}\|_\infty = \max_{i\in\{1,\dots,p\}} |\mathbf{v}_i|$ the maximum norm of $\mathbf{v}$, and by $v_{k,i}$ the $i$th component of the indexed vector $\mathbf{v}_k$.

Given two sequences $A_\nu$ and $B_\nu$, $\nu \to \infty$, we write $A_\nu = O(B_\nu)$ or $A_\nu \preceq B_\nu$, if there exists a positive real $C$ independent of $\nu$ such that $A_\nu/B_\nu \le C$, we write $B_\nu = O(A_\nu)$ or $A_\nu \succeq B_\nu$, if there exists a positive real $C$ independent of $\nu$ such that

$B_\nu/A_\nu \leq C$, and we write $A_\nu \simeq B_\nu$ if there exists a positive real $C$ independent of $\nu$ such that $A_\nu/B_\nu \leq C$ and $B_\nu/A_\nu \leq C$. Similarly, we write $A_\nu = o(B_\nu)$ or $A_\nu \prec B_\nu$, if there exists a positive real $C$ independent of $\nu$ such that $A_\nu/B_\nu < C$, and we write $B_\nu = o(A_\nu)$ or $A_\nu \succ B_\nu$, if there exists a positive real $C$ independent of $\nu$ such that $B_\nu/A_\nu < C$.

We denote by $O_p$ the big-O in probability.

### 2.2. State of the art

Imposing $\mathbf{S}^* = \mathbf{I}_p$, [4] shows that the loading matrix $\mathbf{B}$ and the factor scores $\mathbf{f}_k$, $k \in \{1, \dots, n\}$, are consistently recovered under model (1) as $p \to \infty$ by extracting the top $r$ eigenvectors of $\boldsymbol{\Sigma}_n$, provided that the $r$ eigenvalues of $\mathbf{L}^*$ are $O(p)$. The reason why this method is consistent as $p \to \infty$ can be understood by recalling [27]. In fact, the principal components of $\boldsymbol{\Sigma}_n$ are derived by solving the problem

$$\min_{\mathbf{L}, \mathrm{rk}(\mathbf{L}) \leq r} \frac{1}{p} \|\boldsymbol{\Sigma}_n - \mathbf{L}\|_F^2, \tag{3}$$

which is equivalent to the problem

$$\min_{\mathbf{b}_j, f_{k,j}} \frac{1}{np} \sum_{k=1}^{n} \|\mathbf{x}_k - \mathbf{z}_k\|_2^2, \tag{4}$$

where $\mathbf{z}_k = \sum_{j=1}^{r} \mathbf{b}_j f_{k,j}$, $f_{k,j}$ is the $j$th component of $\mathbf{f}_k$ and the $p \times 1$ column vectors $\mathbf{b}_j$, $j \in \{1, \dots, r\}$, are orthogonal. Intuitively, the solutions to problem (4) are consistent under model (1) if and only if the eigenvalues of $\mathbf{L}^*$ are $O(p)$ as $p \to \infty$, because otherwise the signal $\mathbf{z}_k$ would not be strong enough to be detected.

Full solution vectors $\mathbf{b}_j$, $j \in \{1, \dots, r\}$, can be difficult to interpret in high dimensions. For this reason, [38] introduces Sparse Principal Component Analysis (SPCA), a method based on a version of problem (4) where each $\mathbf{b}_j$ is penalized by a ridge plus lasso penalty. The resulting sparse principal components are no longer orthogonal and represent approximate solutions, which reduce effectively the complexity of estimated components when $p$ is large.

At the same time, as $p$ diverges, the assumption $\mathbf{S}^* = \mathbf{I}_p$ is definitely too strong, as it is unlikely that the latent structure is able to entirely catch the covariance for all pairs of variables. In order to relax that assumption, [15] proposes Principal Component Pursuit (PCP), that is based on the solution of the following problem:

$$\min_{\mathbf{L}+\mathbf{S}=\boldsymbol{\Sigma}_n} \|\mathbf{L}\|_* + \|\mathbf{S}\|_1, \tag{5}$$

where $\|\mathbf{L}\|_*$ is the nuclear norm of $\mathbf{L}$ and $\|\mathbf{S}\|_1$ is the $\ell_1$ norm of $\mathbf{S}$. Problem (5) can be thought of as a robust PCA problem in presence of missing or grossly corrupted data. It is solved by exploiting the singular value thresholding algorithm of [14].

The use of the nuclear norm for rank minimization as an alternative to PCA was first proposed in [25]. The nuclear norm was then successfully applied to matrix completion problems, among which the Netflix problem is the most celebrated one. Within this research strand, we mention [16,26,33,35], which all describe and solve approximate robust PCA problems.

Even if problem (5) is able to bypass the assumption $\mathbf{S}^* = \mathbf{I}_p$, the number of parameters to be recovered may be remarkably high without any further assumption on $\mathbf{S}^*$, particularly if $p$ is large. In order to reduce the parameter space dimensionality, a rough alternative is to impose sparsity on $\boldsymbol{\Sigma}^*$. In the covariance matrix context, for instance, [12] assumes that $\boldsymbol{\Sigma}^*$ is sparse and recover it by solving the problem $\min_{\boldsymbol{\Sigma}} \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}\|_1$. This problem is solved by applying to $\boldsymbol{\Sigma}_n$ the soft-thresholding algorithm of [20], which is consistent for $\boldsymbol{\Sigma}^*$ but does not provide any dimension reduction.

In this paper, we merge dimension reduction and sparsity in a single problem with the aim to explore the performance of the ensuing estimators of factor scores and loadings. First, we recover the two components $\mathbf{L}^*$ and $\mathbf{S}^*$ of $\boldsymbol{\Sigma}^*$ from $\boldsymbol{\Sigma}_n$. This step is performed by solving the following problem [23]:

$$\min_{\mathbf{L},\mathbf{S}} \frac{1}{2} \|\boldsymbol{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_F^2 + \psi \|\mathbf{L}\|_* + \rho \|\mathbf{S}\|_1, \tag{6}$$

where $\|\mathbf{L}\|_*$ is the nuclear norm of $\mathbf{L}$ and $\|\mathbf{S}\|_1$ is the $\ell_1$ norm of $\mathbf{S}$, and $\psi$ and $\rho$ are positive threshold parameters. The feasible set of (6) is the set of all $p \times p$ symmetric positive definite matrices $\mathbf{S}$ and all $p \times p$ symmetric positive semi-definite matrices $\mathbf{L}$. Problem (6) is a least squares one, penalized by a nuclear norm plus $\ell_1$ norm heuristics, which has been proved in [24] to be the tightest convex relaxation of the original NP-hard problem involving $\mathrm{rk}(\mathbf{L})$ and $\|\mathbf{S}\|_0$. Some variants of (6) have been used to estimate under the low rank plus sparse assumption in high dimensions the covariance matrix and its inverse, as well as the spectral density matrix, in [1,10,18], respectively.

Problem (6) can be thought of as an approximate robust PCA problem. In [23], a refined estimation theory for the estimators of $\mathbf{L}^*, \mathbf{S}^*$ and $\boldsymbol{\Sigma}^*$ obtained via (6) is provided assuming the generalized spikiness of the eigenvalues of $\mathbf{L}^*$ and the generalized element-wise sparsity of $\mathbf{S}^*$. The solutions to problem (6) in [23] are called $\widehat{\mathbf{L}}_{\mathrm{ALCE}}$ and $\widehat{\mathbf{S}}_{\mathrm{ALCE}}$, where ALCE stands for ALgebraic Covariance Estimator. ALCE estimates are computed via an alternate thresholding algorithm, composed of a singular value thresholding [14] and a soft-thresholding step [20]. In the Supplement, we report the pseudo-code of the solution algorithm (Section 1) and a criterion to select the optimal threshold pair $(\psi, \rho)$ (Section 3).

ALCE estimates are then re-optimized in order to minimize the spectral loss from the target $\boldsymbol{\Sigma}^*$ given a finite sample. Such task is performed by applying an additional least squares step, which leads to the final covariance matrix estimates $\widehat{\mathbf{L}}_{\mathrm{UNALCE}}$ and $\widehat{\mathbf{S}}_{\mathrm{UNALCE}}$ (where UNALCE stands for UNshrunk ALCE). Under the assumptions of [23], UNALCE estimates converge to ALCE ones as $p$ and $n$ diverge. A characterizing feature of UNALCE and ALCE estimates is that they are both parametrically and algebraically consistent, i.e. covariance matrix estimates are consistent in spectral norm and the latent rank and the residual sparsity pattern are recovered with high probability.

The effectiveness of problem (6) as a factor model estimation method has been recently studied in [9] as far as parametric consistency is concerned, but no algebraic consistency theory is provided therein. Moreover, the latent eigenvalues in [9] must diverge with $p$ in order to ensure parametric consistency, while [23] allows for intermediate degrees of latent eigenvalues spikiness and residual sparsity. An alternative approach is POET [22]. POET covariance matrix estimator is the result of a two-step procedure where $\mathbf{L}^*$ is estimated as the covariance matrix of the top $r$ principal components, and $\mathbf{S}^*$ is estimated by soft-thresholding their orthogonal complement. In [22], the strict pervasiveness of latent factors is assumed, i.e. latent eigenvalues are assumed to be $O(p)$, while residual covariance sparsity is imposed by bounding the $\ell_1$ norm. Under those assumptions, the consistency of factor loadings and scores obtained by OLS is provided.

### 2.3. Paper contributions

In comparison to [9,22], the ALCE estimation framework [23] gives several advantages, e.g., there is no need to use any additional criterion to recover the latent rank, intermediately spiked latent eigenvalues are recovered, a residual sparsity pattern can be recovered, and the sampling theory is relaxed according to the pervasiveness degree of latent factors and the sparsity degree of the residual component. For this reason, in this paper we formally establish the consistency of approximate factor model estimators obtained by ALCE covariance matrix estimates. Consistency is established in a double asymptotic framework where both $p$ and $n$ diverge to infinity, by deriving the conditions ensuring the asymptotic equivalence between the nuclear norm plus $\ell_1$ norm penalized problem (6) and the OLS problem (4). To this end, we extend the theoretical framework of [23] by means of specific assumptions ensuring factor model recovery by exactly characterizing the allowed relative size of $p$ and $n$, the tolerated strength of latent factors and degree of residual sparsity, and the maximum absolute loading and residual covariance magnitude allowed as $p, n \to \infty$. We also include differentiated speeds of divergence for latent eigenvalues as $p \to \infty$. Furthermore, we prove that factor loadings as well as Bartlett's and Thomson's factor scores estimators obtained by UNALCE covariance matrix estimates possess strong optimality properties in finite samples, and we derive the conditions under which they converge to their OLS counterparts. This fully characterizes factor model estimators obtained by solving problem (6), both asymptotically and in finite samples, while retaining the exact recovery of latent rank and residual sparsity pattern.

## 3. Generalized factor model estimation

### 3.1. Derivation of loadings and scores estimates

#### 3.1.1. OLS estimation

Let us define the $n \times r$ matrix $\mathbf{F}$ of factor scores as $\mathbf{F}^\top = [\mathbf{f}_1 \ \dots \ \mathbf{f}_n]$, the $p \times r$ matrix $\mathbf{B}$ of factor loadings as $\mathbf{B}^\top = [\mathbf{b}_1 \ \dots \ \mathbf{b}_p]$, and the $n \times p$ data matrix $\mathbf{X}$ as $\mathbf{X}^\top = [\mathbf{x}_1 \ \dots \ \mathbf{x}_n]$. The estimates of factor loadings and scores based on the OLS are derived as follows:

$$\min_{\mathbf{B},\mathbf{F}} \frac{1}{pn} \sum_{j=1}^{p} \sum_{k=1}^{n} (\mathbf{x}_{kj} - \mathbf{b}_j^\top \mathbf{f}_k)^2. \tag{7}$$

According to [4], minimizing (7) amounts to maximizing $\mathrm{tr}(\mathbf{F}^\top(\mathbf{X}\mathbf{X}^\top)\mathbf{F})$. Under the constraints that $\frac{1}{n}\sum_{k=1}^n \widehat{\mathbf{f}}_k \widehat{\mathbf{f}}_k^\top = \mathbf{I}_r$ and $\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}}$ is diagonal, (7) is solved by $\widehat{\mathbf{F}}_{\mathrm{OLS1}} = \sqrt{n}\mathbf{U}_n$, where $\mathbf{U}_n$ is the $n \times r$ matrix of the top $r$ eigenvectors of the $n \times n$ matrix $\mathbf{X}\mathbf{X}^\top$, and $\widehat{\mathbf{B}}_{\mathrm{OLS1}}^\top = n^{-1}\widehat{\mathbf{F}}_{\mathrm{OLS1}}^\top \mathbf{x}$. It is easy to verify that $n^{-1}\widehat{\mathbf{F}}_{\mathrm{OLS1}}^\top \widehat{\mathbf{F}}_{\mathrm{OLS1}} = n^{-1}\sqrt{n}\mathbf{U}_n^\top \sqrt{n}\mathbf{U}_n = \mathbf{I}_r$, and that $\widehat{\mathbf{B}}_{\mathrm{OLS1}}^\top \widehat{\mathbf{B}}_{\mathrm{OLS1}} = \boldsymbol{\Lambda}_r$, where $\boldsymbol{\Lambda}_r$ is the $r \times r$ diagonal matrix containing the largest $r$ eigenvalues of $\boldsymbol{\Sigma}_n$ in decreasing order.

Let us explore alternative estimates of loadings and factor scores. We denote the spectral decomposition of $\boldsymbol{\Sigma}_n$ as $\mathbf{U}_p \boldsymbol{\Lambda}_p \mathbf{U}_p^\top$, where $\mathbf{U}_p$ contains as columns the $p$ eigenvectors of $\boldsymbol{\Sigma}_n$, and $\boldsymbol{\Lambda}_p$ is a diagonal $p \times p$ matrix containing the associated eigenvalues in decreasing order. We reconsider the principal component problem (3), whose solution is $\mathbf{L}_r = \mathbf{U}_r \boldsymbol{\Lambda}_r \mathbf{U}_r^\top$, where $\mathbf{U}_r$ is the $p \times r$ matrix whose columns are the eigenvectors corresponding to the largest $r$ eigenvalues of $\boldsymbol{\Sigma}_n$, contained in $\boldsymbol{\Lambda}_r$. Then, we choose $\widehat{\mathbf{B}}_{\mathrm{OLS2}} = \mathbf{U}_r \boldsymbol{\Lambda}_r{}^{1/2}$ as the loading matrix estimator, and, conditionally on $\widehat{\mathbf{B}}_{\mathrm{OLS2}}$, we estimate the factor scores via OLS as $\widehat{\mathbf{f}}_{\mathrm{OLS2},k} = (\widehat{\mathbf{B}}_{\mathrm{OLS2}}^\top \widehat{\mathbf{B}}_{\mathrm{OLS2}})^{-1} \widehat{\mathbf{B}}_{\mathrm{OLS2}}^\top \mathbf{x}_k = \boldsymbol{\Lambda}_r^{-1} \widehat{\mathbf{B}}_{\mathrm{OLS2}}^\top \mathbf{x}_k$, for $k \in \{1, \dots, n\}$.

It is worth exploring the relationship between $\widehat{\mathbf{F}}_{\mathrm{OLS1}} = \sqrt{n}\mathbf{U}_n$ and $\widehat{\mathbf{F}}_{\mathrm{OLS2}}$, defined as $\widehat{\mathbf{F}}_{\mathrm{OLS2}}^\top = [\widehat{\mathbf{f}}_{\mathrm{OLS2},1} \ \dots \ \widehat{\mathbf{f}}_{\mathrm{OLS2},n}]$. Denoting the eigenvalues and the eigenvectors of $\boldsymbol{\Sigma}_n$ by $\lambda_i(\boldsymbol{\Sigma}_n)$ and $\mathbf{u}_i$, $i \in \{1, \dots, p\}$, we know (see [13], formula (12).30) that the corresponding eigenvalues and eigenvectors of $n^{-1}\mathbf{X}\mathbf{X}^\top$ are $\lambda_i(\boldsymbol{\Sigma}_n)$ and $[\sqrt{n\lambda_i(\boldsymbol{\Sigma}_n)}]^{-1}\mathbf{X}\mathbf{u}_i$, respectively, with $i \in \{1, \dots, p\}$. Therefore, we can recognize that $\mathbf{U}_n^\top = n^{-1/2}\boldsymbol{\Lambda}_r^{-1/2}\mathbf{U}_r^\top \mathbf{X}^\top$. It follows that $\widehat{\mathbf{F}}_{\mathrm{OLS2}} = \mathbf{X}\widehat{\mathbf{B}}_{\mathrm{OLS2}}\boldsymbol{\Lambda}_r^{-1} = \mathbf{X}\mathbf{U}_r\boldsymbol{\Lambda}_r^{-1/2} = \sqrt{n}\mathbf{U}_n = \widehat{\mathbf{F}}_{\mathrm{OLS1}}$, and, straightforwardly, that $\widehat{\mathbf{B}}_{\mathrm{OLS2}} = \widehat{\mathbf{B}}_{\mathrm{OLS1}}$.

#### 3.1.2. ALCE estimation

We assume that the matrix components $\mathbf{L}^*$ and $\mathbf{S}^*$ come from the following sets of matrices:

$$\mathcal{L}(r) = \{\mathbf{L} \mid \mathbf{L} \succeq 0, \mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{U}^\top, \mathbf{U} \in \mathbb{R}^{p\times r}, \mathbf{U}^\top\mathbf{U} = \mathbf{I}_r, \mathbf{D} \in \mathbb{R}^{r\times r}\text{diagonal}\}, \tag{8}$$

$$\mathcal{S}(s) = \{\mathbf{S} \in \mathbb{R}^{p\times p} \mid \mathbf{S} \succ 0, |\text{supp}(\mathbf{S})| \leq s\}, \tag{9}$$

where $\mathcal{L}(r)$ is the variety of matrices with at most rank $r$, and $\mathcal{S}(s)$ is the variety of (element-wise) sparse matrices with at most $s$ non-zero elements ($\text{supp}(\mathbf{S})$ is the orthogonal complement of $ker(\mathbf{S})$ and $|\text{supp}(\mathbf{S})|$ denotes its dimension). Provided that $\mathbf{L}^* \in \mathcal{L}(r)$ and $\mathbf{S}^* \in \mathcal{S}(s)$, [19] shows that an uncertainty principle holds: if $\mathbf{L}^*$ is nearly sparse, $\mathbf{S}^*$ cannot be recovered; if $\mathbf{S}^*$ is nearly low rank, $\mathbf{L}^*$ cannot be recovered.

Let us denote by $\mathcal{T}(\mathbf{L}^*)$ and $\Omega(\mathbf{S}^*)$ the tangent spaces to $\mathcal{L}(r)$ and $\mathcal{S}(s)$ respectively, and recall the following rank-sparsity measures introduced in [18]:

$$\xi(\mathcal{T}(\mathbf{L}^*)) = \max_{\mathbf{L}\in\mathcal{T}(\mathbf{L}^*), \|\mathbf{L}\|_2\leq 1} \|\mathbf{L}\|_\infty, \tag{10}$$

$$\mu(\Omega(\mathbf{S}^*)) = \max_{\mathbf{S}\in\Omega(\mathbf{S}^*), \|\mathbf{S}\|_\infty\leq 1} \|\mathbf{S}\|_2. \tag{11}$$

Bounding the product between (10) and (11) ensures that the tangent spaces $\mathcal{T}(\mathbf{L}^*)$ and $\Omega(\mathbf{S}^*)$ are close to orthogonality. This guarantees algebraic consistency, i.e. the recovery of latent rank and residual sparsity pattern with high probability via heuristics (6). In particular, according to [18], the identifiability condition to be satisfied is $\xi(\mathcal{T}(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*)) \leq 1/54$.

The pair of ALCE covariance matrix estimates $(\widehat{\mathbf{L}}_{\text{ALCE}}, \widehat{\mathbf{S}}_{\text{ALCE}})$ is derived as

$$(\widehat{\mathbf{L}}_{\text{ALCE}}, \widehat{\mathbf{S}}_{\text{ALCE}}) = \arg\min_{\mathbf{L},\mathbf{S}} \frac{1}{2}\|\boldsymbol{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_F^2 + \psi\|\mathbf{L}\|_* + \rho\|\mathbf{S}\|_1. \tag{12}$$

We define $\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}} = \widehat{\mathbf{L}}_{\text{ALCE}} + \widehat{\mathbf{S}}_{\text{ALCE}}$, $\widehat{r}_A = \text{rk}(\widehat{\mathbf{L}}_{\text{ALCE}})$, and $\widehat{s}_A = |\text{supp}(\widehat{\mathbf{S}}_{\text{ALCE}})|$. The recovered low rank and sparse matrix varieties $\widehat{\mathcal{L}}(\widehat{r}_A)$ and $\widehat{\mathcal{S}}(\widehat{s}_A)$ are then defined as

$$\widehat{\mathcal{L}}(\widehat{r}_A) = \{\mathbf{L} \in \mathbb{R}^{p\times p} \mid \mathbf{L} = \widehat{\mathbf{U}}_{\text{ALCE}}\mathbf{D}\widehat{\mathbf{U}}_{\text{ALCE}}^\top,\ \mathbf{D} \in \mathbb{R}^{\widehat{r}_A\times\widehat{r}_A}\text{diagonal}\}, \tag{13}$$

$$\widehat{\mathcal{S}}(\widehat{s}_A) = \{\mathbf{S} \in \mathbb{R}^{p\times p} \mid |\text{supp}(\mathbf{S})| \leq \widehat{s}_A\}. \tag{14}$$

We denote the spectral decomposition of $\widehat{\mathbf{L}}_{\text{ALCE}}$ by $\widehat{\mathbf{U}}_{\text{ALCE}}\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}\widehat{\mathbf{U}}_{\text{ALCE}}^\top$, where $\widehat{\mathbf{U}}_{\text{ALCE}}$ is the $p \times \widehat{r}_A$ matrix containing as columns the eigenvectors of $\widehat{\mathbf{L}}_{\text{ALCE}}$, and $\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}$ is the $\widehat{r}_A \times \widehat{r}_A$ diagonal matrix containing the eigenvalues of $\widehat{\mathbf{L}}_{\text{ALCE}}$ in decreasing order. Analogously to previously defined OLS factor model estimators, we define the ALCE loading matrix estimator $\widehat{\mathbf{B}}_{\text{ALCE2}}$ as $\widehat{\mathbf{B}}_{\text{ALCE2}} = \widehat{\mathbf{B}}_{\text{ALCE}}$, with $\widehat{\mathbf{B}}_{\text{ALCE}} = \widehat{\mathbf{U}}_{\text{ALCE}}\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}^{1/2}$. Conditionally on $\widehat{\mathbf{B}}_{\text{ALCE2}}$, we then define the corresponding estimator of factor scores for $k \in \{1,\ldots,n\}$ via OLS as $\widehat{\mathbf{f}}_{\text{ALCE2},k} = (\widehat{\mathbf{B}}_{\text{ALCE2}}^\top\widehat{\mathbf{B}}_{\text{ALCE2}})^{-1}\widehat{\mathbf{B}}_{\text{ALCE2}}^\top\mathbf{x}_k = \widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}^{-1}\widehat{\mathbf{B}}_{\text{ALCE2}}^\top\mathbf{x}_k$. Finally, to complete the analogy, we set $\widehat{\mathbf{F}}_{\text{ALCE1}} = \widehat{\mathbf{F}}_{\text{ALCE2}}$ and $\widehat{\mathbf{B}}_{\text{ALCE1}} = \widehat{\mathbf{B}}_{\text{ALCE2}}$.

### 3.2. Consistency of loadings and scores estimates

#### 3.2.1. OLS estimation

Suppose that model (1) holds. We state the sufficient conditions to ensure the parametric consistency of the OLS estimators of factor loadings and scores under intermediate spikiness regimes for the latent eigenvalues and intermediate sparsity regimes for the residual component.

**Assumption 1.** (a) The eigenvalues of the $r \times r$ matrix $\mathbf{B}^\top\mathbf{B}$ are such that $\lambda_i(\mathbf{B}^\top\mathbf{B}) \simeq p^{\alpha_i}$, $i \in \{1,\ldots,r\}$, for some $1/2 < \alpha_r \leq \cdots \leq \alpha_1 \leq 1$; (b) $\|\mathbf{b}_j\|_\infty = O(1)$, $j \in \{1,\ldots,p\}$, and $r$ is finite for all $p \in \mathbb{N}$.

Assumption 1(a) prescribes that the latent eigenvalues are intermediately spiked with respect to (hereafter, *wrt*) the dimension $p$ in the sense of Yu and Samworth ([22], p. 656), thus allowing for intermediately pervasive latent factors as $p$ diverges. This is a characterizing feature of our approach compared to existing factor model estimators in high dimensions, like those in [9,22]. We also allow for different speeds of divergence for latent eigenvalues as $p$ increases. Assumption 1(b) prescribes that the factor loading vectors have bounded maximum norm, and that the latent rank $r$ is finite and independent of $p$.

**Assumption 2.** For all $p \in \mathbb{N}$, there exist $\delta_1 \in [0, 1/2]$ and $\delta_2 > 0$, such that: (a) $\|\mathbf{S}^*\|_{0,v} = \max_{1\leq i\leq p}\sum_{j=1}^p \mathbb{1}(\mathbf{S}_{ij}^* \neq 0) \leq \delta_2 p^{\delta_1}$; (b) $\|\mathbf{S}^*\|_\infty = O(1)$; (c) $p^{1-\delta_1}\|\mathbf{S}^*\|_{\text{min,off}} = o(1)$; (d) $\sum_{j=1}^p \mathbf{S}_{jj}^* = o(p^{\alpha_1})$.

Assumption 2 prescribes that: for all $p \in \mathbb{N}$, the maximum number of non-zeros per row, i.e. the maximum degree of $\mathbf{S}^*$, is bounded; the maximum entry of $\mathbf{S}^*$ in absolute value is $O(1)$; the minimum absolute nonzero element of $\mathbf{S}^*$ is $o(1/p^{1-\delta_1})$; the sum of residual variances is $o(p^\alpha)$. This assumption allows to apply the probabilistic framework of [12] to the residual covariance matrix component, and to make negligible its impact on the overall covariance matrix rate as $n \to \infty$. The resulting sparsity framework explicitly controls the number of non-zeros in the residual component and their position, and imposes a gap between the magnitude of the smallest latent eigenvalue and the largest residual eigenvalue, because $\|\mathbf{S}^*\|_2 \le \|\mathbf{S}^*\|_{1,v} \le \|\mathbf{S}^*\|_{0,v}\|\mathbf{S}^*\|_\infty \le \delta_2 p^{\delta_1}$, and $\delta_1 \le 1/2 < \alpha_r$ by Assumption 1(a). Part (d) actually equals to impose $\mathrm{tr}(\mathbf{S}^*) = o(\mathrm{tr}(\mathbf{L}^*))$, which is a reasonable condition because it is equivalent to prescribe for model (1) that $\mathrm{tr}(\mathbf{L}^*)/\mathrm{tr}(\boldsymbol{\Sigma}^*) \to 1$ as $p \to \infty$, analogously to [19].

**Assumption 3.** In model (1), $\mathrm{E}[\mathbf{f}] = \mathbf{0}_r$, $\mathrm{Var}[\mathbf{f}] = \mathbf{I}_r$, $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}_p$, $\mathrm{Var}[\boldsymbol{\epsilon}] = \mathbf{S}^*$, $\lambda_p(\mathbf{S}^*) > 0$, $\mathrm{E}[\mathbf{f}\boldsymbol{\epsilon}^\top] = \mathbf{0}_{r\times p}$, and there exist $b_1, b_2, c_1, c_2 > 0$ such that, for any $l > 0$, $k \in \{1, \dots, n\}$, $i \in \{1, \dots, r\}$, $j \in \{1, \dots, p\}$:

$$\Pr(|f_{k,i}| > l) \le \exp\{-(l/b_1)^{c_1}\}, \qquad \Pr(|\epsilon_{k,j}| > l) \le \exp\{-(l/b_2)^{c_2}\}.$$

Assumption 3 defines model (1) as an approximate factor model in the sense of [17], and imposes sub-exponential tails to true factors and residuals, as in [22]. This allows to apply large deviation theory to factors, residuals and their interactions. Moreover, it implies that all moments of $f_{k,i}$ and $\epsilon_{k,j}$ exist for all $k \in \{1, \dots, n\}$, $i \in \{1, \dots, r\}$, $j \in \{1, \dots, p\}$.

**Assumption 4.** There exists $M > 0$ such that, for all $j \in \{1, \dots, p\}$ and $k \in \{1, \dots, n\}$: (a) $\mathrm{E}[p^{-1/2}(\boldsymbol{\epsilon}_k^\top\boldsymbol{\epsilon}_k - \mathrm{E}[\boldsymbol{\epsilon}_k^\top\boldsymbol{\epsilon}_k])^4] < M$ and (b) $\mathrm{E}[\|p^{-1/2}\sum_{j=1}^p \mathbf{b}_j\epsilon_{k,j}\|^4] < M$.

Assumption 4 is necessary to estimate loadings and factor scores. It explicitly controls cross-sectional dependence, analogously to [4,22].

We can now focus on factor model estimators based on OLS. We follow the inferential framework of [4], exactly as [22] does. We define the projection matrix onto the orthogonally rotated true factor space as $\mathbf{H}_{\mathrm{OLS1}} = n^{-1}(\boldsymbol{\Lambda}_r)^{-1}\widehat{\mathbf{F}}_{\mathrm{OLS1}}^\top\mathbf{F}\mathbf{B}^\top\mathbf{B}$. Then, the following theorem holds.

**Theorem 1.** *Suppose that Assumptions 1–4 hold. Assume that $\|\mathbf{S}^*\|_1/p \le \delta_2'$ for some $\delta_2' > 0$, $\alpha_1 - \alpha_r \le \delta_1$, and $\max\{p^{2-2\alpha_r}, \ln(p)\}/n = o(1)$. Then, as $p, n \to \infty$, for the estimators $\widehat{\mathbf{B}}_{\mathrm{OLS1}}^\top = [\widehat{\mathbf{b}}_{\mathrm{OLS1},1} \dots \widehat{\mathbf{b}}_{\mathrm{OLS1},p}]$ and $\widehat{\mathbf{F}}_{\mathrm{OLS1}}^\top = [\widehat{\mathbf{f}}_{\mathrm{OLS1},1} \dots \widehat{\mathbf{f}}_{\mathrm{OLS1},n}]$ minimizing (7) with $r$ known it holds:*

(i) $\max_{j\le p} \|\widehat{\mathbf{b}}_{\mathrm{OLS1},j} - \mathbf{H}_{\mathrm{OLS1}}\mathbf{b}_j\| = O_p\left(\sqrt{\frac{\ln(p)}{n}}\right)$;

(ii) $\max_{k\le n} \|\widehat{\mathbf{f}}_{\mathrm{OLS1},k} - \mathbf{H}_{\mathrm{OLS1}}\mathbf{f}_k\| = O_p\left(\frac{n^{1/4}p^{1-\alpha_r}}{p^{1/2}}\right)$;

(iii) $\max_{j\le p,k\le n} \|\widehat{\mathbf{b}}_{\mathrm{OLS1},j}^\top\widehat{\mathbf{f}}_{\mathrm{OLS1},k} - \mathbf{b}_j^\top\mathbf{f}_k\| = O_p\left(\frac{n^{1/4}p^{1-\alpha_r}}{p^{1/2}} + \ln(n)^{\frac{1}{c_2}}\sqrt{\frac{\ln(p)}{n}}\right)$.

Theorem 1, proved in A, provides the uniform error rates in Euclidean norm for loading vectors, factor scores and commonalities estimated by OLS under generalized latent eigenvalues spikiness and residual sparsity, provided that the latent rank $r$ is known. This condition is necessary because, as reported by Yu and Samworth in the discussion of [22], the latent rank may be underestimated by the information criteria of [6] when $\alpha_r < 1$, since in that case $\lim_{p,n\to\infty}\Pr\{IC(r') < IC(r)\} > 0$, $r' < r$ (we refer to Lemma S.1 in the Supplement for more details).

The condition $\max\{p^{2-2\alpha_r}, \ln(p)\}/n = o(1)$ is sufficient to ensure that the estimation errors vanish as $p, n \to \infty$. The condition $\alpha_1 - \alpha_r \le \delta_1$ rises to ensure the annihilation of the residual component contribution to the estimation errors as $p, n \to \infty$. The condition $\|\mathbf{S}^*\|_1/p \le \delta_2'$ is imposed to control the cumulation of residual covariances. By assuming $\alpha_1 = \alpha_r = 1$, we note that we reobtain the rates and the conditions of Theorem 4 in [22] (also see Theorem S.1 in the Supplement for more details), and that the condition $p^{2-2\alpha_r}/n = o(1)$ ceases to be binding as $n \to \infty$, which means that Theorem 1 holds for all $p \in \mathbb{N}$ as $n \to \infty$ and $\ln(p)/n = o(1)$.

**Remark 1.** Part (ii) of Theorem 1 shows that, for the estimation error of factor scores to disappear, it is required that $p^{2-2\alpha_r}/n = o(1)$ and $n^{1/4}p^{1-\alpha_r}/p^{1/2} = o(1)$ hold simultaneously. This implies that it must hold $\underline{\delta}_{n,1}p^{2-2\alpha_r} < n < \overline{\delta}_{n,1}p^{4\alpha_r-2}$ for some $\underline{\delta}_{n,1}, \overline{\delta}_{n,1} > 0$ with $\underline{\delta}_{n,1} < \overline{\delta}_{n,1}$. The inequality $2 - 2\alpha_r \le 4\alpha_r - 2$ thus leads to the condition $\alpha_r \ge 2/3$ for this to hold. If $\alpha_r = 2/3$, it must be $n = O(p^{2/3})$. If $\alpha_r = 3/4$, it must hold $\sqrt{p} = o(n)$ and $n = o(p)$. If $\alpha_r = 1$, it must hold $\ln(p) = o(n)$ and $n = o(p^2)$, consistently with [22].

### *3.2.2. ALCE estimation*

The consistency of ALCE factor model estimators requires to bound the quantities (10) and (11). In order to clarify why, we define: the incoherence of $\mathbf{L}^*$ as $\mathrm{inc}(\mathbf{L}^*) = \max_{j\in\{1,\dots,p\}} \|\mathbb{P}_{\mathbf{L}^*}\mathbf{e}_j\|$, where $\mathbf{e}_j$ is the $j$th canonical basis vector, and $\mathbb{P}_{\mathbf{L}^*}$ is the projection operator onto the row–column space of $\mathbf{L}^*$; the minimum (or maximum) degree of $\mathbf{S}^*$ as $\mathrm{deg}_{min}(\mathbf{S}^*) = \min_{1\le i\le p}\sum_{j=1}^p \mathbb{1}(\mathbf{S}_{ij}^*{\neq}0)$ (or $\mathrm{deg}_{max}(\mathbf{S}^*) = \max_{1\le i\le p}\sum_{j=1}^p \mathbb{1}(\mathbf{S}_{ij}^*{\neq}0)$). The incoherence of $\mathbf{L}^*$ represents the maximum discrepancy between the row–column space of $\mathbf{L}^*$ and the canonical base. According to [19], $\mathrm{inc}(\mathbf{L}^*)$ ranges between $\sqrt{r/p}$ (maximum incoherence) and 1 (minimum incoherence). The minimum (maximum) degree of $\mathbf{S}^*$ represents instead the minimum (maximum) number of non-zeros per row in $\mathbf{S}^*$. We know from [19] that

$$\begin{aligned} &\mathrm{inc}(\mathbf{L}^*) \le \xi(\mathcal{T}(\mathbf{L}^*)) \le 2\mathrm{inc}(\mathbf{L}^*); \\ &\mathrm{deg}_{min}(\mathbf{S}^*) \le \mu(\Omega(\mathbf{S}^*)) \le \mathrm{deg}_{max}(\mathbf{S}^*). \end{aligned} \tag{15}$$

These definitions allow us to properly present the additional assumptions required to ensure the asymptotic consistency of ALCE approach.

**Assumption 5.** For all $p \in \mathbb{N}$, there exist $\kappa_L, \kappa_S > 0$ with $\sqrt{r}\kappa_S/\kappa_L \leq 1/54$, $k_L \geq \sqrt{r}/2$, $\kappa_S \leq \delta_2$, such that $\xi(\mathcal{T}(\mathbf{L}^*)) = \sqrt{r}/\kappa_L p^{\delta_1}$ and $\mu(\Omega(\mathbf{S}^*)) = \kappa_S p^{\delta_1}$.

Assumption 5 essentially imposes the identifiability condition $\xi(\mathcal{T}(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*)) \leq 1/54$ (we refer to Proposition 1 in Appendix A for more details) by controlling the rate of the maximum degree of $\mathbf{S}^*$ *wrt* the dimension $p$. From (15), it is clear that Assumption 2 also implicitly controls the rate of $\mu(\Omega(\mathbf{S}^*))$, by controlling $\deg_{max}(\mathbf{S}^*) = \|\mathbf{S}^*\|_{0,v}$, which is imposed to be not larger than $\delta_2 p^{\delta_1}$ ($\delta_1 \in [0, 1/2]$), thus requiring $\xi(\mathcal{T}(\mathbf{L}^*))$ to scale to $O(p^{-\delta_1})$ to ensure identifiability. This is also why we require $\kappa_S \leq \delta_2$. Assumption 5 is a characterizing feature of ALCE approach compared to alternative approaches in [4,9,22], because those approaches do not explicitly control the underlying algebraic structure in terms of geometric manifolds, as we do to ensure the identifiability of (8) and (9), following [18].

**Example 1.** Let us consider the two extreme values prescribed for $\delta_1$ by Assumption 2. If $\delta_1 = 0$, $\xi(\mathcal{T}(\mathbf{L}^*)) = \sqrt{r}/\kappa_L$ and $\mu(\Omega(\mathbf{S}^*)) = \kappa_S$. This means that $\mathbf{S}^*$ is imposed to be really sparse, because $\|\mathbf{S}^*\|_{0,v} = O(1)$ and consequently $\mu(\Omega(\mathbf{S}^*)) = O(1)$. This case exemplifies a situation where latent eigenvectors are minimally incoherent *wrt* the canonical base, because $\xi(\mathcal{T}(\mathbf{L}^*)) = O(1)$. To tolerate this minimal value of incoherence for $\mathbf{L}^*$, the identifiability condition $\xi(\mathcal{T}(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*)) \leq 1/54$ imposes a minimal number of residual non-zeros in $\mathbf{S}^*$. On the other hand, if $\delta_1 = 1/2$, according to [19] the low rank matrix $\mathbf{L}^*$ is maximally incoherent, which means that the identifiability condition $\xi(\mathcal{T}(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*)) \leq 1/54$ tolerates a maximum number of non-zeros per row $\|\mathbf{S}^*\|_{0,v}$ as large as $O(p^{1/2})$. In plain terms, Assumption 5 controls $\mu(\Omega(\mathbf{S}^*))$ by its upper bound $\|\mathbf{S}^*\|_{0,v}$, which is in turn controlled by Assumption 2.

**Assumption 6.** Define $\psi_0 = 1/\xi(\mathcal{T}(\mathbf{L}^*))\sqrt{\ln(p)/n}$. There exist $\delta_L, \delta_S > 0$ such that (a) the minimum eigenvalue of $\mathbf{L}^*$ ($\lambda_r(\mathbf{L}^*)$) is greater than $\delta_L\psi_0/\xi^2(\mathcal{T}(\mathbf{L}^*))$; (b) the minimum absolute value of the non-zero off-diagonal entries of $\mathbf{S}^*$, $\|\mathbf{S}^*\|_{\min,\mathrm{off}}$, is greater than $\delta_S\psi_0/\mu(\Omega(\mathbf{S}^*))$.

Assumption 6 is the other crucial assumption to ensure algebraic consistency. It prescribes that the smallest latent eigenvalue and the minimum off-diagonal absolute magnitude in the residual component are large enough. In particular, Assumption 6(a) is required for the recovery of the latent rank and Assumption 6(b) is required to recover the residual sparsity pattern.

**Theorem 2.** *Suppose that Assumptions 1–3 and 5–6 hold. Define $\rho_0 = \gamma\psi_0$, where $\gamma \in [9\xi(\mathcal{T}(\mathbf{L}^*)), 1/(6\mu(\Omega(\mathbf{S}^*)))]$. Assume that $\delta_1 \leq \alpha_r/3$ and $\ln(p)/n \to 0$. Then, there exists a positive real $\kappa$ independent of $p$ and $n$ such that, for all $p \in \mathbb{N}$, as $n \to \infty$ the pair of estimators (12) satisfies:*

(i) $\Pr(p^{-\alpha_1}\|\widehat{\mathbf{L}}_{\mathrm{ALCE}} - \mathbf{L}^*\|_2 \leq \kappa\psi_0) \to 1$;
(ii) $\Pr(\|\widehat{\mathbf{S}}_{\mathrm{ALCE}} - \mathbf{S}^*\|_\infty \leq \kappa\rho_0) \to 1$;
(iii) $\Pr(\mathrm{rk}(\widehat{\mathbf{L}}_{\mathrm{ALCE}}) = \mathrm{rk}(\mathbf{L}^*)) \to 1$.

*Further assume that $p^{2-2\delta_1}\ln(p)/n = o(1)$. Then, for all $p \in \mathbb{N}$, as $n \to \infty$ it holds: (iv)* $\Pr(\mathrm{sgn}(\widehat{\mathbf{S}}_{\mathrm{ALCE}}) = \mathrm{sgn}(\mathbf{S}^*)) \to 1$.

Theorem 2, proved in A, is explained in detail below.

**Remark 2.** Theorem 2 differs from Theorem 1 in [23] in what follows. First, Assumption 1, unlike Assumption 1 in [23], now incorporates differentiated speeds of divergence among latent eigenvalues. Second, the sparsity conditions in Assumption 2 have been completely reshaped and simplified, as they are now controlled by means of $\|\mathbf{S}^*\|_{0,v}$, $\|\mathbf{S}^*\|_\infty$, $\|\mathbf{S}^*\|_{\min,\mathrm{off}}$, and $\mathrm{tr}(\mathbf{S}^*)$, unlike the corresponding Assumption 4 in [23]. Third, the theorem is now proved for all $p \in \mathbb{N}$ as $n \to \infty$, thanks to the new condition $\delta_1 \leq \alpha_r/3$, which regulates compatibility between Assumption 1(a) and 6(a) (see Remark 3). Fourth, thanks to these updates, Assumptions 5 and 6 in [23] are no longer needed.

**Remark 3.** The conditions $\delta_1 \leq \alpha_r/3$ and $\ln(p)/n \to 0$ are imposed to ensure that Assumption 6(a) is compatible with Assumption 1(a), which requires, imposing Assumption 5,

$$\lambda_r(\mathbf{L}^*) \geq \delta_L \frac{\psi_0}{\xi^2(\mathcal{T}(\mathbf{L}^*))} \geq \delta_L \left(\frac{\sqrt{r}}{\kappa_L}\right)^3 p^{3\delta_1}\sqrt{\frac{\ln(p)}{n}} \tag{16}$$

under $\lambda_r(\mathbf{L}^*) \simeq p^{\alpha_r}$. Compatibility is always ensured for all $p \in \mathbb{N}$ if $0 \leq \delta_1 \leq \alpha_r/3$, because $n \to \infty$ and $r$ is finite by Assumption 1(b). If $\delta_1 \in (\alpha_r/3, 1/2]$, Theorem 2 continues to hold as $p \to \infty$ with the stricter requirement $p^{6\delta_1-2\alpha_r}/n \to 0$, that comes from the need to ensure that (16) holds under $\lambda_r(\mathbf{L}^*) \simeq p^{\alpha_r}$. This means that a large $\delta_1$, corresponding to more non-zeros in $\mathbf{S}^*$, complicates the recovery of both components in that the condition $n > p$ may be needed to ensure consistency if $\delta_1 \in (\alpha_r/3, 1/2]$.

**Remark 4.** Assumptions 6(b) and 2(b) are always compatible, as Assumption 5 ensures that

$$0 < 54\delta_S\sqrt{\frac{\ln(p)}{n}} \leq \frac{\delta_S}{\xi(\mathcal{T}(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*))}\sqrt{\frac{\ln(p)}{n}} = \delta_S\frac{\psi_0}{\mu(\Omega(\mathbf{S}^*))} < \|\mathbf{S}^*\|_{\min,\text{off}} < \|\mathbf{S}^*\|_\infty = O(1),$$

which is always verified for all $p \in \mathbb{N}$ as $n \to \infty$ and $\ln(p)/n = o(1)$. Similarly, Assumption 6(b) and 2(c) are always compatible as long as $p^{2-2\delta_1}\ln(p)/n = o(1)$, because $54\delta_S\sqrt{\ln(p)/n} < \|\mathbf{S}^*\|_{\min,\text{off}} = o(1/p^{1-\delta_1})$ is always verified for all $p \in \mathbb{N}$ as $n \to \infty$ and $\sqrt{\ln(p)/n} = o(1/p^{1-\delta_1})$, which leads to the condition $p^{2-2\delta_1}\ln(p)/n = o(1)$. Note that such condition is only required to ensure the validity of clause 4 in Theorem 2, i.e., the sparsistency of $\widehat{\mathbf{S}}_{\text{ALCE}}$, because of the specific role of Assumption 6(b) (cf. Theorem 2 in [23]).

**Remark 5.** Assumption 3 imposes sub-exponential tails to $f_{k,i}$, $i = \{1, \dots, r\}$, and $\epsilon_{k,j}$, $j = \{1, \dots, p\}$. This condition is satisfied by the Gaussian distribution but is more general in nature. It is needed to apply Lemmas 3 and 4(a) of [22], and (12) in [12] (together with Assumption 2(b), cf. Section 2 in [12]). Assumptions 1, 2, and 3 are required to prove Lemma 1, which is essential for all the results of this paper. Assumption 4 is a further condition needed to explicitly control cross-sectional dependence. It would be implied by Assumption 3 in the case of cross-sectional independence. It is not needed to prove Theorem 2, because, unlike [22], it is only needed to prove consistency of ALCE estimated loadings and factor scores (see Theorem 3), but not to prove consistency of ALCE covariance matrix.

**Remark 6.** Parts (i) and (ii) of Theorem 2 establish the parametric consistency of the pair of ALCE estimators (12) in $g_\gamma$-norm. The statements of parts (i) and (ii) follow in fact from the underlying thesis of Theorem 2, that is

$$g_\gamma(\widehat{\mathbf{S}}_{\text{ALCE}} - \mathbf{S}^*, \widehat{\mathbf{L}}_{\text{ALCE}} - \mathbf{L}^*) \leq \psi_0,$$

which in turn leads to $p^{-\delta_1}g_\gamma(\widehat{\mathbf{S}}_{\text{ALCE}} - \mathbf{S}^*, \widehat{\mathbf{L}}_{\text{ALCE}} - \mathbf{L}^*) = O_p\left(\sqrt{\ln(p)/n}\right)$, where $\psi = p^{\alpha_1}\psi_0$, $\rho_0 = \rho$, $\gamma = \psi_0/\rho_0$, and $\psi$ and $\rho$ are the threshold parameters in (6). Remark 6 expresses a consistency result for $\widehat{\mathbf{L}}_{\text{ALCE}}$ and $\widehat{\mathbf{S}}_{\text{ALCE}}$ jointly considered in $g_\gamma$-norm (see Definition 1), which is the dual norm of the composite penalty $\psi_0\|\cdot\|_* + \rho_0\|\cdot\|_1$. More explanations can be found in the proof of Theorem 2 and in [18].

**Remark 7.** Parts (iii) and (iv) establish conditions a and b of Definition 1. This result comes at the price of bounding the degree of transversality of the geometric manifolds where $\mathbf{L}^*$ and $\mathbf{S}^*$ lie, i.e. the matrix varieties $\mathcal{L}(r)$ and $\mathcal{S}(s)$. This is precisely the role of Assumption 5, that, together with Assumption 2, is responsible for the factor $p^{\delta_1}$ that appears in the error rate of $\widehat{\mathbf{L}}_{\text{ALCE}}$ and $\widehat{\mathbf{S}}_{\text{ALCE}}$ (see Corollary 1) in spectral norm. The lower bounds imposed by Assumption 6 on $\lambda_r(\mathbf{L}^*)$ and $\|\mathbf{S}\|_{\min,\text{off}}$ are then the key to ensure the identifiability of $\mathcal{L}(r)$ and $\mathcal{S}(s)$ with high probability, where the probability rate also depends on Assumptions 1 and 3.

**Corollary 1.** *Under the assumptions of Theorem 2, for all $p \in \mathbb{N}$, as $n \to \infty$ it holds:* (i) $\Pr(p^{-\delta_1}\|\widehat{\mathbf{S}}_{\text{ALCE}} - \mathbf{S}^*\|_2 \leq \kappa\sqrt{\ln(p)/n}) \to 1$; (ii) $\Pr(p^{-(\alpha_1+\delta_1)}\|\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}} - \boldsymbol{\Sigma}^*\|_2 \leq \kappa\sqrt{\ln(p)/n}) \to 1$; (iii) $\Pr(\lambda_p(\widehat{\mathbf{S}}_{\text{ALCE}}) > 0) \to 1$; (iv) $\Pr(\lambda_p(\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}) > 0) \to 1$. *In addition, supposing that $\lambda_p(\mathbf{S}^*) = O(p^{\alpha_1-1-\varepsilon})$ and $\lambda_p(\boldsymbol{\Sigma}^*) = O(p^{\alpha_1-1-\varepsilon})$ for some $\varepsilon > 0$, the following statements hold for all $p \in \mathbb{N}$ as $n \to \infty$:* (v) $\Pr\left(p^{-\delta_1}p^{-2(1-\alpha_1+\varepsilon)}\|\widehat{\mathbf{S}}_{\text{ALCE}}^{-1} - \mathbf{S}^{*-1}\|_2 \leq \kappa\sqrt{\ln(p)/n}\right) \to 1$; (vi) $\Pr\left(p^{-(\alpha_1+\delta_1)}p^{-2(1-\alpha_1+\varepsilon)}\|\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}^{-1} - \boldsymbol{\Sigma}^{*-1}\|_2 \leq \kappa\sqrt{\ln(p)/n}\right) \to 1.$

Corollary 1, proved in A, is clarified by means of the following remarks.

**Remark 8.** Conditions $\lambda_p(\mathbf{S}^*) = O(p^{\alpha_1-1-\varepsilon})$ and $\lambda_p(\boldsymbol{\Sigma}^*) = O(p^{\alpha_1-1-\varepsilon})$ for some $\varepsilon > 0$ are needed to ensure compatibility with Assumption 2(d).

**Remark 9.** Parts (iii) and (iv) ensure that condition c of Definition 1 holds. Together with parts (iii) and (iv) of Theorem 2, they ensure the algebraic consistency of the pair of ALCE estimators (12) according to Definition 1.

We can now define the projection matrix onto the orthogonally rotated true factor space as $\mathbf{H}_{\text{ALCE1}} = n^{-1}(\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}})^{-1}\widehat{\mathbf{F}}_{\text{ALCE1}}^\top\mathbf{F}\mathbf{B}^\top\mathbf{B}$, and state the main theorem of this section.

**Theorem 3.** *Suppose that all the assumptions and conditions of Theorem 2 hold. Then, for the estimators $\widehat{\mathbf{B}}_{\text{ALCE1}} = [\widehat{\mathbf{b}}_{\text{ALCE1},1}\dots\widehat{\mathbf{b}}_{\text{ALCE1},p}]$ (where $\widehat{\mathbf{B}}_{\text{ALCE1}}$ is $p \times r^*$ with $r^* = \text{rk}(\widehat{\mathbf{L}}_{\text{ALCE}}) = \widehat{r}_A$) and $\widehat{\mathbf{F}}_{\text{ALCE1}} = [\widehat{\mathbf{f}}_{\text{ALCE1},1}\dots\widehat{\mathbf{f}}_{\text{ALCE1},n}]$, Theorem 1 holds with $\mathbf{H}_{\text{ALCE1}}$ in place of $\mathbf{H}_{\text{OLS1}}$.*

Theorem 3, proved in A, provides the uniform error rates in Euclidean norm for loading vectors, factor scores and commonalities estimated by penalized OLS via ALCE approach under the generalized spikiness regime for latent eigenvalues and the generalized sparsity regime for the residual component, relying on the fact that $\widehat{r}_A = r$ with probability tending to one under all the assumptions and conditions of Theorem 2.

**Example 2.** The conditions $\delta_1 \leq \alpha_r/3$ of Theorem 2 and $\alpha_1 - \alpha_r \leq \delta_1$ of Theorem 1 give rise to the inequality $\alpha_1 - \alpha_r \leq \alpha_r/3$, which leads to $\alpha_1 - \frac{4}{3}\alpha_r \leq 0$. Therefore, for Theorem 3 to hold, such condition must be satisfied. For instance, if $\alpha_1 = 1$, it must hold $\alpha_r \in (3/4, 1]$; if $\alpha_1 = 3/4$, it must hold $\alpha_r \in (9/16, 3/4]$; if $\alpha_1 = 2/3$, it must hold $\alpha_r \in (1/2, 2/3]$. The limit case (ruled out by Assumption 1(a)) is $\alpha_r = \alpha_1 = 1/2$.

## 4. Estimation in the finite sample: UNALCE

In Section 3, we derived the asymptotic consistency of OLS and ALCE factor model estimates assuming that latent factors are intermediately pervasive and the residual covariance is intermediately sparse *wrt* the dimension $p$. In this section, we discuss the optimality properties of factor model estimates based on heuristics (6) when the parameters $p$ and $n$ are fixed.

Let us recall the conclusions of Theorem 2 and Corollary 1. Theorem 2 states that the pair of solutions (12) under Assumptions 1–3 and 5–6 is parametrically consistent and recovers the true latent rank and the residual sparsity pattern with probability tending to one, provided that the sample size $n$ and the thresholds $\psi$ and $\rho$ lie in a specific range *wrt* the dimension $p$. The resulting estimators are named $\widehat{\mathbf{L}}_{\text{ALCE}}$, $\widehat{\mathbf{S}}_{\text{ALCE}}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}$. Corollary 1 provides the error bounds in spectral norm and the invertibility conditions for $\widehat{\mathbf{S}}_{\text{ALCE}}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}$. Theorem 2 and Corollary 1 together mean that ALCE estimates are algebraically consistent.

Once fixed the dimension $p$ and the sample size $n$, our new aim is to prove that ALCE estimates can be re-optimized as much as possible in the following sense:

$$\min_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A),\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \frac{1}{2}\|\boldsymbol{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_2, \tag{17}$$

where $\widehat{\mathcal{S}}_{diag}$ is the following set of matrices:

$$\widehat{\mathcal{S}}_{diag} = \{\mathbf{S} \in \mathbb{R}^{p\times p} \mid \mathrm{diag}(\mathbf{L}) + \mathrm{diag}(\mathbf{S}) = \mathrm{diag}(\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}), \mathrm{off}-\mathrm{diag}(\mathbf{S}) = \mathrm{off}-\mathrm{diag}(\widehat{\mathbf{S}}_{\text{ALCE}}), \mathbf{L} \in \widehat{\mathcal{L}}(\widehat{r}_A)\}.$$

In synthesis, for any threshold pair $(\breve{\psi}, \breve{\rho})$ satisfying the conditions of Theorem 2, the recovered matrix varieties $\widehat{\mathcal{L}}(\widehat{r}_A)$ and $\widehat{\mathcal{S}}(\widehat{s}_A)$ (see (13) and (14)) are first recovered by solving the problem $\min_{\mathbf{L},\mathbf{S}} \frac{1}{2}\|\boldsymbol{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_F^2$ under the constraint that $\breve{\psi}\|\mathbf{L}\|_* + \breve{\rho}\|\mathbf{S}\|_1$ is minimum, which leads to the pair of ALCE estimates (12). In a second step, the estimates are then re-optimized by solving problem (17), which entirely depends on the sample covariance matrix $\boldsymbol{\Sigma}_n$, and restricts the low rank solution to lie in $\widehat{\mathcal{L}}(\widehat{r}_A)$ and the residual solution to have the same off-diagonal elements of $\widehat{\mathbf{S}}_{\text{ALCE}}$ and a constrained diagonal *wrt* $\mathrm{diag}(\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}})$.

Relying on the parametric guarantees offered by $\widehat{\mathcal{L}}(\widehat{r}_A)$ and $\widehat{\mathcal{S}}(\widehat{s}_A)$, and conditioning upon the recovered latent rank $\widehat{r}_A$, the residual sparsity pattern $\mathrm{sgn}(\widehat{\mathbf{S}}_{\text{ALCE}})$ and the first step optimization in (12), we prove that it is possible to re-optimize the pair of estimates $(\widehat{\mathbf{L}}_{\text{ALCE}}, \widehat{\mathbf{S}}_{\text{ALCE}})$ to improve the overall fitting as much as possible, constraining the solutions to lie on $\widehat{\mathcal{L}}(\widehat{r}_A)$ and $\widehat{\mathcal{S}}_{diag}$.

**Theorem 4.** *Let us define*

$$\widehat{\mathbf{L}}_{\text{UNALCE}} = \widehat{\mathbf{U}}_{\text{ALCE}}(\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}} + \breve{\psi}\mathbf{I}_r)\widehat{\mathbf{U}}_{\text{ALCE}}^\top,$$

*where $\breve{\psi} > 0$ is a chosen eigenvalue threshold parameter, and $\widehat{\mathbf{S}}_{\text{UNALCE}}$ such that*

$$\mathrm{diag}(\widehat{\mathbf{S}}_{\text{UNALCE}}) = \mathrm{diag}(\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}) - \mathrm{diag}(\widehat{\mathbf{L}}_{\text{UNALCE}}),$$

*and*

$$\mathrm{off}-\mathrm{diag}(\widehat{\mathbf{S}}_{\text{UNALCE}}) = \mathrm{off}-\mathrm{diag}(\widehat{\mathbf{S}}_{\text{ALCE}}).$$

*Then, assuming the sample covariance matrix $\boldsymbol{\Sigma}_n$ as fixed, under all the assumptions and conditions of Theorem 2 the following statements hold:* (i) $\arg\min\max_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A)} \|\mathbf{L}-\mathbf{L}^*\|_2 = \widehat{\mathbf{L}}_{\text{UNALCE}}$; (ii) $\arg\min\max_{\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|\mathbf{S}-\mathbf{S}^*\|_2 = \widehat{\mathbf{S}}_{\text{UNALCE}}$; (iii) $\arg\min\max_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A),\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|(\mathbf{L}+\mathbf{S})-\boldsymbol{\Sigma}^*\|_2 = \widehat{\boldsymbol{\Sigma}}_{\text{UNALCE}}$. *More, if all the conditions of Corollary 1 and Theorem 3 hold, then:* (iv) $\arg\min\max_{\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|\mathbf{S}^{-1}-\mathbf{S}^{*-1}\|_2 = \widehat{\mathbf{S}}_{\text{UNALCE}}^{-1}$; (v) $\arg\min\max_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A),\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|(\mathbf{L}+\mathbf{S})^{-1}-\boldsymbol{\Sigma}^{*-1}\|_2 = \widehat{\boldsymbol{\Sigma}}_{\text{UNALCE}}^{-1}$.

**Remark 10.** In part (i), the minimax can actually be replaced by a simple minimum if $\delta_1 > 0$ and $p$ is large enough, due to algebraic consistency (see Proposition 3 and the proof of Theorem 4 in A).

Theorem 4 states that the UNALCE estimates of $\mathbf{L}^*$, $\mathbf{S}^*$, $\boldsymbol{\Sigma}^*$, $\mathbf{S}^{*-1}$, $\boldsymbol{\Sigma}^{*-1}$ show the minimax errors in spectral norm among algebraically consistent estimates, assuming the sample covariance matrix as fixed. The UNALCE procedure has the effect to un-shrink the eigenvalues of $\widehat{\mathbf{L}}_{\text{ALCE}}$, and to update the diagonal of $\widehat{\mathbf{S}}_{\text{ALCE}}$, while saving the recovered sparsity pattern $\mathrm{sgn}(\widehat{\mathbf{S}}_{\text{ALCE}})$, that is equal to $\mathrm{sgn}(\widehat{\mathbf{S}}_{\text{UNALCE}})$ by definition.

## 5. Bartlett's and Thomson's estimation

In this section, we prove that Bartlett's and Thomson's factor scores estimators based on UNALCE covariance matrix estimates show the tightest possible error bounds in Euclidean norm, given the finite sample. First, we state the optimality of the UNALCE loading matrix estimator, $\widehat{\mathbf{B}}_{\mathrm{UNALCE}} = \widehat{\mathbf{U}}_{\mathrm{UNALCE}}\widehat{\boldsymbol{\Lambda}}_{\mathrm{UNALCE}}^{1/2}$, with $\widehat{\mathbf{U}}_{\mathrm{UNALCE}} = \widehat{\mathbf{U}}_{\mathrm{ALCE}}$ and $\widehat{\boldsymbol{\Lambda}}_{\mathrm{UNALCE}} = \widehat{\boldsymbol{\Lambda}}_{\mathrm{ALCE}} + \breve{\psi}\mathbf{I}_r$, by the following Corollary.

**Corollary 2.** *Let $\mathbf{H}_r$ be any $r \times r$ orthogonal matrix. Under the assumptions of Theorem 4, then*

$$\arg\min \max_{\widehat{\mathbf{B}}|\widehat{\mathbf{L}}=\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top \in \widehat{\mathcal{L}}(\widehat{r}_A)} \|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}_r\| = \widehat{\mathbf{B}}_{\mathrm{UNALCE}}.$$

**Proof.** Let us define the loading matrix estimator $\widehat{\mathbf{B}} = \widehat{\mathbf{U}}_L\widehat{\boldsymbol{\Lambda}}_L^{\frac{1}{2}}$, with $\widehat{\mathbf{U}}_L$ $p \times r$ semi-orthogonal matrix and $\widehat{\boldsymbol{\Lambda}}_L$ diagonal $r \times r$ matrix such that $\widehat{\mathbf{L}} = \widehat{\mathbf{U}}_L\widehat{\boldsymbol{\Lambda}}_L\widehat{\mathbf{U}}_L^\top \in \widehat{\mathcal{L}}(\widehat{r}_A)$. It then holds

$$\|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}_r\| = \lambda_1(\widehat{\mathbf{U}}_L\widehat{\boldsymbol{\Lambda}}_L^{\frac{1}{2}} - \mathbf{U}_L\boldsymbol{\Lambda}_L^{\frac{1}{2}}\mathbf{H}_r).$$

At this stage, we observe that, under the assumptions of Theorem 4,

$$\arg\min \max_{\widehat{\mathbf{L}}\in\widehat{\mathcal{L}}(\widehat{r}_A)} \|\widehat{\mathbf{L}} - \mathbf{L}^*\|_2 = \arg\min_{(\widehat{\mathbf{U}}_L,\widehat{\boldsymbol{\Lambda}}_L)} \max_{\text{s.t. } \widehat{\mathbf{L}}=\widehat{\mathbf{U}}_L\widehat{\boldsymbol{\Lambda}}_L\widehat{\mathbf{U}}_L^\top\in\widehat{\mathcal{L}}(\widehat{r}_A)} \lambda_1(\widehat{\mathbf{U}}_L\widehat{\boldsymbol{\Lambda}}_L\widehat{\mathbf{U}}_L^\top - \mathbf{U}_L\boldsymbol{\Lambda}_L\mathbf{U}_L^\top) = \widehat{\mathbf{L}}_{\mathrm{UNALCE}},$$

if Theorem 2 holds. Setting $\widehat{\mathbf{L}}_{\mathrm{UNALCE}} = \widehat{\mathbf{U}}_{\mathrm{UNALCE}}\widehat{\boldsymbol{\Lambda}}_{\mathrm{UNALCE}}\widehat{\mathbf{U}}_{\mathrm{UNALCE}}^\top$, it follows that the thesis

$$(\widehat{\mathbf{U}}_{\mathrm{UNALCE}}, \widehat{\boldsymbol{\Lambda}}_{\mathrm{UNALCE}}) = \arg\min_{(\widehat{\mathbf{U}}_L,\widehat{\boldsymbol{\Lambda}}_L)} \max_{\text{s.t. } \widehat{\mathbf{L}}=\widehat{\mathbf{U}}_L\widehat{\boldsymbol{\Lambda}}_L\widehat{\mathbf{U}}_L^\top\in\widehat{\mathcal{L}}(\widehat{r}_A)} \lambda_1(\widehat{\mathbf{U}}_L\widehat{\boldsymbol{\Lambda}}_L^{\frac{1}{2}} - \mathbf{U}_L\boldsymbol{\Lambda}_L^{\frac{1}{2}}\mathbf{H}_r),$$

descends from part (i) of Theorem 4. □

Corollary 2 is a direct consequence of Theorem 4. It shows that $\widehat{\mathbf{B}}_{\mathrm{UNALCE}}$ is the optimal loading matrix estimator given the finite sample in terms of fitting performance within the class of algebraically consistent estimates for $\mathbf{B}$, under all the assumptions and conditions of Theorem 4. If $\delta_1 > 0$, the minimax can be replaced by a simple minimum as $p$ is large enough, due to algebraic consistency (see Remark 10).

Then, we define Bartlett's factor scores estimates for the observation $k$, $k \in \{1, \ldots, n\}$, as follows: $\widehat{\mathbf{f}}_{k,B} = (\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\mathbf{x}_k$. They simply are the Generalized Least Squares (GLS) factor scores estimates. The true Bartlett's factors are defined as $\mathbf{f}_{k,B} = (\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{x}_k$. The following result for Bartlett's factor scores and commonalities based on UNALCE covariance matrix estimates holds.

**Theorem 5.** *Let us denote by $\mathbf{H}_r$ any $r \times r$ orthogonal matrix. Under the assumptions of Theorem 4, for $k \in \{1, \ldots, n\}$ the minimax $\min\max_{\widehat{\mathbf{B}},\widehat{\mathbf{L}}=\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top\in\widehat{\mathcal{L}}(\widehat{r}_A),\widehat{\mathbf{S}}\in\widehat{\mathcal{S}}_{diag}} \|\widehat{\mathbf{f}}_{k,B} - \mathbf{H}_r\mathbf{f}_{k,B}\|$ and $\min\max_{\widehat{\mathbf{B}},\widehat{\mathbf{L}}=\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top\in\widehat{\mathcal{L}}(\widehat{r}_A),\widehat{\mathbf{S}}\in\widehat{\mathcal{S}}_{diag}} \|\widehat{\mathbf{B}\widehat{\mathbf{f}}}_{k,B} - \mathbf{B}\mathbf{f}_{k,B}\|$ are achieved for $\widehat{\mathbf{B}} = \widehat{\mathbf{B}}_{\mathrm{UNALCE}}$ and $\widehat{\mathbf{S}} = \widehat{\mathbf{S}}_{\mathrm{UNALCE}}$.*

**Proof.** We start considering the loss $\|\widehat{\mathbf{f}}_{k,B} - \mathbf{H}_r\mathbf{f}_{k,B}\|$. Since $\widehat{\mathbf{f}}_{k,B} = (\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\mathbf{x}_k$, $k \in \{1, \ldots, n\}$, under the assumptions of Theorem 4 we get

$$\|\widehat{\mathbf{f}}_{k,B} - \mathbf{H}_r\mathbf{f}_{k,B}\| \leq \|(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - \mathbf{H}_r(\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{S}^{*-1}\| \times \|\mathbf{x}_k\|.$$

We thus focus on

$$(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - \mathbf{H}_r(\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{S}^{*-1}. \tag{18}$$

We note that, since $\mathbf{H}_r$ is orthogonal, (18) is equivalent to

$$(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - (\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\mathbf{B}^\top\mathbf{S}^{*-1}.$$

Then, via some algebra, we obtain

$$\begin{aligned}
&(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - (\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\mathbf{B}^\top\mathbf{S}^{*-1}\\
&= (\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - (\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\mathbf{B}^\top\mathbf{S}^{*-1} + (\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\mathbf{B}^\top\mathbf{S}^{*-1} - (\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\mathbf{B}^\top\mathbf{S}^{*-1}\\
&= \left[(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1} - (\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\right]\mathbf{B}^\top\mathbf{S}^{*-1} + (\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - \mathbf{B}^\top\mathbf{S}^{*-1}).
\end{aligned}$$

At this point, we add and subtract the quantity $(\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\left[\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - \mathbf{B}^\top\mathbf{S}^{*-1}\right]$, and we get:

$$\begin{aligned}
&(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - (\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\mathbf{B}^\top\mathbf{S}^{*-1} = \left[(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1} - (\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\right]\mathbf{B}^\top\mathbf{S}^{*-1}\\
&+ \left[(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1} - (\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\widehat{\mathbf{B}}^{-1}\right](\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - \mathbf{B}^\top\mathbf{S}^{*-1}) + (\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1} - \mathbf{B}^\top\mathbf{S}^{*-1}).
\end{aligned} \tag{19}$$

We first focus on the error matrix $(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1})$. We write

$$\begin{aligned}(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1}) &= \widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1}+\mathbf{B}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\widehat{\mathbf{S}}^{-1}(\widehat{\mathbf{B}}-\mathbf{B})^\top\widehat{\mathbf{S}}^{-1}+\mathbf{B}^\top(\widehat{\mathbf{S}}^{-1}-\mathbf{S}^{*-1})\\ &=(\widehat{\mathbf{B}}-\mathbf{B})^\top\widehat{\mathbf{S}}^{-1}+\mathbf{B}^\top(\widehat{\mathbf{S}}^{-1}-\mathbf{S}^{*-1})-(\widehat{\mathbf{B}}-\mathbf{B})^\top\mathbf{S}^{*-1}+(\widehat{\mathbf{B}}-\mathbf{B})^\top\mathbf{S}^{*-1}\\ &=(\widehat{\mathbf{B}}-\mathbf{B})^\top(\widehat{\mathbf{S}}^{-1}-\mathbf{S}^{*-1})+\mathbf{B}^\top(\widehat{\mathbf{S}}^{-1}-\mathbf{S}^{*-1})+(\widehat{\mathbf{B}}-\mathbf{B})^\top\mathbf{S}^{*-1}. \end{aligned}\tag{20}$$

By the triangular inequality, Corollary 2 and Theorem 4 (part (iv)), it follows from (20) that

$$\arg\min_{\widehat{\mathbf{B}},\widehat{\mathbf{L}}=\widehat{\mathbf{B}\mathbf{B}}^\top\in\widehat{\mathcal{L}}(\widehat{r}_A),\widehat{\mathbf{S}}\in\widehat{\mathcal{S}}_{diag}}\max(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1})=\left(\widehat{\mathbf{B}}_{\mathrm{UNALCE}},\widehat{\mathbf{S}}_{\mathrm{UNALCE}}\right).\tag{21}$$

We now focus on the term $\left[(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}-(\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\right]$. We recall that

$$\begin{aligned}\|(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1}-(\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1}\| &\le \lambda_1((\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)^{-1})\lambda_1((\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})^{-1})\|(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})-(\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)\|\\ &\le \frac{\|(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})-(\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)\|}{\lambda_r(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})\lambda_r(\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)}.\end{aligned}\tag{22}$$

At this stage, we need to bound $\|(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}})-(\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)\|$. We consider the error matrix $\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}}-\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top$. We write

$$\begin{aligned}\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}}-\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top &= \widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}}-\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top+\mathbf{B}^\top\mathbf{S}^{*-1}\widehat{\mathbf{B}}-\mathbf{B}^\top\mathbf{S}^{*-1}\widehat{\mathbf{B}}\\ &=(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1})\widehat{\mathbf{B}}+\mathbf{B}^\top\mathbf{S}^{*-1}(\widehat{\mathbf{B}}-\mathbf{B}\mathbf{H}_r)\\ &=(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1})\widehat{\mathbf{B}}+\mathbf{B}^\top\mathbf{S}^{*-1}(\widehat{\mathbf{B}}-\mathbf{B}\mathbf{H}_r)-(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1})\mathbf{B}+(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1})\mathbf{B}\\ &=(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1})(\widehat{\mathbf{B}}-\mathbf{B})+\mathbf{B}^\top\mathbf{S}^{*-1}(\widehat{\mathbf{B}}-\mathbf{B}\mathbf{H}_r)+(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}-\mathbf{B}^\top\mathbf{S}^{*-1})\mathbf{B}.\end{aligned}\tag{23}$$

Now, by the triangular inequality, Corollary 2 and Theorem 4 (part (iv)), from (22) and (23), we get

$$\arg\min_{\widehat{\mathbf{B}},\widehat{\mathbf{L}}=\widehat{\mathbf{B}\mathbf{B}}^\top\in\widehat{\mathcal{L}}(\widehat{r}_A),\widehat{\mathbf{S}}\in\widehat{\mathcal{S}}_{diag}}\max(\widehat{\mathbf{B}}^\top\widehat{\mathbf{S}}^{-1}\widehat{\mathbf{B}}-\mathbf{B}^\top\mathbf{S}^{*-1}\mathbf{B}\mathbf{H}_r^\top)=\left(\widehat{\mathbf{B}}_{\mathrm{UNALCE}},\widehat{\mathbf{S}}_{\mathrm{UNALCE}}\right).\tag{24}$$

Finally, from (19), (21), (22) and (24) the statement follows. □

Theorem 5 states that Bartlett's factor scores and commonalities estimated by UNALCE approach are the most precise given the finite sample within the sets of algebraically consistent estimates for $\mathbf{B}$ and $\mathbf{S}^*$, under the assumptions and conditions of Theorem 4.

Suppose now that the bivariate distribution $(\mathbf{x}_k,\mathbf{f}_k)$, $k\in\{1,\dots,n\}$, is multivariate normal (denoted as MNV):

$$\begin{pmatrix}\mathbf{x}_k\\ \mathbf{f}_k\end{pmatrix}\sim\mathrm{MNV}\left[\begin{pmatrix}\mathbf{0}_p\\ \mathbf{0}_r\end{pmatrix},\begin{pmatrix}\mathbf{B}\mathbf{B}^\top+\mathbf{S}^* & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{I}_r\end{pmatrix}\right].$$

As a consequence, from the Bayesian point of view, we can derive the following a posteriori expected value for $\mathbf{f}_k$:

$$\mathrm{E}[\mathbf{f}_k|\mathbf{x}_k]=\mathbf{B}^\top(\mathbf{B}\mathbf{B}^\top+\mathbf{S}^*)^{-1}\mathbf{x}_k.$$

Thomson's estimates of factor scores are the estimates of such expected value: $\widehat{\mathbf{f}}_{k,T}=\widehat{\mathbf{B}}^\top(\widehat{\mathbf{B}\mathbf{B}}^\top+\widehat{\mathbf{S}})^{-1}\mathbf{x}_k$, $k\in\{1,\dots,n\}$. The corresponding Thomson's true factors are defined as $\mathbf{f}_{k,T}=\mathbf{B}^\top(\mathbf{B}\mathbf{B}^\top+\mathbf{S}^*)^{-1}\mathbf{x}_k$, $k\in\{1,\dots,n\}$. The following theorem on the performance of Thomson's estimates of factor scores and commonalities based on UNALCE holds.

**Theorem 6.** *Let us denote by $\mathbf{H}_r$ any $r\times r$ orthogonal matrix. Under the assumptions of Theorem 4, for $k\in\{1,\dots,n\}$ the minimax* $\min\max_{\widehat{\mathbf{B}},\widehat{\mathbf{L}}=\widehat{\mathbf{B}\mathbf{B}}^\top\in\widehat{\mathcal{L}}(\widehat{r}_A),\widehat{\mathbf{S}}\in\widehat{\mathcal{S}}_{diag}}\|\widehat{\mathbf{f}}_{k,T}-\mathbf{H}_r\mathbf{f}_{k,T}\|$ *and* $\min\max_{\widehat{\mathbf{B}},\widehat{\mathbf{L}}=\widehat{\mathbf{B}\mathbf{B}}^\top\in\widehat{\mathcal{L}}(\widehat{r}_A),\widehat{\mathbf{S}}\in\widehat{\mathcal{S}}_{diag}}\|\widehat{\mathbf{B}\mathbf{f}}_{k,T}-\mathbf{B}\mathbf{f}_{k,T}\|$ *are achieved for* $\widehat{\mathbf{B}}=\widehat{\mathbf{B}}_{\mathrm{UNALCE}}$ *and* $\widehat{\mathbf{S}}=\widehat{\mathbf{S}}_{\mathrm{UNALCE}}$.

**Proof.** We start considering the loss $\|\widehat{\mathbf{f}}_{k,T}-\mathbf{f}_{k,T}\|$. For the definition of $\widehat{\mathbf{f}}_{k,T}$, under the assumptions of Theorem 4 we get

$$\|\widehat{\mathbf{f}}_{k,T}-\mathbf{f}_{k,T}\|\le\|\widehat{\mathbf{B}}^\top\widehat{\boldsymbol{\Sigma}}^{-1}-\mathbf{B}^\top\boldsymbol{\Sigma}^{*-1}\|\cdot\|\mathbf{x}_k\|.$$

We first focus on the error matrix $(\widehat{\mathbf{B}}^\top\widehat{\boldsymbol{\Sigma}}^{-1}-\mathbf{B}^\top\boldsymbol{\Sigma}^{*-1})$. We write

$$\begin{aligned}(\widehat{\mathbf{B}}^\top\widehat{\boldsymbol{\Sigma}}^{-1}-\mathbf{B}^\top\boldsymbol{\Sigma}^{*-1}) &= \widehat{\mathbf{B}}^\top\widehat{\boldsymbol{\Sigma}}^{-1}-\mathbf{B}^\top\boldsymbol{\Sigma}^{*-1}+\mathbf{B}^\top\widehat{\boldsymbol{\Sigma}}^{-1}-\mathbf{B}^\top\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\mathbf{B}}-\mathbf{B})^\top\widehat{\boldsymbol{\Sigma}}^{-1}+\mathbf{B}^\top(\widehat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{*-1})\\ &=(\widehat{\mathbf{B}}-\mathbf{B})^\top\widehat{\boldsymbol{\Sigma}}^{-1}+\mathbf{B}^\top(\widehat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{*-1})-(\widehat{\mathbf{B}}-\mathbf{B})^\top\boldsymbol{\Sigma}^{*-1}+(\widehat{\mathbf{B}}-\mathbf{B})^\top\boldsymbol{\Sigma}^{*-1}\\ &=(\widehat{\mathbf{B}}-\mathbf{B})^\top(\widehat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{*-1})+\mathbf{B}^\top(\widehat{\boldsymbol{\Sigma}}^{-1}-\boldsymbol{\Sigma}^{*-1})+(\widehat{\mathbf{B}}-\mathbf{B})^\top\boldsymbol{\Sigma}^{*-1}.\end{aligned}\tag{25}$$

By the triangular inequality, Corollary 2 and Theorem 4 (part (v)), it follows from (25) that

$$\arg\min_{\widehat{\mathbf{B}},\widehat{\mathbf{L}}=\widehat{\mathbf{B}\mathbf{B}}^\top\in\widehat{\mathcal{L}}(\widehat{r}_A),\widehat{\mathbf{S}}\in\widehat{\mathcal{S}}_{diag}}\max(\widehat{\mathbf{B}}^\top(\widehat{\mathbf{L}}+\widehat{\mathbf{S}})^{-1}-\mathbf{B}^\top(\mathbf{L}^*+\mathbf{S}^*)^{-1})=\left(\widehat{\mathbf{B}}_{\mathrm{UNALCE}},\widehat{\boldsymbol{\Sigma}}_{\mathrm{UNALCE}}\right),\tag{26}$$

with $\widehat{\boldsymbol{\Sigma}}_{\text{UNALCE}} = \widehat{\mathbf{B}}_{\text{UNALCE}}\widehat{\mathbf{B}}_{\text{UNALCE}}^\top + \widehat{\mathbf{S}}_{\text{UNALCE}} = \widehat{\mathbf{L}}_{\text{UNALCE}} + \widehat{\mathbf{S}}_{\text{UNALCE}}$.

Then, we note that

$$\widehat{\mathbf{B}\widehat{\mathbf{f}}}_{k,T} - \mathbf{B}\mathbf{H}_r\mathbf{f}_{k,T} = \widehat{\mathbf{B}\widehat{\mathbf{B}}}^\top \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}_k - \widehat{\mathbf{B}}\mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}\mathbf{x}_k = \widehat{\mathbf{L}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}_k - \widehat{\mathbf{L}}\boldsymbol{\Sigma}^{*-1}\mathbf{x}_k.$$

Therefore, under the assumptions of Theorem 4 we can write

$$\|\widehat{\mathbf{B}\widehat{\mathbf{f}}}_{k,T} - \mathbf{B}\mathbf{f}_{k,T}\| \le \|\widehat{\mathbf{B}\widehat{\mathbf{B}}}^\top \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}_k - \widehat{\mathbf{B}}\mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}\| \cdot \|\mathbf{x}_k\|.$$

By some algebra, we get

$$\begin{aligned}\widehat{\mathbf{B}\widehat{\mathbf{f}}}_{k,T} - \mathbf{B}\mathbf{f} &= \widehat{\mathbf{B}\widehat{\mathbf{B}}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} - \mathbf{B}\mathbf{B}^\top \boldsymbol{\Sigma}^{*-1} = \widehat{\mathbf{B}\widehat{\mathbf{B}}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} - \mathbf{B}\mathbf{B}^\top \boldsymbol{\Sigma}^{*-1} + \widehat{\mathbf{B}}\mathbf{B}^\top \boldsymbol{\Sigma}^{*-1} - \widehat{\mathbf{B}}\mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}\\
&= \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} - \mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}) + (\widehat{\mathbf{B}} - \mathbf{B})\mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}\\
&= \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} - \mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}) + (\widehat{\mathbf{B}} - \mathbf{B})\mathbf{B}^\top \boldsymbol{\Sigma}^{*-1} + \mathbf{B}(\widehat{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} - \mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}) - \mathbf{B}(\widehat{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} - \mathbf{B}^\top \boldsymbol{\Sigma}^{*-1})\\
&= (\widehat{\mathbf{B}} - \mathbf{B})(\widehat{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} + \mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}) + \mathbf{B}(\widehat{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} + \mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}) + (\widehat{\mathbf{B}} - \mathbf{B})\mathbf{B}^\top \boldsymbol{\Sigma}^{*-1}.\end{aligned}$$

At this point, by the triangular inequality, Corollary 2 and (26), it follows that

$$\arg\min_{\widehat{\mathbf{B}},\widehat{\mathbf{L}}=\widehat{\mathbf{B}\widehat{\mathbf{B}}}^\top \in \widehat{\mathcal{L}}(\widehat{r}_A),\widehat{\mathbf{S}}\in\widehat{\mathcal{S}}_{diag}} \max (\widehat{\mathbf{B}\widehat{\mathbf{B}}}^\top(\widehat{\mathbf{L}} + \widehat{\mathbf{S}})^{-1} - \mathbf{B}\mathbf{B}^\top(\mathbf{L}^* + \mathbf{S}^*)^{-1}) = \left(\widehat{\mathbf{B}}_{\text{UNALCE}}, \widehat{\boldsymbol{\Sigma}}_{\text{UNALCE}}\right),$$

which completes the proof. □

Theorem 6 states the same optimality properties of Theorem 5 for Thomson's estimates of factor scores and communalities estimated by UNALCE approach. Their proof follows from the results of Theorem 4.

In the end, let us define for each $k \in \{1, \dots, n\}$ the factor scores estimators:

- $\widehat{\mathbf{f}}_{k,B,A} = (\widehat{\mathbf{B}}_{\text{ALCE}}^\top\widehat{\mathbf{S}}_{\text{ALCE}}^{-1}\widehat{\mathbf{B}}_{\text{ALCE}})^{-1}\widehat{\mathbf{B}}_{\text{ALCE}}^\top\widehat{\mathbf{S}}_{\text{ALCE}}^{-1}\mathbf{x}_k$; $\widehat{\mathbf{f}}_{k,T,A} = \widehat{\mathbf{B}}_{\text{ALCE}}^\top(\widehat{\mathbf{B}}_{\text{ALCE}}\widehat{\mathbf{B}}_{\text{ALCE}}^\top + \widehat{\mathbf{S}}_{\text{ALCE}})^{-1}\mathbf{x}_k$;
- $\widehat{\mathbf{f}}_{k,B,U} = (\widehat{\mathbf{B}}_{\text{UNALCE}}^\top\widehat{\mathbf{S}}_{\text{UNALCE}}^{-1}\widehat{\mathbf{B}}_{\text{UNALCE}})^{-1}\widehat{\mathbf{B}}_{\text{UNALCE}}^\top\widehat{\mathbf{S}}_{\text{UNALCE}}^{-1}\mathbf{x}_k$; $\widehat{\mathbf{f}}_{k,T,U} = \widehat{\mathbf{B}}_{\text{UNALCE}}^\top(\widehat{\mathbf{B}}_{\text{UNALCE}}\widehat{\mathbf{B}}_{\text{UNALCE}}^\top + \widehat{\mathbf{S}}_{\text{UNALCE}})^{-1}\mathbf{x}_k$;

and the corresponding $n \times r$ matrix estimators: $\widehat{\mathbf{F}}_{B,A}^\top = [\widehat{\mathbf{f}}_{1,B,A} \ \dots \ \widehat{\mathbf{f}}_{n,B,A}]$; $\widehat{\mathbf{F}}_{T,A}^\top = [\widehat{\mathbf{f}}_{1,T,U} \ \dots \ \widehat{\mathbf{f}}_{n,T,A}]$; $\widehat{\mathbf{F}}_{B,U}^\top = [\widehat{\mathbf{f}}_{1,B,U} \ \dots \ \widehat{\mathbf{f}}_{n,B,U}]$; $\widehat{\mathbf{F}}_{T,U}^\top = [\widehat{\mathbf{f}}_{1,T,U} \ \dots \ \widehat{\mathbf{f}}_{n,T,U}]$. We finally state a corollary which completes our theory, in that it shows that, as $p, n \to \infty$, Bartlett's and Thomson's UNALCE estimates of factor scores converge to the respective ALCE ones and, more importantly, Bartlett's and Thomson's ALCE factor scores estimates converge to their OLS counterparts. This means that the asymptotic rates of OLS factor scores estimates reported in Theorem 1 also hold as $p, n \to \infty$ for Bartlett's and Thomson's UNALCE and ALCE factor scores estimates.

**Corollary 3.** *Under all the assumptions and conditions of Theorems 3 and 4, as $p, n \to \infty$, it holds:* (i) *(a)* $\widehat{\mathbf{F}}_{B,U} \to \widehat{\mathbf{F}}_{B,A}$ *and (b)* $\widehat{\mathbf{F}}_{B,A} \to \widehat{\mathbf{F}}_{\text{OLS2}}$*;* (ii) *(a)* $\widehat{\mathbf{F}}_{T,U} \to \widehat{\mathbf{F}}_{T,A}$ *and (b)* $\widehat{\mathbf{F}}_{T,A} \to \widehat{\mathbf{F}}_{\text{OLS2}}$.

**Proof.** Let us define the spectral decomposition of $\widehat{\mathbf{L}}_{\text{ALCE}}$ as $\widehat{\mathbf{U}}_{\text{ALCE}}\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}\widehat{\mathbf{U}}_{\text{ALCE}}^\top$, and $\widehat{\mathbf{B}}_{\text{ALCE}} = \widehat{\mathbf{U}}_{\text{ALCE}}\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}^{\frac{1}{2}}$. We recall that $\widehat{\mathbf{F}}_{\text{OLS2}} = \mathbf{X}\widehat{\mathbf{B}}_{\text{OLS2}}\boldsymbol{\Lambda}_r^{-1}$, with $\widehat{\mathbf{B}}_{\text{OLS2}} = \mathbf{U}_r\boldsymbol{\Lambda}_r^{\frac{1}{2}}$, and $\widehat{\mathbf{F}}_{\text{ALCE2}} = \mathbf{X}\widehat{\mathbf{B}}_{\text{ALCE2}}\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}^{-1}$, with $\widehat{\mathbf{B}}_{\text{ALCE2}} = \widehat{\mathbf{B}}_{\text{ALCE}}$. First, we observe that, under all the assumptions and conditions of Theorem 3, problems (3) and (6) are equivalent, because $\psi/p \to 0$. As a consequence, $\widehat{\mathbf{B}}_{\text{ALCE2}} \to \widehat{\mathbf{B}}_{\text{OLS2}}$ and $\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}} \to \boldsymbol{\Lambda}_r$ as $p, n \to \infty$.

For part i(a), $\widehat{\mathbf{F}}_{B,U} \to \widehat{\mathbf{F}}_{B,A}$ is implied by the condition $\psi/p \to 0$ as $p, n \to \infty$. We thus examine $\widehat{\mathbf{F}}_{B,A} = \mathbf{X}\widehat{\mathbf{S}}_A^{-1}\widehat{\mathbf{B}}_A(\widehat{\mathbf{B}}_A^\top\widehat{\mathbf{S}}_A^{-1}\widehat{\mathbf{B}}_A)^{-1}$. We note that, under Assumption 2, the proportion of non-zeros in $\mathbf{S}^*$, which is $O\left(p^{1+\delta_1}/p^2\right)$, tends to 0 as $p \to \infty$. This means that $\mathbf{S}^* \to \mathbf{D}$ as $p \to \infty$, where $\mathbf{D}$ is a $p \times p$ diagonal matrix, because $\lambda_p(\mathbf{S}^*) > 0$ by Assumption 3. More, $\widehat{\mathbf{S}}_A \to \mathbf{S}^*$ as $p, n \to \infty$ if Theorem 3 holds. Consequently, since $\widehat{\mathbf{B}}_{\text{ALCE}} = \widehat{\mathbf{B}}_{\text{ALCE2}}$, as $p, n \to \infty$ $\widehat{\mathbf{F}}_{B,A} = \mathbf{X}\widehat{\mathbf{S}}_A^{-1}\widehat{\mathbf{B}}_A(\widehat{\mathbf{B}}_A^\top\widehat{\mathbf{S}}_A^{-1}\widehat{\mathbf{B}}_A)^{-1} \to \mathbf{X}\widehat{\mathbf{B}}_A(\widehat{\mathbf{B}}_A^\top\widehat{\mathbf{B}}_A)^{-1} = \mathbf{X}\widehat{\mathbf{B}}_A\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}^{-1} = \widehat{\mathbf{F}}_{\text{ALCE2}}$. Part i(b) then follows by Theorem 3.

For part ii(a), $\widehat{\mathbf{F}}_{T,U} \to \widehat{\mathbf{F}}_{T,A}$ is implied by the condition $\psi/p \to 0$ as $p, n \to \infty$. We thus examine $\widehat{\mathbf{F}}_{T,A} = \mathbf{X}\widehat{\boldsymbol{\Sigma}}_A^{-1}\widehat{\mathbf{B}}_A$, where $\widehat{\boldsymbol{\Sigma}}_A = \widehat{\mathbf{L}}_A + \widehat{\mathbf{S}}_A$. We recall that, under Assumptions 2 and 3, $\mathbf{S}^* \to \mathbf{D}$, where $\mathbf{D}$ is a $p \times p$ diagonal matrix, and $\widehat{\mathbf{S}}_A \to \mathbf{S}^*$ by Theorem 3 as $p, n \to \infty$. At this point, $\widehat{\boldsymbol{\Sigma}}_A = \widehat{\mathbf{B}}_A\widehat{\mathbf{B}}_A^\top + \widehat{\mathbf{S}}_A \to \widehat{\mathbf{B}}_A\widehat{\mathbf{B}}_A^\top + \mathbf{D} = \widehat{\mathbf{U}}_{\text{ALCE}}\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}\widehat{\mathbf{U}}_{\text{ALCE}}^\top + \mathbf{D}$ as $p, n \to \infty$. Under the conditions of Theorem 3, we know that $\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}$ tends to $\boldsymbol{\Lambda}_r$, and $\widehat{\mathbf{U}}_{\text{ALCE}}$ tends to $\mathbf{U}_r$. Then, since the smallest eigenvalue of $\boldsymbol{\Lambda}_r$ is $O(p^{\alpha_r})$ by Lemma 1, and the eigenvalues of $\mathbf{D}$ are $o(p^{\alpha_r})$ by Assumption 2(a), it follows that, as $p, n \to \infty$, the $p \times p$ positive definite matrix $\widehat{\mathbf{U}}_{\text{ALCE}}\widehat{\boldsymbol{\Lambda}}_{\text{ALCE}}\widehat{\mathbf{U}}_{\text{ALCE}}^\top + \mathbf{D}$ tends to the $p \times p$ $r$-rank matrix $\mathbf{U}_r\boldsymbol{\Lambda}_r\mathbf{U}_r^\top$. Therefore, $\widehat{\boldsymbol{\Sigma}}_A^{-1}$ tends to $\mathbf{U}_r\boldsymbol{\Lambda}_r^{-1}\mathbf{U}_r^\top$, and $\widehat{\mathbf{F}}_{T,A}$ tends to $\mathbf{X}\mathbf{U}_r\boldsymbol{\Lambda}_r^{-1}\mathbf{U}_r^\top\mathbf{U}_r\boldsymbol{\Lambda}_r^{\frac{1}{2}} = \mathbf{X}\mathbf{U}_r\boldsymbol{\Lambda}_r^{-\frac{1}{2}} = \widehat{\mathbf{F}}_{\text{OLS2}}$. Part ii(b) then follows. □

Summing up, we have proved that, under the conditions of Theorems 3 and 4, Bartlett's and Thomson's UNALCE and ALCE estimators of factor scores $\widehat{\mathbf{F}}_{B,A}$, $\widehat{\mathbf{F}}_{B,U}$, $\widehat{\mathbf{F}}_{T,A}$, and $\widehat{\mathbf{F}}_{T,U}$ converge to the OLS factor scores estimator $\widehat{\mathbf{F}}_{\text{OLS2}}$. This implies that $\widehat{\mathbf{F}}_{B,U}$ and $\widehat{\mathbf{F}}_{T,U}$ are optimal estimators of factor scores in the finite sample, and that those estimators tend to $\widehat{\mathbf{F}}_{\text{OLS2}}$ as $p, n \to \infty$, exactly as $\widehat{\mathbf{F}}_{B,A}$ and $\widehat{\mathbf{F}}_{T,A}$, under the conditions of Theorem 3. Therefore, the UNALCE approach provides optimality guarantees in terms of Euclidean error norm when $p$ and $n$ are fixed, while retaining the algebraic consistency properties of ALCE and the asymptotic rates of OLS as $p, n \to \infty$. Importantly, unlike OLS, the UNALCE/ALCE approach is able to recover the latent rank and the residual sparsity pattern with probability tending to one. UNALCE factor model estimates are thus parametrically and algebraically consistent, as well as minimally biased in the finite sample.

## 6. Conclusions

In this paper, we propose to estimate high-dimensional approximate factor models with element-wise sparse residual covariance matrix by nuclear norm plus $\ell_1$ norm penalization. We provide the conditions on the respective magnitude of the dimension $p$ and the sample size $n$, as well as on the allowed degree of spikiness for latent eigenvalues and of sparsity for residual covariance, ensuring consistency for the estimators of factor loadings, scores and common components such derived. These conditions guarantee sparsistency, i.e., the residual sparsity pattern recovery, and the latent rank recovery, even when the latent factors are not strictly pervasive with respect to the dimension $p$. We derive a finite sample version of those factor model estimators, presenting strong optimality properties in terms of minimax Euclidean error bound for factor loadings and scores (estimated both via Bartlett's and Thomson's method). We finally prove that those finite sample estimators converge to the respective OLS counterparts as $p, n \to \infty$.

## Acknowledgments

The authors gratefully thank Editor, Associate Editor and two anonymous referees who relevantly improved the paper by their constructive comments, and acknowledge the support of the Air Force Office of Scientific Research, USA under award number FA9550-17-1-0103.

## Appendix A. Proofs

**Lemma 1.** *Let $\lambda_r(\boldsymbol{\Sigma}_n)$ be the $r$-th largest eigenvalue of the sample covariance matrix $\boldsymbol{\Sigma}_n = n^{-1}\sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k^\top$. Under Assumptions 1, 2 and 3, $\lambda_r(\boldsymbol{\Sigma}_n) \simeq p^{\alpha_r}$ with probability approaching 1 as $n \to \infty$.*

**Proof.** On one hand, we note that, since $r + p - p = r \leq p$, dual Weyl inequality (see [36]) can be applied, leading to

$$\lambda_r(\boldsymbol{\Sigma}^*) \geq \lambda_r(\mathbf{L}^*) + \lambda_p(\mathbf{S}^*). \tag{A.1}$$

From (A.1), we can write

$$\lambda_r(\boldsymbol{\Sigma}^*) \succeq O(p^{\alpha_r}) + O(p^{\delta_1}) = O(p^{\alpha_r}),$$

because $\lambda_r(\mathbf{L}^*) \simeq p^{\alpha_r}$ by Assumption 1(a), and $\lambda_p(\mathbf{S}^*) = O(p^{\delta_1})$ by Assumption 2, with $\delta_1 < \alpha_r$.

On the other hand, Lidskii inequality (see [36]) leads to

$$\lambda_r(\boldsymbol{\Sigma}^*) \leq \lambda_r(\mathbf{L}^*) + \sum_{j=1}^{r} \lambda_j(\mathbf{S}^*). \tag{A.2}$$

From (A.2), we can write

$$\lambda_r(\boldsymbol{\Sigma}^*) \preceq O(p^{\alpha_r}) + O(rp^{\delta_1}) = O(p^{\alpha_r}),$$

because $\lambda_r(\mathbf{L}^*) \simeq p^{\alpha_r}$ by Assumption 1(a), $\lambda_1(\mathbf{S}^*) = O(p^{\delta_1})$ with $\delta_1 < \alpha_r$ by Assumption 2(a), and $r$ is finite for all $p \in \mathbb{N}$ by Assumption 1(b). It follows that $\lambda_r(\boldsymbol{\Sigma}^*) \simeq p^{\alpha_r}$.

Recalling that $\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k^\top$ and $\mathbf{x}_k = \mathbf{B}\mathbf{f}_k + \boldsymbol{\epsilon}_k$, where $\mathbf{f}_k$ and $\boldsymbol{\epsilon}_k$, $k \in \{1, \ldots, n\}$, are respectively the vectors of factor scores and residuals for each observation, we can decompose the error matrix $\mathbf{E}_n = \boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*$ in four components as follows (see [22]):

$$\mathbf{E}_n = \boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^* = \mathbf{D}_1 + \mathbf{D}_2 + \mathbf{D}_3 + \mathbf{D}_4,$$

where $\mathbf{D}_1 = n^{-1}\mathbf{B}\left(\sum_{k=1}^n \mathbf{f}_k\mathbf{f}_k^\top - \mathbf{I}_r\right)\mathbf{B}^\top$, $\mathbf{D}_2 = n^{-1}\sum_{k=1}^n \left(\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k^\top - \mathbf{S}^*\right)$, $\mathbf{D}_3 = n^{-1}\mathbf{B}\sum_{k=1}^n \mathbf{f}_k\boldsymbol{\epsilon}_k^\top$, $\mathbf{D}_4 = \mathbf{D}_3^\top$.

Following [22], we note that

$$\|\mathbf{D}_1\|_2 \leq \left\| \frac{1}{n}\left(\sum_{k=1}^n \mathbf{f}_k\mathbf{f}_k^\top - \mathbf{I}_r\right)\right\|_2 \|\mathbf{B}\mathbf{B}^\top\|_2 \leq rp^{\alpha_1}\max_{i,j\leq r}\left|\frac{1}{n}\sum_{k=1}^n f_{i,k}f_{j,k} - \mathrm{E}[f_{i,k}f_{j,k}]\right|,$$

since $\mathrm{E}[\mathbf{f}] = \mathbf{0}_r$ and $\mathrm{Var}[\mathbf{f}] = \mathbf{I}_r$ by Assumption 3, $\|\mathbf{B}\mathbf{B}^\top\|_2 = O(p^{\alpha_1})$ by Assumption 1(a), and

$$\left\| \frac{1}{n}\left(\sum_{k=1}^n \mathbf{f}_k\mathbf{f}_k^\top - \mathbf{I}_r\right)\right\|_2 \leq r\left\| \frac{1}{n}\left(\sum_{k=1}^n \mathbf{f}_k\mathbf{f}_k^\top - \mathbf{I}_r\right)\right\|_\infty,$$

with $r$ finite and independent of $p$ by Assumption 1(b).

Under Assumption 3, we can apply Lemma 4(a) in [22], which claims

$$\max_{i,j\leq r}\left|\frac{1}{n}\sum_{k=1}^{n} f_{i,k}f_{j,k} - \mathrm{E}[f_{i,k}f_{j,k}]\right| \leq \tilde{C}\frac{1}{\sqrt{n}}, \tag{A.3}$$

with probability $1-O(1/n^2)$ ($\tilde{C}$ is a real positive constant). Consequently, we obtain

$$\|\mathbf{D}_1\|_2 \leq \tilde{C}r\frac{p^{\alpha_1}}{\sqrt{n}} \leq C'\frac{p^{\alpha_1}}{\sqrt{n}}, \tag{A.4}$$

with $C' = \tilde{C}r= O(1)$ by Assumption 1(b).

Then, we note that the diagonal elements of the matrix $\mathbf{S}^*$ are bounded by a finite constant, due to Assumption 2(b). Under Assumption 3 and the condition $\ln(p)/n \to 0$, (12) in [12] thus holds for the matrix $\mathbf{S}^*$, leading to:

$$\|\mathbf{D}_2\|_\infty = \max_{i,j\leq p}\left|\frac{1}{n}\sum_{k=1}^{n}\epsilon_{i,k}\epsilon_{j,k} - \mathrm{E}(\epsilon_{i,k}\epsilon_{j,k})\right| \leq \tilde{C}_2\sqrt{\frac{\ln(p)}{n}}, \tag{A.5}$$

that holds with probability $1-O(1/n^2)$.

Now, by the triangular inequality we can write

$$\|\mathbf{D}_2\|_2 \leq \|\mathbf{D}_2^{(1)}\|_2 + \|\mathbf{D}_2^{(2)}\|_2, \tag{A.6}$$

where $\mathbf{D}_2^{(1)} = \mathcal{T}^{(H)}_{\|\mathbf{S}^*\|_{\mathrm{min,off}}}(\mathbf{D}_2)$, with $\mathcal{T}^{(H)}$ hard-thresholding operator of parameter $\|\mathbf{S}^*\|_{\mathrm{min,off}}$, and $\mathbf{D}_2^{(2)} = \mathbf{D}_2 - \mathbf{D}_2^{(1)}$. Since by Assumption 2(a) $\|\mathbf{S}^*\|_{0,v} = O(p^{\delta_1})$, it follows from (A.5) that, as $n\to\infty$,

$$\|\mathbf{D}_2^{(1)}\|_2 \leq \|\mathbf{D}_2^{(1)}\|_{0,v}\|\mathbf{D}_2^{(1)}\|_\infty \leq \tilde{C}_2\delta_2 p^{\delta_1}\sqrt{\frac{\ln(p)}{n}}. \tag{A.7}$$

Similarly, it follows from (A.5) that, as $n\to\infty$,

$$\|\mathbf{D}_2^{(2)}\|_2 \leq \|\mathbf{D}_2{}^{(2)}\|_{0,v}\|\mathbf{D}_2{}^{(2)}\|_\infty < p\|\mathbf{S}^*\|_{\mathrm{min,off}}. \tag{A.8}$$

Imposing the condition $p\|\mathbf{S}^*\|_{\mathrm{min,off}} = o(p^{\delta_1}\sqrt{\ln(p)/n})$, we get $p^{1-\delta_1}\|\mathbf{S}^*\|_{\mathrm{min,off}} = o(\sqrt{\ln(p)/n})$, which leads to $p^{1-\delta_1}\|\mathbf{S}^*\|_{\mathrm{min,off}} = o(1)$ as $\ln(p)/n \to 0$. Therefore, combining (A.7) and (A.8), by Assumption 2(c) it follows from (A.6) that, as $n\to\infty$,

$$\|\mathbf{D}_2\|_2 \leq \tilde{C}_2\delta_2 p^{\delta_1}\sqrt{\frac{\ln(p)}{n}}. \tag{A.9}$$

At this stage, we consider $\left\|n^{-1}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k^\top\right\|_2$. We first observe that $\left\|n^{-1}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k^\top\right\|_2 \leq \left\|n^{-1}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k^\top\right\|_F$. We then write $\left\|n^{-1}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k^\top\right\|_F = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{p}\widehat{\mathrm{Cov}}(f_i,\epsilon_j)^2} \leq \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{p}\widehat{\mathrm{V}}(f_i)\widehat{\mathrm{V}}(\epsilon_j)}$. We note that $\sqrt{\sum_{i=1}^{r}\sum_{j=1}^{p}\widehat{\mathrm{V}}(f_i)\widehat{\mathrm{V}}(\epsilon_j)}$ converges to $\sqrt{\sum_{i=1}^{r}\sum_{j=1}^{p}\mathrm{V}(f_i)\mathrm{V}(\epsilon_j)}$ for each $i\in\{1,\ldots,r\}$ and $j\in\{1,\ldots,p\}$ as $n\to\infty$ and $\ln(p)/n\to 0$ with probability $1-O(1/n^2)$, by (A.3) and (A.5). Therefore, we can write $\left\|n^{-1}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k^\top\right\|_F \leq \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{p}\mathrm{V}(f_i)\mathrm{V}(\epsilon_j)} \leq \sqrt{r\sum_{j=1}^{p}\mathrm{V}(\epsilon_j)} = \sqrt{ro(p^{\alpha_1})} = o(p^{\alpha_1/2})$, by Cauchy–Schwartz inequality and Assumption 1(b), 2(d) and 3. It follows that

$$\left\|n^{-1}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k^\top\right\|_2 = o(p^{\alpha_1/2})$$

as $n\to\infty$ and $\ln(p)/n\to 0$.

Consequently, we obtain with probability $1-O(1/n^2)$ the following claim

$$\|\mathbf{D}_3\|_2 \leq \left\|\frac{1}{n}\sum_{k=1}^{n}\mathbf{f}_k\boldsymbol{\epsilon}_k^\top\right\|_2 \cdot \|\mathbf{B}\| = o(p^{\frac{\alpha_1}{2}})O\left(p^{\frac{\alpha_1}{2}}\right) = o(p^{\alpha_1}) \tag{A.10}$$

because $\|\mathbf{B}\| = O(p^{\alpha_1/2})$ by Assumption 1(a).

Putting (A.4), (A.9), and (A.10) together, the following bound is proved with probability $1-O(1/n^2)$:

$$\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_2 \leq C'\frac{p^{\alpha_1}}{\sqrt{n}},$$

because $\delta_1 < \alpha_r \leq \alpha_1$ by Assumptions 1(a) and 2(a). It follows that

$$\frac{1}{p^{\alpha_1}}\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_2 \xrightarrow{n\to\infty} 0,$$

which proves the lemma. □

**Proof of Theorem 1.**

According to [4], setting $\widehat{\mathbf{F}} = \widehat{\mathbf{F}}_{\text{OLS1}}$ and $\mathbf{H} = \mathbf{H}_{\text{OLS1}} = n^{-1}\boldsymbol{\Lambda}_r^{-1}\widehat{\mathbf{F}}^\top\mathbf{F}\mathbf{B}^\top\mathbf{B}$, we can write, for each $\widehat{\mathbf{f}}_{k_1}$, $k_1 \in \{1, \dots, n\}$:

$$\widehat{\mathbf{f}}_{k_1} - \mathbf{H}\mathbf{f}_{k_1} = \frac{1}{n}\left(\frac{\boldsymbol{\Lambda}_r}{p}\right)^{-1}\left\{\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathrm{E}[\boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}]/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\varsigma_{k_1k_2} + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\eta_{k_1k_2} + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\xi_{k_1k_2}\right\}, \tag{A.11}$$

where $\xi_{k_1k_2} = (\mathbf{f}_{k_1}^\top\sum_{j=1}^{p}\mathbf{b}_j\epsilon_{k_2,j})/p$, $\eta_{k_1k_2} = (\mathbf{f}_{k_2}^\top\sum_{j=1}^{p}\mathbf{b}_j\epsilon_{k_1,j})/p$, $\varsigma_{k_1k_2} = \boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}/p - \mathrm{E}[\boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}]/p$, and $\boldsymbol{\Lambda}_r$ is the diagonal matrix containing the top $r$ eigenvalues of $\boldsymbol{\Sigma}_n$ in decreasing order. Expression (A.11) holds conditioning on the assumption that the latent rank $r$ is known.

In order to verify (A.11), we observe that the r.h.s of (A.11) (divided by $n^{-1}(\boldsymbol{\Lambda}_r/p)^{-1}$) can be written as

$$\begin{aligned}
&\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathrm{E}[\boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}]/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\varsigma_{k_1k_2} + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\eta_{k_1k_2} + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\xi_{k_1k_2}\\
&= \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathrm{E}[\boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}]/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}/p - \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathrm{E}[\boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}]/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}(\mathbf{f}_{k_2}^\top\sum_{j=1}^{p}\mathbf{b}_j\epsilon_{k_1,j})/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}(\mathbf{f}_{k_1}^\top\sum_{j=1}^{p}\mathbf{b}_j\epsilon_{k_2,j})/p\\
&= \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}(\mathbf{f}_{k_2}^\top\sum_{j=1}^{p}\mathbf{b}_j\epsilon_{k_1,j})/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}(\mathbf{f}_{k_1}^\top\sum_{j=1}^{p}\mathbf{b}_j\epsilon_{k_2,j})/p\\
&= \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}(\mathbf{f}_{k_2}^\top\mathbf{B}^\top\epsilon_{k_1})/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}(\mathbf{f}_{k_1}^\top\mathbf{B}^\top\epsilon_{k_2})/p.
\end{aligned}$$

We know that $\epsilon_{k_1} = \mathbf{x}_{k_1} - \mathbf{B}\mathbf{f}_{k_1}$ and $\epsilon_{k_2} = \mathbf{x}_{k_2} - \mathbf{B}\mathbf{f}_{k_2}$. Then, we can write

$$\begin{aligned}
&\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\boldsymbol{\epsilon}_{k_2}^\top\boldsymbol{\epsilon}_{k_1}/p = \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}(\mathbf{x}_{k_2} - \mathbf{B}\mathbf{f}_{k_2})^\top(\mathbf{x}_{k_1} - \mathbf{B}\mathbf{f}_{k_1})/p\\
&= \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{x}_{k_2}^\top\mathbf{x}_{k_1}/p - \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{x}_{k_2}^\top\mathbf{B}\mathbf{f}_{k_1}/p - \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_2}^\top\mathbf{B}^\top\mathbf{x}_{k_1}/p + \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_2}^\top\mathbf{B}^\top\mathbf{B}\mathbf{f}_{k_1}/p.
\end{aligned}$$

We can also write

$$\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}(\mathbf{f}_{k_2}^\top\mathbf{B}^\top\epsilon_{k_1})/p = \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_2}^\top\mathbf{B}^\top(\mathbf{x}_{k_1} - \mathbf{B}\mathbf{f}_{k_1})/p = \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_2}^\top\mathbf{B}^\top\mathbf{x}_{k_1}/p - \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_2}^\top\mathbf{B}^\top\mathbf{B}\mathbf{f}_{k_1}/p$$

and

$$\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}(\mathbf{f}_{k_1}^\top\mathbf{B}^\top\epsilon_{k_2})/p = \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_1}^\top\mathbf{B}^\top(\mathbf{x}_{k_2} - \mathbf{B}\mathbf{f}_{k_2})/p = \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_1}^\top\mathbf{B}^\top\mathbf{x}_{k_2}/p - \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_1}^\top\mathbf{B}^\top\mathbf{B}\mathbf{f}_{k_2}/p.$$

Therefore, due to cancellation effects, the r.h.s. of (A.11) (divided by $n^{-1}(\boldsymbol{\Lambda}_r/p)^{-1}$) reduces to $\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{x}_{k_2}^\top\mathbf{x}_{k_1}/p - \sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_1}^\top\mathbf{B}^\top\mathbf{B}\mathbf{f}_{k_2}/p$. We finally observe that $\mathbf{f}_{k_1}^\top\mathbf{B}^\top\mathbf{B}\mathbf{f}_{k_2} = \mathbf{f}_{k_2}^\top\mathbf{B}^\top\mathbf{B}\mathbf{f}_{k_1}$, because it is a scalar, and that $\widehat{\mathbf{f}}_{k_2} = \boldsymbol{\Lambda}_r^{-1}\mathbf{B}^\top\mathbf{x}_{k_2}$. It follows that, under the conditions of Lemma 1, as $n \to \infty$,

$$\sum_{k_2=1}^{n}\boldsymbol{\Lambda}_r^{-1}\mathbf{B}^\top\mathbf{x}_{k_2}\mathbf{x}_{k_2}^\top\mathbf{x}_{k_1}/p = \boldsymbol{\Lambda}_r^{-1}\mathbf{B}^\top\sum_{k_2=1}^{n}\mathbf{x}_{k_2}\mathbf{x}_{k_2}^\top\mathbf{x}_{k_1}/p = n\boldsymbol{\Lambda}_r^{-1}\mathbf{B}^\top\mathbf{B}\mathbf{B}^\top\mathbf{x}_{k_1}/p = n\boldsymbol{\Lambda}_r\boldsymbol{\Lambda}_r^{-1}\mathbf{B}^\top\mathbf{x}_{k_1}/p = n\boldsymbol{\Lambda}_r\widehat{\mathbf{f}}_{k_1}/p,$$

and

$$\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathbf{f}_{k_2}^{\top}\mathbf{B}^{\top}\mathbf{B}\mathbf{f}_{k_1}/p=\widehat{\mathbf{F}}^{\top}\mathbf{F}\mathbf{B}^{\top}\mathbf{B}\mathbf{f}_{k_1}/p=n\boldsymbol{\Lambda}_r\mathbf{H}\mathbf{f}_{k_1}/p,$$

where $\mathbf{H}=\mathbf{H}_{\mathrm{OLS}1}=n^{-1}\boldsymbol{\Lambda}_r^{-1}\widehat{\mathbf{F}}^{\top}\mathbf{F}\mathbf{B}^{\top}\mathbf{B}$. Re-multiplying $n\boldsymbol{\Lambda}_r\widehat{\mathbf{f}}_{k_1}/p$ and $n\boldsymbol{\Lambda}_r\mathbf{H}\mathbf{f}_{k_1}/p$ by $n^{-1}(\boldsymbol{\Lambda}_r/p)^{-1}$, Eq. (A.11) is verified.

Recalling that $\mathbf{f}_{k_1}$ and $\boldsymbol{\epsilon}_{k_1}$ are uncorrelated processes over the observations $k_1=\{1,\ldots,n\}$ by Assumption 3, we can prove the following lemmas by simply applying the corresponding proofs in [22] and imposing Assumption 4.

**Lemma 2.** *For* $i\in\{1,\ldots,r\}$*:* (i) $n^{-1}\sum_{k_2=1}^{n}\left(\frac{1}{np}\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2,i}\mathrm{E}[\boldsymbol{\epsilon}_{k_2}^{\top}\boldsymbol{\epsilon}_{k_1}]\right)^2=O_p\left(n^{-1}\right)$*;* (ii) $n^{-1}\sum_{k_2=1}^{n}n^{-1}\left(\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2,i}\varsigma_{k_1k_2}\right)^2=O_p\left(p^{-1}\right)$*;* (iii) $n^{-1}\sum_{k_2=1}^{n}n^{-1}\left(\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2,i}\eta_{k_1k_2}\right)^2=O_p\left(p^{-1}\right)$*;* (iv) $n^{-1}\sum_{k_2=1}^{n}n^{-1}\left(\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2,i}\xi_{k_1k_2}\right)^2=O_p\left(p^{-1}\right)$.

**Proof.** The proof of part (i) is analogous to the proof of Lemma 8(a) in [22], where

$$\frac{1}{np}\max_{k_2,k_1}|\mathrm{E}[\boldsymbol{\epsilon}_{k_2}^{\top}\boldsymbol{\epsilon}_{k_1}]|=\max_{k_1\le n}\frac{1}{n}\sum_{k_2=1}^{n}\frac{1}{p}|\mathrm{E}[\boldsymbol{\epsilon}_{k_2}^{\top}\boldsymbol{\epsilon}_{k_1}]|\le\frac{1}{np}\|\mathbf{S}^*\|_1=O_p\left(\frac{1}{n}\right)$$

by the condition $\|\mathbf{S}^*\|_1/p\le\delta_2'$ for some $\delta_2'>0$, and because $\epsilon_{k_1}$ is uncorrelated across observations.

The proof of part (ii) is analogous to the proof of Lemma 8(b) in [22] under Assumption 4(a).

The proof of parts (iii) and (iv) are analogous to the proof of Lemma 8(c) and 8(d) in [22] under Assumption 4(b). □

**Lemma 3.** (i) $\max_{k_1\le n}\left\|(np)^{-1}\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\mathrm{E}[\boldsymbol{\epsilon}_{k_2}^{\top}\boldsymbol{\epsilon}_{k_1}]\right\|=O_p\left(n^{-1/2}\right)$*;* (ii) $\max_{k_1\le n}\left\|n^{-1}\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\varsigma_{k_1k_2}\right\|=O_p\left(n^{1/4}/p^{1/2}\right)$*;* (iii) $\max_{k_1\le n}\left\|n^{-1}\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\eta_{k_1k_2}\right\|=O_p\left(n^{1/4}/p^{1/2}\right)$*;* (iv) $\max_{k_1\le n}\left\|n^{-1}\sum_{k_2=1}^{n}\widehat{\mathbf{f}}_{k_2}\xi_{k_1k_2}\right\|=O_p\left(n^{1/4}/p^{1/2}\right)$.

**Proof.** The proof is analogous to the Proof of Lemma 2 with reference to the proof of Lemma 9 in [22]. □

Observing that the eigenvalues of $(\boldsymbol{\Lambda}_r/p)^{-1}$ scale to $O(p^{1-\alpha_r})$, due to Lemma 1, we obtain the following Lemma.

**Lemma 4.** (i)$\max_{i\le r}n^{-1}\sum_{k=1}^{n}(\widehat{\mathbf{f}}_{k,i}-\mathbf{H}\mathbf{f}_{k,i})^2=O_p\left(p^{2(1-\alpha_r)}/n+p^{2(1-\alpha_r)}/p\right)$*;*
(ii) $n^{-1}\sum_{k=1}^{n}\|\widehat{\mathbf{f}}_k-\mathbf{H}\mathbf{f}_k\|^2=O_p\left(p^{2(1-\alpha_r)}/n+p^{2(1-\alpha_r)}/p\right)$*;*
(iii) $\max_{k\le n}\sum_{k=1}^{n}\|\widehat{\mathbf{f}}_k-\mathbf{H}\mathbf{f}_k\|=O_p\left(p^{1-\alpha_r}/\sqrt{n}+p^{1-\alpha_r}n^{1/4}/p^{1/2}\right)$.

**Proof.** The proof is analogous to the proof of Lemma 10 in [22], and follows from the fact $\widehat{r}=r$, Lemma 1, Lemmas 2 and 3. Note that part (iii) derives a bound for the uniform rate of $\widehat{\mathbf{f}}_k-\mathbf{H}\mathbf{f}_k$ over $k\in\{1,\ldots,n\}$, that leads to part (ii) of Theorem 1, because $p^{2-2\alpha_r}/n=o(1)$. □

**Lemma 5.** (i) $\mathbf{H}\mathbf{H}^{\top}=\mathbf{I}_r+O_p\left(\frac{p^{1-\alpha_r}}{\sqrt{n}}+\frac{p^{1-\alpha_r}}{p^{1/2}}\right)$*;* (ii) $\mathbf{H}^{\top}\mathbf{H}=\mathbf{I}_r+O_p\left(\frac{p^{1-\alpha_r}}{\sqrt{n}}+\frac{p^{1-\alpha_r}}{p^{1/2}}\right)$.

**Proof.** The proof is analogous to the proof of Lemma 11 in [22], and follows from the fact $\widehat{r}=r$, Lemma 1, inequality (A.3), and Lemma 4. □

At this stage, following [22], we can observe that $\widehat{\mathbf{b}}_j-\mathbf{H}\mathbf{b}_j$ can be decomposed as follows:

$$\widehat{\mathbf{b}}_j-\mathbf{H}\mathbf{b}_j=\mathbf{A}_{1,j}+\mathbf{A}_{2,j}+\mathbf{A}_{3,j},$$

with

$$\mathbf{A}_{1,j}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{H}\mathbf{f}_k\epsilon_{k,j};\quad \mathbf{A}_{2,j}=\sum_{k=1}^{n}x_{k,j}(\widehat{\mathbf{f}}_k-\mathbf{H}\mathbf{f}_k);\quad \mathbf{A}_{3,j}=\mathbf{H}\sum_{k=1}^{n}(\widehat{\mathbf{f}}_k\widehat{\mathbf{f}}_k^{\top}-\mathbf{I}_r)\mathbf{b}_j.$$

We note from Lemma 5 that $\|\mathbf{H}\|=O_p(1)$ if $n\succ p^{2-2\alpha_r}$ and $\alpha_r>1-1/2=1/2$. The second condition is always verified, because $\alpha_r>1/2$ by Assumption 1(a). Under the first condition, inequality (A.10) allows to conclude that

$$\max_{j\le p}\|\mathbf{A}_{1,j}\|=O_p\left(p^{\frac{\delta_1}{2}}\sqrt{\frac{\ln(p)}{n}}\right).\tag{A.12}$$

From Lemma 4, it follows that

$$\max_{j\le p}\|\mathbf{A}_{2,j}\| = O_p\left(\frac{p^{1-\alpha_r}}{n^{\frac{1}{2}}} + \frac{p^{1-\alpha_r}}{p^{\frac{1}{2}}}\right), \tag{A.13}$$

because $\mathrm{E}[x_{k,j}] = O(1)$ for each $k \in \{1,\dots,n\}$ by Assumptions 1(b) and 2(b).

Inequality (A.3) and Assumption 1(b) then ensure that, since $\|\mathbf{H}\| = O_p(1)$ by Lemma 5,

$$\max_{j\le p}\|\mathbf{A}_{3,j}\| = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{A.14}$$

At this point, since $\alpha_1 - \alpha_r \le \delta_1$, the condition $\delta_1/2 \le 1-\alpha_r$ is always verified, because it must hold

$$\alpha_1 - \alpha_r \le \delta_1 \le 2 - 2\alpha_r,$$

that leads to $\alpha_1 \le 2-\alpha_r$ which is always true under Assumption 1(a).

Therefore, putting together (A.12), (A.13), (A.14), under the condition $p^{2-2\alpha_r}/n = o(1)$ we obtain that

$$\max_{j\le p}\|\widehat{\mathbf{b}}_j - \mathbf{H}\mathbf{b}_j\| = O_p(\omega_n), \tag{A.15}$$

where $\omega_n = \sqrt{\ln(p)/n}$ as $p \to \infty$, because $1-\alpha_r-\frac{1}{2} < 0$ is always verified under Assumption 1(a), due to the condition $\alpha_r > 1/2$. Part (i) of Theorem 1 then follows.

Applying parts (i) and (ii), and Lemma 5, we can prove by Assumption 3, analogously to [22], that for each $k \in \{1,\dots,n\}$:

$$\max_{j\le p, i\le r}\|\widehat{\mathbf{b}}_j^\top\widehat{\mathbf{f}}_k - \mathbf{b}_j^\top\mathbf{f}_k\| = O_p\left(\frac{n^{1/4}p^{1-\alpha_r}}{p^{1/2}} + \ln(n)^{\frac{1}{c_2}}\sqrt{\frac{\ln(p)}{n}}\right),$$

from which part (iii) of Theorem 1 follows. □

**Lemma 6.** *Under Assumptions 1(b), 2(b) and 3,*

$$\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_\infty \le C'\sqrt{\frac{\ln(p)}{n}}.$$

*with probability approaching one as $n \to \infty$.*

**Proof.** Under Assumptions 1(b) and 3, with probability $1 - O(1/n^2)$,

$$\|\mathbf{D}_1\|_\infty \le \left\|\frac{1}{n}\left(\sum_{k=1}^n \mathbf{f}_k\mathbf{f}_k^\top - \mathbf{I}_r\right)\right\|_\infty \|\mathbf{B}\mathbf{B}^\top\|_\infty \le C'\sqrt{\frac{1}{n}}, \tag{A.16}$$

because $\|\mathbf{B}\mathbf{B}^\top\|_\infty \le (\max_{j\in\{1,\dots,p\}}\|\mathbf{b}_j\|)^2 \le r^2\|\mathbf{B}\|_\infty^2 = O(1)$ for all $p \in \mathbb{N}$.

Under Assumptions 2(b) and 3, (A.5) ensures that, with probability $1 - O(1/n^2)$,

$$\|\mathbf{D}_2\|_\infty = \max_{i,j\le p}\left|\frac{1}{n}\sum_{k=1}^n \epsilon_{i,k}\epsilon_{j,k} - \mathrm{E}(\epsilon_{i,k}\epsilon_{j,k})\right| \le C'\sqrt{\frac{\ln(p)}{n}}. \tag{A.17}$$

Under Assumptions 1(b), 2(b) and 3, from (A.16) and (A.17) we get

$$\|\mathbf{D}_3\|_\infty = \left\|\frac{1}{n}\sum_{k=1}^n \mathbf{f}_k\boldsymbol{\epsilon}_k^\top\right\|_\infty \le C'\sqrt{\frac{\ln(p)}{n}}, \tag{A.18}$$

with probability $1 - O(1/n^2)$. Putting together (A.16), (A.17), (A.18), the statement follows. □

**Proof of Theorem 2.**

Let us define the following measure of transversality between two matrix varieties $\mathcal{T}_1$ and $\mathcal{T}_2$:

$$\varrho(\mathcal{T}_1,\mathcal{T}_2) = \max_{\|\mathbf{N}\|_2\le 1}\|\mathcal{P}_{\mathcal{T}_1}\mathbf{N} - \mathcal{P}_{\mathcal{T}_2}\mathbf{N}\|_2,$$

where $\mathcal{P}_{\mathcal{T}_1}$ and $\mathcal{P}_{\mathcal{T}_2}$ are the projection operators onto $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. Given two conformable matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, we define the standard Euclidean inner product $\langle\mathbf{M}_1,\mathbf{M}_2\rangle = \mathrm{tr}(\mathbf{M}_1'\mathbf{M}_2) = \mathrm{tr}(\mathbf{M}_2'\mathbf{M}_1)$, and we call $\mathcal{A}$ the addition operator, such that $\mathcal{A}(\mathbf{M}_1,\mathbf{M}_2) = \mathbf{M}_1 + \mathbf{M}_2$, and $\mathcal{A}^\dagger$ its adjoint operator *wrt* the inner product above defined (see also page 1946 in [18]).

Hereafter, let $\Omega = \Omega(\mathbf{S}^*)$ and $\mathcal{T} = \mathcal{T}(\mathbf{L}^*)$, where $\Omega$ is the space tangent to $\mathcal{S}(s)$ (see (9)) at $\mathbf{S}^*$ and $\mathcal{T}$ is the space tangent to $\mathcal{L}(r)$ (see (8)) at $\mathbf{L}^*$. We define the Cartesian product $\mathcal{Y} = \Omega \times \mathcal{T}'$, where $\mathcal{T}'$ is a manifold such that $\varrho(\mathcal{T},\mathcal{T}') \le \xi(\mathcal{T})/2$.

In light of these definitions, the following identities hold:

$$\mathcal{A}^\dagger\mathcal{A}(\mathbf{S},\mathbf{L})=(\mathbf{S}+\mathbf{L},\mathbf{S}+\mathbf{L});\quad \mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\mathbf{S},\mathbf{L})=(\mathbf{S}+\mathcal{P}_{\Omega}\mathbf{L},\mathcal{P}_{\mathcal{T}'}\mathbf{S}+\mathbf{L});\quad \mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\mathbf{S},\mathbf{L})=(\mathbf{S}+\mathcal{P}_{\Omega^\perp}\mathbf{L},\mathcal{P}_{\mathcal{T}'^\perp}\mathbf{S}+\mathbf{L}).$$

We consider the following norm $g_\gamma$:

$$g_\gamma(\widehat{\mathbf{L}}-\mathbf{L}^*,\widehat{\mathbf{S}}-\mathbf{S}^*)=\max\left(\frac{\|\widehat{\mathbf{S}}-\mathbf{S}^*\|_\infty}{\gamma},\frac{\|\widehat{\mathbf{L}}-\mathbf{L}^*\|_2}{\|\mathbf{L}^*\|_2}\right),\tag{A.19}$$

with $\psi_0=(1/\xi(\mathcal{T}(\mathbf{L}^*)))\sqrt{\ln(p)/n}$, $\psi=p^{\alpha_1}\psi_0$, $\rho_0=\rho$, $\gamma=\rho_0/\psi_0$, where $\psi$ and $\rho$ are the thresholds in (6). The norm (A.19) is the dual norm of the composite penalty $\psi_0\|\cdot\|_*+\rho_0\|\cdot\|_1$, with which the direct sum $\mathcal{L}(r)\oplus\mathcal{S}(s)$ is naturally equipped. Obviously this $g_\gamma$-consistency implies consistency in $\ell_2$ norm.

**Proposition 1.** *Suppose that $\gamma\in[9\xi(\mathcal{T}(\mathbf{L}^*)),1/(6\mu(\Omega(\mathbf{S}^*)))]$. Then, under Assumption 5, for all $(\mathbf{S},\mathbf{L})\in\mathcal{Y}$ such that $\mathcal{Y}=\Omega\times\mathcal{T}'$ with $\varrho(\mathcal{T},\mathcal{T}')\leq\xi(\mathcal{T})/2$, the following holds:*

(i) $g_\gamma(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\mathbf{S},\mathbf{L}))\geq\frac{1}{2}g_\gamma(\mathbf{S},\mathbf{L})$;
(ii) $g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\mathbf{S},\mathbf{L}))\leq\frac{1}{2}g_\gamma(\mathbf{S},\mathbf{L})$.

**Proof.** Since $\mathbf{L}\in\mathcal{T}'$, $\mathbf{S}\in\Omega$, $\gamma\in[9\xi(\mathcal{T}),1/6\mu(\Omega)]$, and Assumption 5 ensures that the condition $\xi(\mathcal{T})\mu(\Omega)\leq 1/54$ holds, the proof of Proposition 3.3 in [18] follows. □

We now consider the solution of the following algebraic problem:

$$(\widehat{\mathbf{S}}_\Omega,\widehat{\mathbf{L}}_{\mathcal{T}'})=\arg\min_{\underline{\mathbf{L}}\in\mathcal{T}',\underline{\mathbf{S}}\in\Omega}\frac{1}{2p^{\alpha_1}}\|\boldsymbol{\Sigma}_n-(\underline{\mathbf{L}}+\underline{\mathbf{S}})\|_F^2+\psi_0\|\underline{\mathbf{L}}\|_*+\rho_0\|\underline{\mathbf{S}}\|_1.\tag{A.20}$$

This is the constrained version of the minimization (6) which for convenience is rescaled by $p^{\alpha_1}$. The additional constraints are needed to ensure that the Hessian of $\frac{1}{2}\|\boldsymbol{\Sigma}_n-(\underline{\mathbf{L}}+\underline{\mathbf{S}})\|_F^2$ is positive definite, such that the optimum of (A.20) is unique.

**Proposition 2.** *Let $\varrho(\mathcal{T}',\mathcal{T})\leq\xi(\mathcal{T})/2$ and define $\widetilde{r}=\max\{4[g_\gamma(A^\dagger\boldsymbol{\Delta}_n)+g_\gamma(A^\dagger\mathbf{C}_{\mathcal{T}'})+\psi_0],\|\mathbf{C}_{\mathcal{T}'}\|_2\}$ where $\mathbf{C}_{\mathcal{T}'}=\mathcal{P}_{\mathcal{T}'^\perp}(\mathbf{L}^*)$ and $\boldsymbol{\Delta}_n=\boldsymbol{\Sigma}_n-\boldsymbol{\Sigma}^*$. Then, under the conditions of Proposition 1, the solution of problem (A.20) $(\widehat{\mathbf{S}}_\Omega,\widehat{\mathbf{L}}_{\mathcal{T}'})$ satisfies*

$$g_\gamma(\widehat{\mathbf{S}}_\Omega-\mathbf{S}^*,\widehat{\mathbf{L}}_{\mathcal{T}'}-\mathbf{L}^*)\leq 2\widetilde{r}.$$

**Proof.** The proof, analogous to the proof of Proposition 5.2 in [18], follows from Proposition 1 (cf. also Proposition 12 in [31]). □

Let us define the tangent space to $\mathcal{L}(r)$ in a generic $\widetilde{\mathbf{L}}\neq\mathbf{L}^*$:

$$\widetilde{\mathcal{T}}(\widetilde{\mathbf{L}})=\{\mathbf{M}\in\mathbb{R}^{p\times p}\mid\mathbf{M}=\mathbf{U}\mathbf{Y}_1^\top+\mathbf{Y}_2\mathbf{U}^\top\mid\mathbf{Y}_1,\mathbf{Y}_2\in\mathbb{R}^{p\times r},\mathbf{U}\in\mathbb{R}^{p\times r},\mathbf{U}^\top\mathbf{U}=\mathbf{I}_r,\mathbf{U}^\top\widetilde{\mathbf{L}}\mathbf{U}\in\mathbb{R}^{r\times r}\text{diagonal},\widetilde{\mathbf{L}}\in\mathcal{L}(r)\}.$$

Consider the solution pair

$$(\widehat{\mathbf{S}}_\Omega,\widehat{\mathbf{L}}_{\widetilde{\mathcal{T}}})=\arg\min_{\substack{\underline{\mathbf{L}}\in\widetilde{\mathcal{T}}\\ \underline{\mathbf{S}}\in\Omega}}\frac{1}{2p^{\alpha_1}}\|\boldsymbol{\Sigma}_n-(\underline{\mathbf{L}}+\underline{\mathbf{S}})\|_F^2+\psi_0\|\underline{\mathbf{L}}\|_*+\rho_0\|\underline{\mathbf{S}}\|_1.\tag{A.21}$$

**Proposition 3.** *Let $\gamma\in[9\xi(\mathcal{T}(\mathbf{L}^*)),1/(6\mu(\Omega(\mathbf{S}^*)))]$ and suppose that the minimum eigenvalue of $\mathbf{L}^*$ is such that $\lambda_r(\mathbf{L}^*)>\delta_L\psi_0/\xi^2(T)$ and $\|\mathbf{S}^*\|_{\min,\text{off}}>\delta_S\psi_0/\mu(\Omega)$ with $\delta_L$ and $\delta_S$ finite positive reals. Suppose also that*

$$g_\gamma(\mathcal{A}^\dagger\boldsymbol{\Delta}_n)\leq\frac{\psi_0}{18},$$

*with $\boldsymbol{\Delta}_n=\boldsymbol{\Sigma}_n-\boldsymbol{\Sigma}^*$. Then, there exists a unique $\widetilde{\mathcal{T}}$ satisfying Proposition 1 when setting $\widetilde{\mathcal{T}}=\mathcal{T}'$ therein, and a corresponding unique solution pair $\left(\widehat{\mathbf{S}}_\Omega,\widehat{\mathbf{L}}_{\widetilde{\mathcal{T}}}\right)$ of (A.21), such that:* (i) *$\varrho(\mathcal{T},\widetilde{\mathcal{T}})\leq\xi(\mathcal{T})/4$;* (ii) *$\text{rk}(\widehat{\mathbf{L}}_{\widetilde{\mathcal{T}}})=r$, $\text{sgn}(\widehat{\mathbf{S}}_{\Omega,ij})=\text{sgn}(\mathbf{S}^*_{ij})$ for all $i,j\in\{1,\ldots,p\}$;* (iii) *$g_\gamma(A^\dagger\mathbf{C}_{\widetilde{\mathcal{T}}})\leq\psi_0/18$, with $\mathbf{C}_{\widetilde{\mathcal{T}}}=\text{Pr}_{\widetilde{\mathcal{T}}^\perp}(\mathbf{L}^*)$;* (iv) *$\left(\widehat{\mathbf{S}}_\Omega,\widehat{\mathbf{L}}_{\widetilde{\mathcal{T}}}\right)$ is also the unique solution of problem (6).*

**Proof.** The proof is analogous to the proof of Proposition 5.3 in [18], by noticing that $\lambda_r(\mathbf{L}^*)>\delta_L\psi_0/\xi^2(T)$ and $\|\mathbf{S}^*\|_{\min,\text{off}}>\delta_S\psi_0/\mu(\Omega)$ hold under Assumption 6, Propositions 1 and 2 hold under Assumption 5 with $\gamma\in[9\xi(\mathcal{T}(\mathbf{L}^*)),1/(6\mu(\Omega(\mathbf{S}^*)))]$, and $\frac{1}{2}\|\boldsymbol{\Sigma}_n-(\underline{\mathbf{L}}+\underline{\mathbf{S}})\|_F$ has 2nd derivative *wrt* $\underline{\mathbf{L}}$ and $\underline{\mathbf{S}}$ equal to $\mathbf{I}_p\otimes\mathbf{I}_p$. □

At this point, we note that, if $\delta_1\leq\alpha_r/3$, Assumptions 1, 2, 5, and 6 are always compatible for all $p\in\mathbb{N}$ as $n\to\infty$ and $p^{2-2\delta_1}\ln(p)/n=o(1)$ (see Remarks 3 and 4). Then, we note that under Assumptions 1, 2, 3, with probability tending to one as $n \to \infty$, it holds:

$$g_\gamma(\mathcal{A}^\dagger \boldsymbol{\Delta}_n) = g_\gamma(\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*, \boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*) \leq \max\left(\frac{\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_\infty}{\gamma}, \frac{\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_2}{\|\boldsymbol{\Sigma}^*\|_2}\right)$$
$$\leq \max\left(\frac{\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_\infty}{\gamma}, \frac{\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_2}{p^{\alpha_1}}\right) \leq \frac{C'}{9\xi(\mathcal{T})}\sqrt{\frac{\ln(p)}{n}}$$

This result descends from Lemma 6, from the condition $\gamma \in [9\xi(\mathcal{T}(\mathbf{L}^*)), 1/(6\mu(\Omega(\mathbf{S}^*)))]$ of Proposition 1, and from Lemma 1, under Assumptions 1, 2, 3.

Since we have set $\psi_0 = (1/\xi(\mathcal{T}))\sqrt{\ln(p)/n}$, the condition of Proposition 3 can be written as $g_\gamma(\mathcal{A}^\dagger \boldsymbol{\Delta}_n) \leq \psi_0/18 \leq (p^{\delta_1} k_L/(18\sqrt{r}))(\sqrt{\ln(p)/n})$ by Assumption 5. Therefore, setting $C = k_L/(18\sqrt{r})$ and $C' = 1/2$, under Assumptions 1, 2, 3, 5 and 6 Proposition 3 (parts (i), (iii) and (iv)) ensures that the solution $(\widehat{\mathbf{S}}, \widehat{\mathbf{L}})$ of (6) satisfies

$$g_\gamma(\widehat{\mathbf{S}} - \mathbf{S}^*, \widehat{\mathbf{L}} - \mathbf{L}^*) \leq C\frac{80}{9}\psi_0 \leq \kappa\frac{p^{\delta_1}}{\sqrt{n}},$$

where $\kappa = (80 \times k_L)/(9 \times 18\sqrt{r})$. Recalling the definition of $g_\gamma$ in (A.19), we can thus write $\|\widehat{\mathbf{L}} - \mathbf{L}^*\|_2 \leq Cp^{\alpha_1}\frac{80}{9}\psi_0 \leq \kappa p^{\alpha_1+\delta_1}\sqrt{\frac{\ln(p)}{n}}$, $\|\widehat{\mathbf{S}} - \mathbf{S}^*\|_\infty \leq C\frac{80}{9}\gamma\psi_0 \leq \kappa\sqrt{\frac{\ln(p)}{n}}$. This proves parts (i) and (ii) of Theorem 2. Finally, Proposition 3 (parts (ii) and (iv)) ensures that $\Pr(\mathrm{rk}(\widehat{\mathbf{L}}) = r) \to 1$, $\Pr(\mathrm{sgn}(\widehat{\mathbf{S}}) = \mathrm{sgn}(\mathbf{S}^*)) \to 1$, as $n \to \infty$. This proves parts (iii) and (iv) of Theorem 2. □

**Proof of Corollary 1.**

Suppose that all the assumptions and conditions of Theorem 2 hold, and recall that the pair $(\widehat{\mathbf{S}}, \widehat{\mathbf{L}})$ is the solution of (6). Then, part (i) holds true because of Theorem 2 part (ii) and Assumption 2(a), as

$$\|\widehat{\mathbf{S}} - \mathbf{S}^*\|_2 \leq \kappa\delta_2 p^{\delta_1}\sqrt{\frac{\ln(p)}{n}},$$

where we used arguments analogous to (A.7), (A.8), (A.9).

Part (ii) holds true under Assumption 1(a) and 2(a) because

$$\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_2 \leq \|\widehat{\mathbf{L}} - \mathbf{L}^*\|_2 + \|\widehat{\mathbf{S}} - \mathbf{S}^*\|_2 \leq \kappa p^{\alpha_1+\delta_1}\sqrt{\frac{\ln(p)}{n}} + \kappa\delta_2 p^{\delta_1}\sqrt{\frac{\ln(p)}{n}}.$$

Then, Proposition 3 (part (iv)) ensures part (iii) of the Corollary, as $\widehat{\mathbf{S}} \succ 0$ because $\widehat{\mathbf{S}} \in \mathcal{S}(s)$ as $n \to \infty$. Part (iv) of the Corollary descends by Proposition 3 (parts (ii) and (iv)), because $\Pr(\mathrm{rk}(\widehat{\mathbf{L}}) = r) \to 1$ as $n \to \infty$ (part (iii) of Theorem 2), and $\widehat{\boldsymbol{\Sigma}} \succ 0$ because $\lambda_p(\widehat{\boldsymbol{\Sigma}}) \geq \lambda_p(\widehat{\mathbf{L}}) + \lambda_p(\widehat{\mathbf{S}}) > 0 + \lambda_p(\widehat{\mathbf{S}}) > 0$, by dual Lidksii inequality and part (iii) of the Corollary.

Part (v) of the Corollary holds because $\|\widehat{\mathbf{S}}^{-1} - \mathbf{S}^{*-1}\| \leq \lambda_p(\mathbf{S}^*)^{-1}\lambda_p(\widehat{\mathbf{S}})^{-1}\|\widehat{\mathbf{S}} - \mathbf{S}^*\|$, $\lambda_p(\mathbf{S}^*) = O(p^{\alpha_1-1-\varepsilon})$ for some $\varepsilon > 0$ by assumption, and $\lambda_p(\widehat{\mathbf{S}})$ tends to $\lambda_p(\mathbf{S}^*)$ as $n \to \infty$. Analogously, part (vi) holds because $\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{*-1}\| \leq \lambda_p(\boldsymbol{\Sigma}^*)^{-1}\lambda_p(\widehat{\boldsymbol{\Sigma}})^{-1}\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|$, $\lambda_p(\boldsymbol{\Sigma}^*) = O(p^{\alpha_1-1-\varepsilon})$ for some $\varepsilon > 0$ by assumption, and $\lambda_p(\widehat{\boldsymbol{\Sigma}})$ tends to $\lambda_p(\boldsymbol{\Sigma}^*)$ as $n \to \infty$. □

**Proof of Theorem 3.**

We start by noticing that, under all the assumptions and conditions of Theorems 1 and 2, problems (3) and (6) are equivalent as $p, n \to \infty$, provided that $\psi/p \to 0$. To see that, it is enough to consider problem (A.20), and to observe that, under Assumption 5, $\psi_0 = O\left(p^{\delta_1}/\sqrt{n}\right)$ and $\rho_0 = O(1)$. It follows that, in problem (6), $\psi = O\left(p^{\alpha_1+\delta_1}/\sqrt{n}\right)$. As a consequence, in order to have $\psi/p \to 0$, there is an additional error term to be controlled for, that is $O\left(p^{1-\alpha_1-\delta_1}/\sqrt{n}\right)$. This leads to the condition $p^{2(1-\alpha_1-\delta_1)}/n \to 0$ as $p, n \to \infty$. At this stage, we can observe that such condition is inactive under the condition $p^{2(1-\alpha_r)}/n \to 0$ unless $\alpha_1 - \alpha_r > \delta_1$. It follows that the assumption $\alpha_1 - \alpha_r \leq \delta_1$ is enough to make Theorem 1 hold for $\widehat{\mathbf{B}}_{\mathrm{ALCE}1}, \widehat{\mathbf{F}}_{\mathrm{ALCE}1}, \mathbf{H}_{\mathrm{ALCE}1}$ in the place of $\widehat{\mathbf{B}}_{\mathrm{OLS}1}, \widehat{\mathbf{F}}_{\mathrm{OLS}1}$, and $\mathbf{H}_{\mathrm{OLS}1}$. Since the assumption $\alpha_1 - \alpha_r \leq \delta_1$ is already imposed in Theorem 1, Theorem 3 follows. □

**Proof of Theorem 4.**

Let us define $\mathbf{Y}_{\mathrm{pre}}$ and $\mathbf{Z}_{\mathrm{pre}}$ as the last updates of Algorithm 1 (Section 1 in the Supplement), with $\boldsymbol{\Sigma}_{\mathrm{pre}} = \mathbf{Y}_{\mathrm{pre}} + \mathbf{Z}_{\mathrm{pre}}$. We note that the matrices $\mathbf{Y}_{\mathrm{pre}}$, $\mathbf{Z}_{\mathrm{pre}}$ and $\boldsymbol{\Sigma}_{\mathrm{pre}}$ completely depend on the sample covariance matrix $\boldsymbol{\Sigma}_n$. Therefore, assuming as in Theorem 4 that $\boldsymbol{\Sigma}_n$ is fixed, we can decompose the minimization problem (17) assuming as fixed the matrices $\mathbf{Y}_{\mathrm{pre}}$ and $\mathbf{Z}_{\mathrm{pre}}$.

We start considering the loss $\|\boldsymbol{\Sigma}_n - (\mathbf{L} + \mathbf{S})\|_2$ with reference to $\mathbf{Y}_{\mathrm{pre}}$ and $\mathbf{Z}_{\mathrm{pre}}$. By the triangular inequality, we can write

$$\begin{aligned}\|\boldsymbol{\Sigma}_n - (\mathbf{L} + \mathbf{S})\|_2 &= \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_{\mathrm{pre}} + \boldsymbol{\Sigma}_{\mathrm{pre}} - (\mathbf{L} + \mathbf{S})\|_2 \leq \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_{\mathrm{pre}}\| + \|\boldsymbol{\Sigma}_{\mathrm{pre}} - (\mathbf{L} + \mathbf{S})\|_2 \\ &\leq \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_{\mathrm{pre}}\|_2 + \|\mathbf{L} - \mathbf{Y}_{\mathrm{pre}}\|_2 + \|\mathbf{S} - \mathbf{Z}_{\mathrm{pre}}\|_2, \end{aligned} \tag{A.22}$$

where $\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_{\text{pre}}$ is fixed. Therefore, in order to minimize $\|\boldsymbol{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_2$ under the given constraints, we can focus on the problems $\arg\min_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A)} \|\mathbf{L}-\mathbf{Y}_{\text{pre}}\|_2$ and $\arg\min_{\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|\mathbf{S}-\mathbf{Z}_{\text{pre}}\|_2$.

First, we can note that $\arg\min_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A)} \|\mathbf{L}-\mathbf{Y}_{\text{pre}}\|_2 = \widehat{\mathbf{L}}_{\text{UNALCE}}$, because of the optimal approximation property of principal components, proved in [21], and since $\widehat{\mathbf{L}}_{\text{UNALCE}}$ is derived by the top $\widehat{r}$ principal components of $\mathbf{Y}_{\text{pre}}$. Part (i) of Theorem 4 then follows because, by the triangular inequality,

$$\|\mathbf{L}-\mathbf{L}^*\|_2 = \|\mathbf{L}-\mathbf{Y}_{\text{pre}}+\mathbf{Y}_{\text{pre}}-\mathbf{L}^*\|_2 \le \|\mathbf{L}-\mathbf{Y}_{\text{pre}}\|_2 + \|\mathbf{Y}_{\text{pre}}-\mathbf{L}^*\|_2, \tag{A.23}$$

and $\mathbf{Y}_{\text{pre}}-\mathbf{L}^*$ is fixed. Note that, as explained in Remark 10, the inequality in (A.23) tends to an equality if $\delta_1 > 0$ and $p$ is large enough, because $\mathbf{Y}_{\text{pre}} \in \widehat{\mathcal{L}}(\widehat{r}_A)$, $\mathbf{L}^* \in \mathcal{L}(r)$, and Proposition 3 part (i) ensures that $\varrho(\mathcal{L}(r), \mathcal{L}(\widehat{r}_A)) \to 0$ as $p \to \infty$.

The problem in $\mathbf{S}$ can be rewritten as follows. Suppose that, by exploiting the assumption of Theorem 4 $\text{diag}(\mathbf{L}) + \text{diag}(\mathbf{S}) = \text{diag}(\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}})$, we constrain our search within the set of matrices $\mathbf{S}$ such that $\text{diag}(\mathbf{S}) = \text{diag}(\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}) - \text{diag}(\mathbf{L})$, with $\mathbf{L} \in \widehat{\mathcal{L}}(\widehat{r}_A)$. Then, also assuming the invariance of the off-diagonal elements in $\widehat{\mathbf{S}}$ as in Theorem 4, we can write

$$\begin{aligned}\min_{\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|\mathbf{S}-\mathbf{Z}_{\text{pre}}\|_2 &= \min_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A)} \|\text{diag}(\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}-\mathbf{L}) - \text{diag}(\boldsymbol{\Sigma}_{\text{pre}}-\mathbf{Y}_{\text{pre}})\|_2\\ &\le \min_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A)} \|\text{diag}(\mathbf{L}-\mathbf{Y}_{\text{pre}})\|_2 + \|\text{diag}(\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}-\boldsymbol{\Sigma}_{\text{pre}})\|_2 \le r \min_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A)} \|\mathbf{L}-\mathbf{Y}_{\text{pre}}\|_2 + \|\text{diag}(\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}}-\boldsymbol{\Sigma}_{\text{pre}})\|_2.\end{aligned}$$

Since $\widehat{\boldsymbol{\Sigma}}_{\text{ALCE}} - \boldsymbol{\Sigma}_{\text{pre}}$ is fixed, from part (i) it follows that $\arg\min\max_{\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|\mathbf{S}-\mathbf{Z}_{\text{pre}}\|_2 = \widehat{\mathbf{S}}_{\text{UNALCE}}$. From the triangular inequality

$$\|\mathbf{S}-\mathbf{S}^*\|_2 = \|(\mathbf{S}-\mathbf{Z}_{\text{pre}})+(\mathbf{Z}_{\text{pre}}-\mathbf{S}^*)\|_2 \le \|\mathbf{S}-\mathbf{Z}_{\text{pre}}\|_2 + \|\mathbf{Z}_{\text{pre}}-\mathbf{S}^*\|_2$$

part (ii) of Theorem 4 then follows, because $\mathbf{Z}_{\text{pre}}-\mathbf{S}^*$ is fixed.

From parts (i) and (ii), it also follows from (A.22) that

$$\arg\min \max_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A),\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|\boldsymbol{\Sigma}_n - (\mathbf{L}+\mathbf{S})\|_2 = \left(\widehat{\mathbf{L}}_{\text{UNALCE}}, \widehat{\mathbf{S}}_{\text{UNALCE}}\right),$$

under all the assumptions and conditions of Theorem 4.

The problem in $\boldsymbol{\Sigma} = \mathbf{L}+\mathbf{S}$ can be rewritten as

$$\min_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A),\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|(\mathbf{L}+\mathbf{S})-\boldsymbol{\Sigma}^*\|_2 = \min_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A),\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|\mathbf{L}-\mathbf{L}^*+\mathbf{S}-\mathbf{S}^*\|_2 \le \min_{\mathbf{L}\in\widehat{\mathcal{L}}(\widehat{r}_A)} \|\mathbf{L}-\mathbf{L}^*\|_2 + \min_{\mathbf{S}\in\widehat{\mathcal{S}}_{diag}} \|\mathbf{S}-\mathbf{S}^*\|_2$$

such that, from parts (i) and (ii) of Theorem 4, part (iii) follows.

Finally, the same optimality properties are transmitted to $\widehat{\mathbf{S}}_{\text{UNALCE}}^{-1}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{UNALCE}}^{-1}$. In fact, under the conditions of Corollary 1, it holds $\|\widehat{\mathbf{S}}^{-1} - \mathbf{S}^{*-1}\|_2 \le \|\widehat{\mathbf{S}}^{-1}\|_2\|\widehat{\mathbf{S}}-\mathbf{S}^*\|_2\|\mathbf{S}^{*-1}\|_2 \le \lambda_p(\widehat{\mathbf{S}})^{-1}\lambda_p(\mathbf{S}^*)^{-1}\|\widehat{\mathbf{S}}-\mathbf{S}^*\|_2$ with $\lambda_p(\widehat{\mathbf{S}}_{\text{UNALCE}}) > \lambda_p(\widehat{\mathbf{S}}_{\text{ALCE}}) - \lambda_p(\widehat{\mathbf{U}}_{\text{ALCE}}\breve{\psi}\mathbf{I}_r\widehat{\mathbf{U}}_{\text{ALCE}}^\top)$, and

$$-\frac{r\breve{\psi}}{p} \le -\lambda_p(\widehat{\mathbf{U}}_{\text{ALCE}}\breve{\psi}\mathbf{I}_r\widehat{\mathbf{U}}_{\text{ALCE}}^\top) \le 0, \tag{A.24}$$

where $r\breve{\psi}/p$ is close to 0 if Theorem 3 holds, provided that $p$ is large enough. Part (iv) then follows.

Similarly, it holds $\|\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{*-1}\|_2 \le \|\widehat{\boldsymbol{\Sigma}}^{-1}\|_2\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^*\|_2 \times \|\boldsymbol{\Sigma}^{*-1}\|_2 \le \lambda_p(\widehat{\boldsymbol{\Sigma}})^{-1}\lambda_p(\boldsymbol{\Sigma}^*)^{-1}\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}^*\|_2$ with $\lambda_p(\widehat{\boldsymbol{\Sigma}}_{\text{UNALCE}}) > \lambda_p(\widehat{\mathbf{L}}_{\text{UNALCE}}) + \lambda_p(\widehat{\mathbf{S}}_{\text{UNALCE}}) > 0 + \lambda_p(\widehat{\mathbf{S}}_{\text{UNALCE}})$, such that part (v) follows under the conditions of Corollary 1 and Theorem 3 by (A.24). □

## Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2023.105244. Additional information for this article is available, in the form of an online supplement containing the pseudo-code of the solution algorithm of problem (6), some further technical results, the criterion used to select the optimal threshold pair $(\psi, \rho)$ while solving (6), a wide simulation study and a real data example.

## References

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

[2] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1958.

[3] T.W. Anderson, H. Rubin, Statistical inference in factor analysis, in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Univ of California Press, 1956, pp. 111–150.

[4] J. Bai, Inferential theory for factor models of large dimensions, Econometrica 71 (1) (2003) 135–171.

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

[6] J. Bai, S. Ng, Determining the number of factors in approximate factor models, Econometrica 70 (1) (2002) 191–221.

[7] J. Bai, S. Ng, Large dimensional factor analysis, Found. Trends Econom. 3 (2) (2008) 89–163.

[8] J. Bai, S. Ng, Principal components estimation and identification of static factors, J. Econometrics 176 (1) (2013) 18–29.

[9] J. Bai, S. Ng, Rank regularized estimation of approximate factor models, J. Econometrics 212 (1) (2019) 78–96.
[10] M. Barigozzi, M. Farnè, An algebraic estimator for large spectral density matrices, J. Amer. Statist. Assoc. (2022) 1–13.
[11] M.S. Bartlett, The statistical conception of mental factors, Br. J. Psychol. 28 (1) (1937) 97.
[12] P.J. Bickel, E. Levina, Covariance regularization by thresholding, Ann. Statist. (2008) 2577–2604.
[13] C.M. Bishop, N.M. Nasrabadi, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[14] J.-F. Cai, E.J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (4) (2010) 1956–1982.
[15] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM 58 (3) (2011) 1–37.
[16] E.J. Candès, T. Tao, The power of convex relaxation: Near-optimal matrix completion, IEEE Trans. Inform. Theory 56 (5) (2010) 2053–2080.
[17] G. Chamberlain, M. Rothschild, Arbitrage, factor structure, and mean–variance analysis on large asset markets, Econometrica 51 (5) (1983) 1281–1304.
[18] V. Chandrasekaran, P.A. Parrilo, A.S. Willsky, Latent variable graphical model selection via convex optimization, Ann. Statist. 40 (4) (2012) 1935–1967.
[19] V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, A.S. Willsky, Rank-sparsity incoherence for matrix decomposition, SIAM J. Optim. 21 (2) (2011) 572–596.
[20] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Comm. Pure Appl. Math. 57 (11) (2004) 1413–1457.
[21] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, Psychometrika 1 (3) (1936) 211–218.
[22] J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements, J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (4) (2013) 603–680.
[23] M. Farnè, A. Montanari, A large covariance matrix estimator under intermediate spikiness regimes, J. Multivariate Anal. 176 (2020) 104577.
[24] M. Fazel, Matrix Rank Minimization with Applications (Ph.D. thesis), Stanford University, 2002.
[25] M. Fazel, H. Hindi, S.P. Boyd, A rank minimization heuristic with application to minimum order system approximation, in: Proceedings of the 2001 American Control Conference, IEEE, 2001, pp. 4734–4739.
[26] T. Hastie, R. Mazumder, J.D. Lee, R. Zadeh, Matrix completion and low-rank svd via fast alternating least squares, J. Mach. Learn. Res. 16 (1) (2015) 3367–3402.
[27] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417.
[28] I.T. Jolliffe, Principal Component Analysis, in: Springer Series in Statistics, Springer, New York, 2002.
[29] K.G. Jöreskog, Some contributions to maximum likelihood factor analysis, Psychometrika 32 (4) (1967) 443–482.
[30] D.N. Lawley, A.E. Maxwell, Factor Analysis As a Statistical Method, Butterworths, London, 1971.
[31] X. Luo, High dimensional low rank and sparse covariance matrix estimation via convex minimization, 2011, arXiv preprint.
[32] V.A. Marčnko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Mat. Sb. 114 (4) (1967) 507–536.
[33] R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, J. Mach. Learn. Res. 11 (2010) 2287–2322.
[34] A. Onatski, Asymptotics of the principal components estimator of large factor models with weakly influential factors, J. Econometrics 168 (2) (2012) 244–258.
[35] N. Srebro, J. Rennie, T.S. Jaakkola, Maximum-margin matrix factorization, in: Advances in Neural Information Processing Systems, Vol. 17, MIT Press, 2004.
[36] T. Tao, Topics in Random Matrix Theory, in: Graduate Studies in Mathematics, vol. 132, American Mathematical Society, 2012.
[37] G.H. Thomson, The definition and measurement of g (general intelligence), J. Educ. Psychol. 26 (4) (1935) 241.
[38] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2) (2006) 265–286.