# LOW RANK MATRIX ESTIMATION AND ASYMPTOTIC THEORY OF BOOTSTRAP

MAYUKH CHOUDHURY, DEBRAJ DAS, AND ARINDAM CHATTERJEE

ABSTRACT. In this paper, we study inferential aspects of low-rank matrix linear models under nuclear-norm regularization. While existing literature largely focuses on estimation error bounds and non-asymptotic guarantees, the asymptotic distributional theory of nuclear-norm penalized estimators remains underdeveloped even in classical fixed-dimension settings. We establish consistency and derive the asymptotic distribution of the nuclear-norm penalized estimator when the number of predictors and responses are fixed and the sample size diverges. The limiting law is shown to be non-standard, arising as the minimizer of a random convex objective involving Gaussian matrix perturbations and singular-vector projections, which complicates direct inference. To enable valid inference, we develop a Perturbation Bootstrap procedure tailored to the nuclear-norm penalized estimator. A key obstacle is the lack of rank-selection consistency of the estimator, which invalidates naive bootstrap approximations. We overcome this by introducing a thresholded version of the estimator that achieves strong rank-selection consistency without prior knowledge of the true rank. Centering the bootstrap pivot at this thresholded estimator, we prove that the conditional bootstrap distribution consistently approximates the unconditional asymptotic law of the estimator. We also characterize the limiting behavior of the estimator's singular values and show that a suitably centered Perturbation Bootstrap consistently reproduces these limits for inference. The proposed method provides a theoretically justified approach for constructing confidence regions and hypothesis tests for the coefficient matrix in low-rank multivariate regression models.

## 1. INTRODUCTION

Consider the multivariate linear regression model:

$$Y = XA + E, \tag{1.1}$$

where $Y \in \mathbf{R}^{n\times T}$ is the response matrix with $j$-th column denoting $n$ independent observations on $j$-th response variable. $X \in \mathbf{R}^{n\times p}$ is the non-random design matrix and $A \in \mathbf{R}^{p\times T}$ is the unknown parameter matrix. the randomness in the model is governed by the $n \times T$ error matrix $E$. Clearly, $n$, $p$ and $T$ respectively denote the number of observations, the number of predictors and the number of responses. Throughout the paper, we consider $p$ and $T$ to be fixed and $n$ can grow to $\infty$. Multivariate linear regression model (1.1) is widely used when several related responses are observed for the same set of predictors. In econometrics, such models are employed to jointly analyze multiple economic indicators while accounting for correlation across responses, as studied in the framework of seemingly unrelated regressions by Zellner (1962). In environmental and climate sciences, they are used to relate meteorological covariates to multiple climate variables such as temperature, precipitation, and wind speed (cf. Wilks (2019)). In machine learning and applied statistics, these models form the foundation of multi-output and multi-task regression, where related prediction problems are solved jointly by Breiman and Friedman

*Key words and phrases.* Low rank matrix, Matrix perturbation, Multivariate regression, Nuclear norm, PB, SVD. .

(1997). Similar approaches arise in biomedical and chemometrics applications, where correlated biological or chemical responses are modeled simultaneously to improve prediction and interpretation (cf. Reinsel and Velu (1998)). Development of the statistical inference methodologies for the model (1.1) under varied generalizations is an active field of research, although the foundational work has long been done by considering the model (1.1) an extension of the usual multiple linear regression model. Mardia et al. (1979), Muirhead (1982), and Anderson (2003) developed distributional theory, likelihood-based inference techniques, and hypothesis testing procedures under multivariate normal assumption of the errors. In particular, Muirhead (1982) provided comprehensive results on estimation and testing for the parameter matrix $A$, while Mardia et al. (1979) discussed multivariate regression, principal component analysis, and canonical correlations. Later works such as Johnson and Wichern (2007) and Rencher and Christensen (2012) extended classical multivariate analysis by developing unified estimation and hypothesis testing procedures for generalized MANOVA and growth curve models, particularly in longitudinal and repeated-measures settings with complex covariance structures.

Let us start with the special case of $T = 1$ in the model (1.1), i.e., we are in the multiple linear regression setup. Now when the number of covariates $p$ is considerably large, then it is natural to expect that many of these covariates are actually irrelevant. Such a setting can be identified by imposing sparsity in the multiple linear regression model. The most well-known method of imposing sparsity in (1.1) with $T = 1$ and to identify the relevant covariates is the Lasso, introduced by Tibshirani (1996). However, when $T > 1$ then the relevance of a particular covariate depends on the joint structure of the response variables, since a covaraite may have no effect on one response variable while remain influential for another response variable. Therefore, entry-wise sparsity of the coefficient matrix $A$ does not adequately capture predictor irrelevance in the multi-response framework. Instead, one may assume that $A$ has some low rank structure. For example, if $\mathrm{rank}(A) = r$ with $r \ll \min(p, T)$, then there exist matrices $U \in \mathbf{R}^{p\times r}$ and $V \in \mathbf{R}^{T\times r}$ such that $A = UV^\top$. Under this factorization, the multivariate regression model (1.1) reduces to

$$Y = XUV^\top + E = X^*V^\top + E = (X^*V_1^\top, X^*V_2^\top) + E, \tag{1.2}$$

where $X^* = XU$ and $V = (V_1^\top, V_2^\top)^\top$ with $V_1 \in \mathbf{R}^{r\times r}$. When $r < p$, then (1.2) indicates that $X^* \in \mathbf{R}^{n\times p}$ can be thought of as the latent design matrix. In other words, only $r$ ($< p$) many linear combinations of the the original covariates are sufficient to explain $T$ many response variables. On the other hand, when $r < T$, then (1.2) essentially implies that it is enough to analyze only $r$ ($< T$) many response variables because on an average the remaining $(T - r)$ ones are actually linear combination of those $r$ many response variables. Thus, the low rank structure of the parameter matrix $A$ can effectively identify linear dependence among the covariates as well as among the response variables in model (1.1). Hence, it is very natural to assume a low rank structure in (1.1) when either the number of covariates ($p$) or the number of responses ($T$) is large. Early investigations into low-rank multivariate regression were carried out by Izenman (1975) and Davies and Tso (1982), who studied least-squares estimators under explicit rank constraints on the coefficient matrix (known as reduced rank estimator) and established their inferential properties. The same was further developed by Reinsel and Velu (1998),Bach (2008), Giraud (2011) and Giraud (2021) who studied oracle inequalities and prediction guarantees for reduced-rank estimators. Across these works, the rank is either assumed to be known a priori or selected via model selection or penalization schemes, depending on the specific formulation (cf. Bach (2008); Giraud (2021)).

Least square with rank constraint is a non-convex problem leading to non-uniqueness of the reduced rank estimator and its theoretical analysis generally requires the knowledge of the low rank structure of the model (1.1). The convex relaxation of the rank constraint is the nuclear norm penalty. Therefore, a practical alternative to the reduced rank estimator is the nuclear norm penalized (hereafter referred to as NNP) estimator which is defined as

$$\begin{aligned}\tilde{A}_n &= \arg\min_{B\in\mathbf{R}^{p\times T}} \Big\{\|Y - XB\|_F^2 + \lambda_n\|\sigma(B)\|_1\Big\} \\ &= \arg\min_{B\in\mathbf{R}^{p\times T}} \Big\{\sum_{i=1}^{n}\|\boldsymbol{Y}_i - \boldsymbol{X}_i B\|_2^2 + \lambda_n\|\sigma(B)\|_1\Big\}, \end{aligned} \tag{1.3}$$

where $\boldsymbol{Y}_i \in \mathbf{R}^{1\times T}$ and $\boldsymbol{X}_i \in \mathbf{R}^{1\times p}$ denote the $i$-th rows of $Y$ and $X$, respectively. $\lambda_n$ ($> 0$) is the penalty parameter. Thus, $\tilde{A}_n$ is the minimizer of the NNP least squared criterion function. Some preliminary theoretical results on the NNP estimator $\tilde{A}_n$ and on its implementation are due to Goldberg et al. (2010), Bunea et al. (2011), Chen and Huang (2012),Chen et al. (2013); among others. Particularly, Bunea et al. (2011) establish oracle inequalities and optimal rates for reduced-rank and NNP estimators in high-dimensional multivariate regression. Chen et al. (2013) develop adaptive nuclear norm methods, while Chen and Huang (2012) study convex relaxations with statistical guarantees for low-rank estimation. These works focus on estimation and prediction bounds under rank-regularized or convex formulations, often under more structured design assumptions than the general regression setting considered here. However, relatively less is known on the distributional properties of $\tilde{A}_n$ and its different smooth and non-smooth functions (e.g. entries, different types of norms, minimum of the absolute row or column sums of $\tilde{A}_n$ etc.) even when both $p$ and $T$ are fixed. In this paper, we try to fill this gap by establishing the asymptotic distribution of $\tilde{A}_n$ when the parameter matrix $A$ has low rank structure with no knowledge of its true rank.

We first establish consistency properties of $\tilde{A}_n$ under suitable regularity conditions and characterize the role of the penalty sequence $\lambda_n$ (cf. Theorem 3.1). In particular, we demonstrate that when $n^{-1}\lambda_n \to \lambda_0 > 0$, the estimator $\tilde{A}_n$ converges in probability to the minimizer of a penalized population risk function, thereby introducing an asymptotic bias induced by the nuclear-norm regularization. However, when $n^{-1}\lambda_n \to 0$, the penalization term becomes asymptotically negligible, and $\tilde{A}_n$ coincides with the least squares estimator, resulting in consistency for $A$ but without enforcing a low-rank structure. Subsequently in Theorem 3.2, we derive the asymptotic distribution of the scaled estimator $\sqrt{n}(\tilde{A}_n - A)$ and show that its limiting distribution is characterized as the minimizer of a random convex objective function involving Gaussian matrix perturbations and singular-vector projections of $A$. The resulting limit law is inherently non-standard and typically involves both continuous and discrete components, which prevents direct derivation of marginal limiting distributions for individual entries of $\tilde{A}_n$. These structural complications highlight the limitations of classical inference and motivate the development of resampling-based procedures. In particular, bootstrap methods are explored to approximate the sampling distribution of the estimator and facilitate valid statistical inference in the low-rank penalized multivariate regression setting.

Bootstrap methods for matrix-valued and multivariate regression models have been studied as practical tools for inference and uncertainty quantification. In structured matrix settings, Milan & Whittaker (1995) explored parametric bootstrap approaches based on singular value decomposition frameworks. Subsequently, Portier & Delyon (2014) develop a weighted bootstrap testing procedure for rank determination in least-squares constrained

estimation and establish its consistency under suitable singular value separation conditions. Josse & Wager (2016) then study bootstrap methods in low-rank matrix denoising, focusing on stabilizing singular value estimates under noise. In multivariate regression settings, Eck (2018) develop residual and paired bootstrap procedures for inference on the coefficient matrix, with methodological roots in classical regression bootstrap ideas of Freedman (1981). More recently, Saegusa et al. (2022) construct parametric bootstrap confidence intervals in multivariate Fay–Herriot models. However, these approaches do not address bootstrap validity for non-smooth penalized estimators in low-rank matrix linear regression models from a fully inferential perspective.

In this paper, we will devise the Perturbation Bootstrap (PB) estimator $\check{A}_n^*$ for NNP estimator. See section 4 for the definition. We first address the incompetency of the PB pivotal quantity $\sqrt{n}(\check{A}_n^* - \tilde{A}_n)$ centered at original $\tilde{A}_n$ to approximate the original pivotal quantity. The discussion reveals a fundamental obstacle in bootstrap approximation for NNP estimators. The NNP estimator $\tilde{A}_n$ is not rank-selection consistent (hereafter referred to as RSC) by default. That is the singular values of $\tilde{A}_n$ corresponding to $\sigma_j = 0$ for all $j > r^*(= \text{rank}(A))$, may be positive or 0 with positive probability, erring to capture the true rank. To circumvent this issue, we introduce a thresholded version $\tilde{A}_n^{(thr)}$ of the nuclear–norm estimator that enforces rank identification at an appropriate rate. In Proposition 4.1, we establish that this thresholded estimator is RSC almost surely, that is,

$$\text{rank}\big(\tilde{A}_n^{(\text{thr})}\big) = r^* \quad \text{eventually almost surely.}$$

Building on this stabilization, we propose a modified PB pivotal quantity $\sqrt{n}(\tilde{A}_n^{*(mod)} - \tilde{A}_n^{(thr)})$ centered at the thresholded estimator with $\check{A}_n^* = \tilde{A}_n^{*(mod)}$ (see section 4). Finally in Theorem 4.1, we prove that the conditional distribution of the modified PB pivot converges almost surely to the same fixed limit law that is governed by $\sqrt{n}(\tilde{A}_n - A)$. From an inferential standpoint, this result provides a rigorous justification for bootstrap-based confidence regions and hypothesis tests for the unknown parameter matrix $A$ or its derived functionals in low-rank multivariate models. Importantly, the procedure does not require prior knowledge of the true rank. Additionally we establish the asymptotic distribution and its respective Bootstrap validity of properly centered and scaled singular values of $A$ (cf. Theorem 3.3 and Corollary 4.1).

The rest of the paper is organized as follows. Section 2 introduces the notations and regularity conditions required for main results. Section 3 establishes consistency properties and asymptotic distribution of the NNP estimator. The PB estimator in this context is defined in section 4 and rank consistency of the thresholded estimator is addressed. This section also studies the asymptotic validity of the PB procedure. Section 5 contains extensive simulation study on finite sample results. Proofs of all the requisite lemmas and main results have been attributed to section 5.

## 2. Notations and Regularity Conditions

In this section we first provide requisite notations for our main results. Let $\{X_n\}_{n\geq 1}$ be a sequence of random matrices in $\mathbf{R}^{p\times T}$ and let $X$ be a $\mathbf{R}^{p\times T}$–valued random matrix. We write $\text{vec}(A) \in \mathbf{R}^{pT}$ for the usual column-stacking vectorization of a matrix $A$. When we say variance of a matrix $A$, we mean variance of $vec(A)$. We use the matrix and its vectorization interchangeably. For a random matrix $\boldsymbol{W} \in \mathbf{R}^{p\times T}$, we will denote the matrix-normal representation as,

$$\mathbf{W} \sim \mathcal{MN}_{p\times T}\big(\mathbf{0}_{p\times T},\, U = \sigma^2 C,\, V = I_T\big),$$

Also by definition, $\mathrm{vec}(\mathcal{MN}(M,U,V))$ is equivalent in distribution to $N_{pT}(\mathrm{vec}(M),\, V\otimes U)$. Here $\otimes$ denote the usual Kronecker's product between two matrices. $\mathbf{E}_*(\cdot)$ denote the conditional expectation given the noise matrix $E$ or its entries $(\varepsilon_{ij})$. Define the linear term, gram matrix and residuals as:

$$W_n = \frac{1}{\sqrt{n}}\, X^\top E, \qquad C_n = \frac{1}{n} X^\top X, \qquad \tilde{E} = Y - X\tilde{A}_n.$$

Now we are ready to list the assumptions.

- (C.1) Let $E_{n\times T} = (\varepsilon_{i,j})$ and assume that $\{\varepsilon_{i,j} : i \geq 1,\ j \geq 1\}$ are i.i.d. with mean 0, variance $\sigma^2$ and $\mathbf{E}(|\varepsilon_{1,1}|^3) < \infty$.
- (C.2) $\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{X}_i\|^3 = O(1)$.
- (C.3) There exists a positive definite matrix $C \in \mathbf{R}^{p\times p}$ such that $\frac{1}{n}X^\top X \ \longrightarrow\ C$.
- (C.4) $A$ admits the SVD representation: $A = \sum_{j=1}^{r^*} \boldsymbol{u}_j\, \sigma_j(A)\, (\boldsymbol{v}_j)^\top$, where $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are the left and right singular vectors associated with the ordered singular values $\sigma_1(A) > \sigma_2(A) > .... > \sigma_{r^*}(A) > 0$ and $\sigma_j(A) = 0$ for all $j > r^*$.

We now briefly discuss the above regularity conditions. Conditions (C.1) and (C.2) together are required to control stochastic fluctuation terms of the form $n^{-1/2}X^\top E$ and related empirical quantities arising in multivariate regression. Those controls are in turn required for asymptotic normality and bootstrap validity. Similar regularity conditions are standard in the literature on high-dimensional penalized M-estimation and low-rank matrix recovery; see, for example, Negahban and Wainwright (2011), Bunea et al. (2011), Rohde and Tsybakov (2011), and Koltchinskii et al. (2011). However, one can drop the identical distributed structure of the enries of $E$ consider the following condition instead:

(C.1)′: The random error vectors $\{\boldsymbol{E}_i\}_{i=1}^n$ are independent with $\mathbf{E}(\boldsymbol{E}_i) = 0, \mathbf{Cov}(\boldsymbol{E}_i) = \bar{\Sigma}_i$, such that $\frac{1}{n}\sum_{i=1}^n \mathbf{E}\|\boldsymbol{E}_i\|^3 = O(1)$ and $\frac{1}{n}\sum_{i=1}^n \bar{\Sigma}_i \ \longrightarrow\ \bar{\Sigma}$, for some positive definite matrix $\Sigma$.

Condition (C.3) ensures stability of the quadratic approximation to the objective function and is essential for applying the argmin continuous mapping theorem and deriving asymptotic distributions. This condition is a natural one and is considered almost always in the asymptotic analysis of Lasso regression; see for example Knight and Fu (2000) and Chatterjee and Lahiri (2011). Assumption (C.4) formalizes the low-rank structure underlying the nuclear norm penalization framework. The singular value decomposition plays a central role in the analysis of trace-norm regularized estimators; see Negahban and Wainwright (2011), Koltchinskii et al. (2011), and Candès and Recht (2012). These structural decompositions are fundamental in establishing oracle inequalities, consistency, rank recovery, and perturbation expansions for low-rank estimators. The strict ordering of the nonzero singular values in (C.4) implies that the positive singular values are simple which ensures identifiability of the singular vectors together required for smooth perturbation expansions of singular values and singular subspaces. Without spectral separation, the singular vectors are not uniquely identifiable and hence perturbation expansions may fail; see for example Davis and Kahan (1970), Stewart (1990), and Mardia et al. (1979). These perturbation expansions are crucial for obtaining explicit form of the asymptotic distribution of any functional of $\tilde{A}_n$, in particular of the singular values of $\tilde{A}_n$, based on Hadamard directional differentiability (see for example section B). We denote $\sigma_j(A)$ as $\sigma_j$ from now.

## 3. Main Results on Asymptotic Distribution of $\tilde{A}_n$

We explore the asymptotic properties of the NNP estimator $\tilde{A}_n$. First, we show that $\tilde{A}_n$ is consistent. We summarize the result in the next theorem.

**Theorem 3.1.** *Assume that the conditions (C.1)–(C.3) hold and* $n^{-1}\lambda_n \to \lambda_0 \in [0, \infty)$. *Then* $\tilde{A}_n$ *defined in (1.3) satisfies*

$$\tilde{A}_n \xrightarrow{P} \arg \min_{B \in \mathbf{R}^{p \times T}} Z(B), \qquad Z(B) := \operatorname{tr}\big((B - A)^\top C (B - A)\big) + \lambda_0 \left\|\sigma(B)\right\|_1.$$

The proof of Theorem 3.1 is presented in section C.1. Theorem 3.1 tells us that if $\lambda_n$ grows slower than $n$ then $\tilde{A}_n$ is consistent for $A$. On the other hand, if $\lambda_0 > 0$, i.e., if $\lambda_n$ does not grow slower than $n$, then the NNP estimator $\tilde{A}_n$ is an inconsistent estimator of $A$, because of the presence of the additional term $\lambda_0 \left\|\sigma(B)\right\|_1$. Moreover, consistency of the coefficient matrix immediately implies consistency of its singular values. Indeed, since singular values are continuous functions of the matrix entries, the continuous mapping theorem yields

$$\sigma_j(\tilde{A}_n) \xrightarrow{P} \sigma_j, \quad j = 1, \ldots, \min(p, T), \;\; \text{provided } \lambda_n = o(n).$$

Therefore, when the penalty parameter grows slower than $n$, the NNP estimator consistently recovers not only the coefficient matrix but also its underlying singular value structure. Importantly, Theorem 3.1 is an extension of Theorem 1 of Knight and Fu (2000), where the consistency of the Lasso estimator was explored. The next theorem finds out the asymptotic distribution of $\tilde{A}_n$.

**Theorem 3.2.** *Suppose the assumptions (C.1)–(C.4) hold true. Assume that* $n^{-1/2}\lambda_n \to \lambda_0 \in [0, \infty)$, *as* $n \to \infty$. *Then we have* $\sqrt{n}\,(\tilde{A}_n - A) \xrightarrow{d} \arg\min_{U \in \mathbf{R}^{p \times T}} H(U)$, *where,*

(3.1)

$$H(U) = -2\operatorname{tr}(U^\top W) \;+\; \operatorname{tr}(U^\top C U) \;+\; \lambda_0 \sum_{j=1}^{r^*} \big(\boldsymbol{u}_j{}^\top U\, \boldsymbol{v}_j\big) + \; \lambda_0 \left\| \boldsymbol{\sigma}\Big(U_0^\top U V_0\Big) \right\|_1,$$

*with* $U_0, V_0$ *being respectively the left and right singular vector blocks spanning the null space of* $A$ *and* $W = [W_1, \ldots, W_T]$ *is a* $p \times T$ *random matrix with* $W_j \overset{iid}{\sim} N_p(\mathbf{0}, \sigma^2 C)$.

The proof of Theorem 3.2 is presented in section C.2. Theorem 3.2 provides a weak limit for the matrix-valued pivotal quantity $\sqrt{n}(\tilde{A}_n - A)$. This theorem is an extension of Theorem 2 of Knight and Fu (2000). Clearly, the limiting distribution is not tractable and hence an alternative like Bootstrap needs to be explored. However, we establish the asymptotic distribution of the singular values of $\tilde{A}_n$ before moving to explore Bootstrap. Given Theorem 3.2, it is natural to use some form of delta method in order to derive the asymptotic distribution of singular values of $\tilde{A}_n$. The appropriate tool for applying the delta method is the notion of Hadamard differentiability of the smooth function. The theory of functional delta method based on Hadamard differentiability is well-developed and one can consult Shapiro (1991), Dümbgen (1993), chapter 3.1 of Van der Vaart and Wellner (1996), Fang & Santos (2019) among others. We now present the result on the asymptotic distribution of the singular values of $\tilde{A}_n$.

**Theorem 3.3.** *Suppose the assumptions (C.1)–(C.4) hold true. Assume that* $n^{-1/2}\lambda_n \to \lambda_0 \in [0, \infty)$, *as* $n \to \infty$. *Let* $\sigma_1(\tilde{A}_n) \geq \sigma_2(\tilde{A}_n) \geq \ldots \geq \sigma_m(\tilde{A}_n)$ *be singular values of* $\tilde{A}_n$ *where* $m = \min\{p, T\}$ *and* $U^* = \arg\min_{U \in \mathbf{R}^{p \times T}} H(U)$ *be the limit in Theorem 3.2. Also assume that* $U_0, V_0$ *are respectively the left and right singular vector blocks spanning the null spaces of* $A$. *Then*

*(a) for each* $j = 1, \ldots, r^*$*, (where* $r^* = \mathrm{rank}(A)$*) we have*

$$\sqrt{n}\big(\sigma_j(\tilde{A}_n) - \sigma_j\big) \xrightarrow{d} (\boldsymbol{u}_j)^\top U^* \boldsymbol{v}_j.$$

*(b) for each* $j = r^* + 1, \ldots, m$*,*

$$\sqrt{n}\,\sigma_j(\tilde{A}_n) \xrightarrow{d} \sigma_{j-r^*}\big(U_0^\top U^* V_0\big),$$

*where* $\sigma_1(U_0^\top U^* V_0) \geq \sigma_2(U_0^\top U^* V_0) \geq \ldots \geq \sigma_{m-r^*}(U_0^\top U^* V_0)$ *are* $(m - r^*)$ *singular values of* $(U_0^\top U^* V_0)$*.*

The proof of Theorem 3.3 is presented in section C.3. One can note the difference between the limits in part (a) and (b). This is because the positive singular values $\{\sigma_1 \ldots, \sigma_{r^*}\}$ of $A$ are all simple whereas the singular value 0 has multiplicity $(m - r^*)$. To establish part (a) of the above theorem, we show that for $j = 1, \ldots, r^*$, $\sigma_j(\cdot)$ admits a first order perturbation bound as follows: for any perturbation $S \in \mathbf{R}^{p \times T}$,

$$\sigma_j(A + S) = \sigma_j + (\boldsymbol{u}_j)^\top S\, \boldsymbol{v}_j + O(\|S\|^2); \tag{3.2}$$

based on the section 4 of Stewart (1990). Here $(\boldsymbol{u}_j, \boldsymbol{v}_j)$ is the singular vector pair corresponding to $\sigma_j$. The above expansion implies that the map $A \mapsto \sigma_j(A)$ is Hadamard differentiable at $A$ with derivative $H \mapsto (\boldsymbol{u}_j)^\top H \boldsymbol{v}_j$. Consequently, the functional delta method transfers the matrix limit in Theorem 3.2 to the first $r^*$ many singular values of $\tilde{A}_n$. Now in case of part (b), i.e., when the true singular value is 0 with multiplicity $(m - r^*)$, then the corresponding singular vectors are not unique because of possible multiplicity being $> 1$. Instead there is an entire singular subspace spanned by blocks $U_0, V_0$. This makes the map $A \mapsto \sigma_j(A)$ not Hadamard differentiable at $A$. Hence a linear expansion of the form (3.2) does not exist. Although the map is Hadamard directionally differentiable (cf. Definition 2 and Lemma B.4). Therefore, the perturbation must be analyzed inside the whole singular subspace which involves the singular values of the projected perturbation matrix (cf. Lemma A.15). Subsequently, we obtain the asymptotic result of part (b).

**Remark 3.1.** *Our motivation behind considering the nuclear normed penalized estimator is that the parameter matrix* $A$ *in multivariate regression model (1.1) may be rank deficient. Therefore, a natural question is whether the NNP estimator* $\tilde{A}_n$ *effectively identifies the rank asymptotically, i.e., whether* $\mathrm{rank}(\tilde{A}_n)$ *equals to* $r^*$ $(= \mathrm{rank}(A))$ *in probability or almost surely, as* $n \to \infty$*. Indeed, under the setup of Theorem 3.1 with* $\lambda_0 = 0$*,* $\tilde{A}_n \xrightarrow{P} A$*. This again implies that* $\sigma_j(\tilde{A}_n) \xrightarrow{P} \sigma_j$ *for every* $j = 1, \ldots, m$ *due to Weyl's inequality (cf. Corollary 7.3.5 of Horn and Johnson (2013)). This essentially shows that* $\mathrm{rank}(\tilde{A}_n) = r^*$ *in probability, as* $n \to \infty$*, i.e. the estimator* $\tilde{A}_n$ *is RSC. Same conclusion can be drawn also from Theorem 3.3 when* $n^{-1/2}\lambda_n$ *converges to some non-negative real number. However, notice that we can at most conclude from part (b) of 3.3 that* $\sigma_j(\tilde{A}_n) = O_p(n^{-1/2})$*, for any* $j > r^*$*. This does not guarantee that* $\mathrm{rank}(\tilde{A}_n) = r^*$ *almost surely, as* $n \to \infty$*. This has serious consequences in the Bootstrap approximation, as discussed in the next section.*

## 4. Bootstrap Approximation

In this section, we develop a Bootstrap approximation of the distribution of $\sqrt{n}(\tilde{A}_n - A)$ to bypass the intractability of its asymptotic distribution which we observed in Theorem 3.2. We consider the Perturbation Bootstrap method (hereafter referred to as PB). The Bootstrapped estimator in PB is obtained by introducing a set of random variables which are independent of the data generating process. Let that set of random variables be $\{G_1^*, \ldots, G_n^*\}$ where $G_i^*$'s are i.i.d. non-negative and non-degenerate random variables

with $\mathbf{E}(G_1^*) = \mu_{G^*}, \quad \mathrm{Var}(G_1^*) = \mu_{G^*}^2, \quad \mathbf{E}(G_1^{*3}) < \infty$. Now suppose that we would like to center our Bootstrapped estimator around $\check{A}_n$. Then we define the pseudo response matrix as $Z^*$ of order $n \times T$ where the $i$-th row of $Z^*$ is defined as

$$\boldsymbol{Z}_i^* = \boldsymbol{X}_i \check{A}_n + \left( \frac{G_i^*}{\mu_{G^*}} - 1 \right) \left( \boldsymbol{Y}_i - \boldsymbol{X}_i \check{A}_n \right).$$

Here $\boldsymbol{X}_i \in \mathbf{R}^{1\times p}$ and $\boldsymbol{Y}_i \in \mathbf{R}^{1\times T}$ denote the $i$-th rows of $X$ and $Y$, respectively. Based on these pseudo observations, the PB version of the NNP estimator at penalty $\lambda_n^* > 0$ is defined as

$$\check{A}_n^* = \arg \min_{B \in \mathbf{R}^{p\times T}} \left\{ \|Z^* - XB\|_F^2 + \lambda_n^* \|\sigma(B)\|_1 \right\}. \tag{4.1}$$

If $\check{A}_n$ is the original NNP estimator $\tilde{A}_n$ or the usual least square estimator, then Bootstrap may fail. To get an idea on why, note that

$$\sqrt{n}(\check{A}_n^* - \tilde{A}_n) = \mathrm{argmin}_{V \in \mathbf{R}^{p\times T}} \Bigg[ -2\,\mathrm{tr}\big(V^\top \bar{W}_{n,G}^*\big) + \mathrm{tr}\big(V^\top C_n V\big) + \frac{\lambda_n^*}{\sqrt{n}} \left( \frac{\|\sigma(\tilde{A}_n + n^{-1/2}V)\|_1 - \|\sigma(\tilde{A}_n)\|_1}{n^{-1/2}} \right) \Bigg],$$

where $C_n = n^{-1} X^\top X$, $\bar{W}_{n,G}^* = \frac{1}{\sqrt{n}} X^\top \bar{D}_{G^*} \tilde{E}$, $\tilde{E} = Y - X\tilde{A}_n$ and the matrix of multipliers $\bar{D}_{G^*} = \mathrm{diag}\Big( \frac{G_1^*}{\mu_{G^*}} - 1, \ldots, \frac{G_n^*}{\mu_{G^*}} - 1 \Big)$. If $\arg\min_{U \in \mathbf{R}^{p\times T}} \tilde{H}^*(U)$ is the distributional limit of $\sqrt{n}(\tilde{A}_n^* - \tilde{A}_n)$ conditional on the data, then $\tilde{H}^*(U)$ must match with the form of $H(U)$ (see Theorem 3.2) for the validity of Bootstrap. For that it is required to have $(m - r^*)$ many singular values of $\tilde{A}_n$ to be exactly 0 with probability 1, for the validity of Bootstrap. However, from Theorem 3.3, we can only conclude that for any $j \in \{r^* + 1, \ldots, m\}$,

$$\sigma_j(\tilde{A}_n) = O_p(n^{-1/2}),$$

which does not guarantee that $P(\sigma_j(\tilde{A}_n) = 0) = 1$ for any $j \in \{r^* + 1, \ldots, m\}$. It may happen that some of those singular values remain strictly positive or negative, even for large $n$, implying that '$\mathrm{rank}(\tilde{A}_n) = r^*$ with probability 1' is not guaranteed. A similar phenomenon is true while defining PB method for Lasso in GLM (see Choudhury and Das (2026)). To circumvent the issue, we bring the idea of thresholding similar to Choudhury and Das (2026) and consider a suitable thresholded version of $\tilde{A}_n$ as $\check{A}_n$, as defined in the next paragraph. We show in Proposition 4.1 that the thresholded estimator is RSC almost surely. Thus, we can conclude that the notion of RSC in almost sure sense is an extension of the notion of variable selection consistency (or VSC) when we move from Lasso in (1.1) with $T = 1$ to the NNP estimator in (1.1) with general $T > 1$. Note that '$\check{A}_n$ is RSC in almost sure sense' is equivalent to

'$P(\sigma_j(\check{A}_n) > 0$ infinitely often$) = 0$, for any $j$ such that $\sigma_j = 0$'.

Keeping this in mind, for left and right singular vector pair $(\tilde{\boldsymbol{u}}_j^{(n)},\ \tilde{\boldsymbol{v}}_j^{(n)})$ of $\tilde{A}_n$ corresponding to $\sigma_j(\tilde{A}_n)$, we may define the thresholded version of $\tilde{A}_n$ as

$$\tilde{A}_n^{(\mathrm{thr})} := \sum_{j=1}^{m} \Big[ \sigma_j(\tilde{A}_n)\, \mathbf{1}\{\sigma_j(\tilde{A}_n) > a_n\} \Big]\, \tilde{\boldsymbol{u}}_j^{(n)} (\tilde{\boldsymbol{v}}_j^{(n)})^\top, \tag{4.2}$$

i.e., $\tilde{A}_n^{(\mathrm{thr})}$ is obtained from $\tilde{A}_n$ by singular value thresholding based on a deterministic sequence $\{a_n\}_{n\geq 1}$. The sequence $\{a_n\}_{n\geq 1}$ satisfies

$$a_n + \frac{n^{-1/2}\log n}{a_n} \to 0. \tag{4.3}$$

A typical choice of the sequence is $\{n^{-1/3}\}_{n\geq 1}$ which is also what we use in the simulations presented in the section 5. The reason behind the general choice of $\{a_n\}_{n\geq 1}$ of (4.2) is the order of fluctuations for $\sigma_j(\tilde{A}_n)$ around true $\sigma_j$ in almost sure sense. Lemma A.8 and Weyl's inequality together imply that $|\sigma_j(\tilde{A}_n) - \sigma_j| = o(n^{-1/2}\log n)$ almost surely, for any $j$. Thus for $j = 1, \ldots, r^*$, i.e., when $\sigma_j$ is positive then we need $a_n \to 0$ to keep $\sigma_j(\tilde{A}_n)$ unaffected. On the contrary, when $\sigma_j = 0$ for all $j > r^*$, one needs $a_n$ to grow faster than the order $n^{-1/2}\log n$ such that the thresholding forces $\sigma_j(\tilde{A}_n)$ to be exactly 0 eventually. Clearly, the choice of the thresholding is analogous to what we use to find a VSC estimator from the the Lasso estimator (see Chatterjee and Lahiri (2011); Choudhury and Das (2026)). It is important to point out that the singular value thresholding has huge importance in the matrix completion problems. A seminal paper in that area is the paper by Chatterjee (2015) where a mechanism of matrix completion is proposed based on thresholding the singular values. In the next proposition, we establish that $\tilde{A}_n^{(\mathrm{thr})}$ is RSC in almost sure sense.

**Proposition 4.1.** *Suppose that all the regularity conditions of Theorem 3.3 are true. Consider the estimator $\tilde{A}_n^{(\mathrm{thr})}$ defined in (4.2). Then,*

$$\mathbf{P}\Big(\operatorname{rank}\big(\tilde{A}_n^{(\mathrm{thr})}\big) \neq r^*, \textit{ infinitely often}\Big) = 0.$$

Proof of Proposition 4.1 is provided in section C.4. Proposition 4.1 together with part (a) of Theorem 3.3 essentially establish that the non-zero singular values of $\tilde{A}_n^{(\mathrm{thr})}$ are close to that of $\tilde{A}_n$ and $(m - r^*)$ many singular values of $\tilde{A}_n^{(\mathrm{thr})}$ are essentially 0 for large enough $n$. This is exactly what we need for the term $\check{A}_n$ in defining the Bootstrap pivotal quantity. We establish in the next theorem that the distribution of $\sqrt{n}\big(\tilde{A}_n^{*(mod)} - \tilde{A}_n^{(thr)}\big)$ conditional on the data can be used to approximate the distribution of $\sqrt{n}\big(\tilde{A}_n - A\big)$, where from now on we denote $\check{A}_n^*$ by $\tilde{A}_n^{*(mod)}$ whenever $\check{A}_n = \tilde{A}_n^{(thr)}$. Let $\hat{F}_n^{(\mathrm{thr})}$ denote the conditional distribution of $\sqrt{n}\,\big(\tilde{A}_n^{*(mod)} - \tilde{A}_n^{(\mathrm{thr})}\big)$ given the response $Y$ and $\xrightarrow{d_*}$ denote the distribution convergence in the Bootstrap regime.

**Theorem 4.1.** *Suppose the assumptions (C.1)–(C.4) hold true. Assume that $n^{-1/2}\lambda_n^* \to \lambda_0$, as $n \to \infty$, i.e., the limit of $n^{-1/2}\lambda_n^*$ is same as $n^{-1/2}\lambda_n$. Then we have*

$$\mathbf{P}\Big(\hat{F}_n^{(\mathrm{thr})}(\cdot) \xrightarrow{d_*} \arg\min_{U\in\mathbf{R}^{p\times T}} H(U)\Big) = 1,$$

*where $H(\cdot)$ is as defined in Theorem 3.2.*

The proof is provided in section C.5. Theorem 4.1 together with Theorem 3.2 imply that the distribution of original pivot $\sqrt{n}\big(\tilde{A}_n - A\big)$ and the conditional distribution of $\sqrt{n}\,\big(\tilde{A}_n^{*(mod)} - \tilde{A}_n^{(\mathrm{thr})}\big)$ are close for sufficiently large $n$. Thus, our PB method can be used to draw inference for the parameter $A$. For example, for any $\alpha \in (0,1)$, if $c_{n,1-\alpha}$ denotes the $(1-\alpha)$ conditional quantile of $\|\sqrt{n}\,\big(\tilde{A}_n^{*(mod)} - \tilde{A}_n^{(\mathrm{thr})}\big)\|$ given the data, then the PB percentile confidence region

$$\Big\{A :\ \sqrt{n}\,\|\tilde{A}_n - A\| \leq c_{n,1-\alpha}\Big\}$$

achieve nominal coverage of $(1-\alpha)$ asymptotically. Additionally, PB can be used even if we do not have any prior knowledge of the rank of the true parameter $A$. This is because the thresholded estimator $\tilde{A}_n^{(thr)}$ is capable of identifying the true rank in almost sure sense. thresholding ensures almost sure stabilization of the model dimension. As an application of Theorem 4.1, we can establish the validity of PB in approximating the distribution of the singular values of $\tilde{A}_n$. We present the result a the following corollary.

**Corollary 4.1** (Bootstrap $\sqrt{n}-$consistency of singular values)**.** *Consider the setup of Theorem 3.2. Then we have*

*(a) for each* $j = 1, \ldots, r^*$,

$$\mathbf{P}\Big(\sqrt{n}\big(\sigma_j(\tilde{A}_n^{*(\mathrm{mod})}) - \sigma_j(\tilde{A}_n^{(\mathrm{thr})})\big) \xrightarrow{d^*} (\boldsymbol{u}_j)^\top U^* \boldsymbol{v}_j\Big) = 1.$$

*(b) for* $j = r^* + 1, \ldots, m$,

$$\mathbf{P}\Big(\sqrt{n}\,\sigma_j(\tilde{A}_n^{*(\mathrm{mod})}) \xrightarrow{d^*} \sigma_{j-r^*}\big(U_0^\top U^* V_0\big)\Big) = 1.$$

This result admits two complementary interpretations. First, conditionally, the singular values of $\tilde{A}_n^{*(\mathrm{mod})}$ reproduce the same asymptotic distribution as those of the NNP estimator $\tilde{A}_n$. Second the Bootstrap singular values are consistent for the singular values of $\tilde{A}_n^{(\mathrm{thr})}$. This is coherent since the thresholding ensures that the two targets $\tilde{A}_n$ and $\tilde{A}_n^{(\mathrm{thr})}$ are asymptotically equivalent in the sense that they asymptotically have the same singular value decomposition. To prove the corollary 4.1, we work on the event where the thresholded estimator $\tilde{A}_n^{(\mathrm{thr})}$ is RSC almost surely for all sufficiently large $n$. On this event, $\tilde{A}_n^{(\mathrm{thr})}$ has the same SVD block structure as $A$, precisely, it has rank $r^*$, a positive spectral gap $\sigma_{r^*} > 0$ for the singular values and the same singular subspaces $U_0, V_0$. Conditionally on the data, $\tilde{A}_n^{*(\mathrm{mod})}$ is a $n^{-1/2}$-perturbation of a sequence with stable singular structure. Hence, the first-order SVD perturbation expansion (cf. Stewart (1990)) and Hadamard differentiability of singular values apply conditionally, yielding the stated limits.

## 5. Simulation Study

In this section, we study the finite-sample performance of the NNP estimator and the proposed Perturbation Bootstrap approximation in the multivariate regression model (1.1) based on empirical coverage probabilities. To be precise through these simulation settings, we verify our proposed PB method by finding empirical coverages of 90% one-sided and two-sided confidence intervals for functionals of $A$. In particular, we observe confidence intervals for:

(i) first component $A_{1,1}$ and last component $A_{p,T}$ of the parameter matrix,
(ii) $\max_{1\le j\le T}\sum_{i=1}^{p}|A_{i,j}|$ and $\min_{1\le j\le T}\sum_{i=1}^{p}|A_{i,j}|$
(iii) $\max_{1\le j\le T}\|A_{\cdot,j}\|_2$ and $\min_{1\le j\le T}\|A_{\cdot,j}\|_2$
(iv) largest and smallest non-zero singular values ($\sigma_1$ and $\sigma_{r^*}$) of $A$,
(v) last zero singular value $\sigma_{min(p,T)}$ of $A$.

The functional delta method applies to all these real-valued functionals of the matrix argument. All the maps are Hadamard directionally differentiable functions (cf. Lemma B.1, B.2, B.3 ). Therefore we are allowed to utilize generalized functional delta method corresponding to these maps and their Bootstap counterparts (cf. remark B.1). Throughout we used the Bootstrap percentile method. We try to capture finite sample performance under

the following set-up:

$$(p, T, r^*) \in \Big\{(20, 10, 2);\ (40, 15, 4);\ (60, 20, 6)\Big\},$$

over varying sample sizes $n \in \{100,\ 300,\ 500,\ 800\}$. We generate the true coefficient matrix $A$ once and keep it fixed across all Monte Carlo replications and sample sizes, for a fixed set-up $(p, T, r^*)$. More precisely, the true parameter matrix $A$ is constructed in a fixed low-rank form as follows. We first generate random Gaussian matrices of dimensions $p\times r^*$ and $T \times r^*$, and orthonormalize them via `QR` decomposition to obtain left singular matrix $U_{r^*} \in \mathbf{R}^{p\times r^*}$ and right singular matrix $V_{r^*} \in \mathbf{R}^{T\times r^*}$. A fixed diagonal matrix $D_{r^*} = \mathrm{diag}(5, 3)$ for $r^* = 2$, $D_{r^*} = \mathrm{diag}(5, 3, 2, 1)$ for $r^* = 4$ and $D_{r^*} = \mathrm{diag}(9, 6, 5, 3, 2, 1)$ for $r^* = 6$, are then specified to control the signal strength along each latent component. The true parameter matrix is defined as: $A = U_{r^*} D_{r^*} V_{r^*}^\top$. This construction is similar to Bunea et al. (2011); Chen et al. (2013). To recall this matrix for remaining simulation set-ups over $n$, we utilize `saveRDS` in `R`. The design matrix $X \in \mathbf{R}^{n\times p}$ is generated once from a multivariate normal distribution with mean zero and a structured covariance matrix and kept fixed for remaining simulation set-ups. Specifically, $\boldsymbol{X}_i \sim \mathcal{N}_p(0, \Sigma),\quad i = 1, \ldots, n$, where the covariance matrix $\Sigma$ is defined as: $\Sigma_{jk} = 0.3^{|j-k|},\ 1 \le j, k \le p$. We consider generating $500$ Monte-Carlo (MC) replicates to iterate the entire data and PB quantities for valid inference. Now for each MC replicates, the error matrix $E^{n\times T}$ is generated such that its each entry is iid $\mathcal{N}(0, 1)$. Hence the response matrix $Y$ is generated as per definition (1.1). For the thresholding, we consider $a_n = n^{-1/3}$.

**On finding the NNP estimator:** In our simulation, $A$ is explicitly constructed to be low-rank. Hence, this penalty is correctly aligned with the true data-generating mechanism. We write the objective function of (1.3) as a composite function:

$$f(A) = \|Y - XA\|_F^2, \qquad g(A) = \lambda_n \|\sigma(A)\|_1,$$

The function $f(A)$ is a quadratic form. So its gradient $\nabla f(A) = 2X^\top(XA - Y)$. is linear. Hence $\nabla f$ is Lipschitz continuous with constant: $L = 2\|X^\top X\|_2 = 2\|X\|_2^2$. Hence $f$ is smooth with Lipschitz gradient, $g$ is convex but non-differentiable. Thus the problem fits the standard composite optimization (FISTA) framework of Beck and Teboulle (2009). From Beck and Teboulle (2009), we recall the definition of proximal operator at $\boldsymbol{x}$ with step size $t$:

$$\mathrm{prox}_t(g)(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left\{ g(\boldsymbol{u}) + \frac{1}{2t}\|\boldsymbol{u} - \boldsymbol{x}\|^2 \right\}. \tag{5.1}$$

Applying FISTA on matrix valued map with constant step size $t_k = \frac{1}{L}$, we get:

$$A_{k+1} = \mathrm{prox}_{t_k = \frac{1}{L}}(g)\left(Z_k - \frac{1}{L}\nabla f(Z_k)\right).$$

Define: $W_k := Z_k - \frac{1}{L}\nabla f(Z_k)$. Now following the definition (5.1) we can write,

$$A_{k+1} = \arg\min_A \left\{ \lambda_n \|\sigma(A)\|_1 + \frac{L}{2}\|A - W_k\|_F^2 \right\} = \arg\min_A \left\{ \frac{\lambda_n}{L}\|\sigma(A)\|_1 + \frac{1}{2}\|A - W_k\|_F^2 \right\}$$

Let the singular value decomposition of $W_k$ be $W_k = U\,\mathrm{diag}(\sigma_i)\,V^\top$. By Theorem 2.1 of Cai et al. (2010),

$$A_{k+1} = U\,\mathrm{diag}\big((\sigma_i(W_k) - \lambda_n/L)_+\big)\,V^\top. \tag{5.2}$$

Finally, acceleration is introduced via:

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \quad Z_{k+1} = A_{k+1} + \frac{t_k - 1}{t_{k+1}}(A_{k+1} - A_k). \tag{5.3}$$

Following Beck and Teboulle (2009), the algorithm is initialized as: $A_0 \in \mathbf{R}^{p\times T}, \quad Z_1 = A_0, \quad t_1 = 1$. Under Lipschitz continuity of $\nabla f$, the FISTA iterates satisfy (cf. Theorem 3.1 of Beck and Teboulle (2009)) $F(A_k) - F(A) = O\left(\frac{1}{k^2}\right)$, and convergence is guaranteed asymptotically. In our simulation practice, the iterations are terminated when the relative change between successive iterates is sufficiently small:

$$\frac{\|A_{k+1} - A_k\|_F}{\|A_k\|_F + \varepsilon} < \mathrm{tol},$$

or when a maximum number of iterations $k_{\max}$ is reached. The parameters tolerance $\mathrm{tol}$ and maximum number of iterations $k_{\max}$ control termination of the iterative optimization procedure. This criterion measures the relative change between successive iterates. The use of a relative (rather than absolute) criterion ensures scale invariance, making the stopping rule robust across different magnitudes of $A_k$. The chosen stopping criterion is motivated by the fixed-point characterization of proximal gradient methods: $A = \mathrm{prox}_\tau(g)\big(A - \tau\nabla f(A)\big)$. Near the optimum, successive iterates satisfy $A_{k+1} \approx A_k$, so the norm $\|A_{k+1} - A_k\|_F$ becomes small. This fixed-point characterization of proximal gradient methods is standard in convex optimization; see, e.g., Nesterov (2004); Beck and Teboulle (2009); Beck (2017). In our simulation purpose, we have considered $\varepsilon = 10^{-8}$, $\mathrm{tol} = 10^{-7}$ and $k_{max}$ to be $150$ for appropriate numerical accuracy. Now we summarize the entire optimization step in the algorithm 1.

**Algorithm 1** FISTA for finding NNP estimator

1: **Input:** Data $(X, Y)$, penalty $\lambda_n$, Lipschitz constant $L = 2\|X\|_2^2$, $\varepsilon$, tolerance $\mathrm{tol}$, maximum iterations $k_{\max}$
2: **Initialization:** $A_0 \in \mathbf{R}^{p\times T}, \quad Z_1 = A_0, \quad t_1 = 1$
3: **for** $k = 1, 2, \ldots, k_{\max}$ **do**
4: **Gradient computation:** $\nabla f(Z_k) = 2X^\top(XZ_k - Y)$
5: **Gradient step:** $W_k = Z_k - \frac{1}{L}\nabla f(Z_k)$
6: **Singular Value Decomposition:** $W_k = U\,\mathrm{diag}(\sigma_i(W_k))\,V^\top$
7: **Proximal step (SVT):** $A_k = U\,\mathrm{diag}\big((\sigma_i(W_k) - \lambda_n/L)_+\big)\,V^\top$
8: **Momentum update:**

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \quad Z_{k+1} = A_k + \frac{t_k - 1}{t_{k+1}}(A_k - A_{k-1})$$

9: **Stopping rule:**

$$\frac{\|A_k - A_{k-1}\|_F}{\|A_{k-1}\|_F + \varepsilon} < \mathrm{tol} \quad \text{(terminate if satisfied)}$$

10: **end for**
11: **Output:** $\tilde{A}_n = A_k$

**Selection of the Regularization Parameter:** The performance of the NNP estimator $\tilde{A}_n$ critically depends on the choice of the tuning parameter $\lambda_n$. In practice, choice of such $\lambda_n$ is usually made data-driven for various types of regularization problems. Among the existing methods, cross-validation is the most popular. We employ $K$-fold cross-validation

to estimate the prediction error associated with each $\lambda$. Specifically, define the cross-validated prediction error for NNP at penalty $\lambda$ as,

$$\mathrm{CV}(\lambda) = \frac{1}{K}\sum_{k=1}^{K} \underbrace{\frac{1}{|I_k|}\sum_{i\in I_k}\left\|\boldsymbol{Y}_i - \boldsymbol{X}_i\tilde{A}_{n,-k}(\lambda)\right\|_2^2}_{:=CV_k(\lambda)}, \tag{5.4}$$

where we define; $\tilde{A}_{n,-k}(\lambda) = \mathrm{Argmin}_{A\in\mathbf{R}^{p\times T}}\left\{\sum_{i\notin I_k}\left\|\boldsymbol{Y}_i - \boldsymbol{X}_i A\right\|_2^2 + \lambda\|\boldsymbol{\sigma}(A)\|_1\right\}$, and consider $[I_k : k\in\{1,..,K\}]$ to be a partition of the set $\{1,..,n\}$. Now for the purpose of simulation, we consider a sequence of candidate values $\{\lambda_m\}_{m=1}^M$ defined on a logarithmic scale: $\lambda_m \in \Big[0.001\,\lambda_{\max},\ 2\,\lambda_{\max}\Big]$, where $\lambda_{\max} = \|X^\top Y\|_2$. This choice is motivated by the fact that $\lambda_{\max}$ is the smallest value for which the trivial solution $A = 0$ is optimal (see Friedman et al. (2010)). The logarithmic spacing ensures adequate exploration of both strong and weak regularization regimes. For each $\lambda_m$, the estimator $\tilde{A}_n(\lambda_m)$ is computed using the FISTA algorithm. Let $\hat{\lambda}_{\min}$ denote the minimizer of the cross-validation error: $\widehat{\lambda}_{\min} = \arg\min_{\lambda_m}\mathrm{CV}(\lambda_m)$. Define $\mathrm{SE}(\lambda_m) = \mathrm{sd}\big(\{\mathrm{CV}_k(\lambda_m)\}_{k=1}^K\big)$, which is the standard deviation among $K$ many training CV errors. Cross-validation provides an estimate of prediction risk, while the one-standard-error rule selects the simplest model whose estimated risk is within one standard error of the minimum, thereby favoring more regularized (lower-rank) solutions with comparable predictive performance (cf. Hastie et al. (2015)). Then the selected tuning parameter is given in our set-up as:

$$\lambda_{\mathrm{opt}} = \max\left\{\lambda_m : \mathrm{CV}(\lambda_m) \le \mathrm{CV}(\widehat{\lambda}_{\min}) + \mathrm{SE}(\widehat{\lambda}_{\min})\right\}. \tag{5.5}$$

This rule is justified by the fact that differences within one standard error are not statistically significant. Choosing the largest among them favors a more regularized and parsimonious model, which is particularly important in nuclear norm penalization to avoid rank overestimation. The NNP estimator used in the simulation is therefore: $\tilde{A}_n = \tilde{A}_n(\lambda_{\mathrm{opt}})$. The entire procedure is implemented in `R` through the functions `nnp_lambda_seq`, `cv_nnp` and `nnp_solver_path`. The use of warm starts and FISTA ensures computational efficiency even when solving the problem repeatedly across folds and Monte Carlo replications.

**Bootstrap percentile intervals:** Let $T : \mathbf{R}^{p\times T} \to \mathbf{R}$ be any scalar functional of the underlying matrices we have considered so far. For $B$ many PB replicates, we define,

$$\Delta_b^* = \sqrt{n}\Big(T(\tilde{A}_{n,b}^{*(mod)}) - T(\tilde{A}_n^{(thr)})\Big), \qquad b = 1,\ldots,B.$$

For given level of significance $\alpha$, we denote the $\alpha$-th quantile for the PB distribution of $\{\Delta_b^*\}_{b=1}^B$ by $q_\alpha$. Therefore $100(1-\alpha)\%$ two-sided confidence interval for $T(A)$ is given by:

$$\left[T(\tilde{A}_n) - \frac{q_{1-\alpha/2}}{\sqrt{n}},\ T(\tilde{A}_n) - \frac{q_{\alpha/2}}{\sqrt{n}}\right].$$

Similarly we can define left and right sided confidence intervals. Tables 1–3 summarize the empirical coverages and average widths of the Bootstrap confidence intervals for nine representative functionals of the target matrix $A$ under the three considered combinations of dimension and rank, $(p, T, r^*)$. Several clear patterns emerge and remain consistent across all settings. For the smallest sample size ($n = 100$), the empirical coverages are generally slightly below the nominal $90\%$ level, reflecting the finite-sample difficulty of accurately approximating the sampling distribution of the estimators. Nevertheless, the

coverages remain above $0.80$ for all functionals, indicating that the proposed procedure already provides a reasonable degree of uncertainty quantification even in relatively small samples.

TABLE 1. Empirical coverages for nominal $90\%$ two-sided and right sided confidence intervals under $(p, T, r^*) = (20, 10, 2)$ along with average widths inside parentheses.

| Functional | Coverage | Sample size $n$ | | | |
|---|---|---|---|---|---|
| | | 100 | 300 | 500 | 800 |
| $A_{1,1}$ | TS | 0.826 (1.368) | 0.868 (0.942) | 0.886 (0.821) | 0.902 (0.752) |
| | RS | 0.832 | 0.862 | 0.878 | 0.896 |
| $A_{p,T}$ | TS | 0.848 (1.125) | 0.870 (0.894) | 0.892 (0.715) | 0.910 (0.623) |
| | RS | 0.820 | 0.856 | 0.874 | 0.898 |
| $\max_{1\le j\le T}\sum_{i=1}^{p}\lvert A_{ij}\rvert$ | TS | 0.838 (1.129) | 0.866 (0.816) | 0.882 (0.705) | 0.904 (0.613) |
| | RS | 0.840 | 0.872 | 0.890 | 0.908 |
| $\min_{1\le j\le T}\sum_{i=1}^{p}\lvert A_{ij}\rvert$ | TS | 0.846 (1.301) | 0.878 (0.984) | 0.886 (0.821) | 0.898 (0.721) |
| | RS | 0.836 | 0.860 | 0.880 | 0.894 |
| $\max_{1\le j\le T}\lVert A_{\cdot,j}\rVert_2$ | TS | 0.820 (1.102) | 0.858 (0.897) | 0.884 (0.703) | 0.896 (0.653) |
| | RS | 0.834 | 0.862 | 0.882 | 0.890 |
| $\min_{1\le j\le T}\lVert A_{\cdot,j}\rVert_2$ | TS | 0.852 (1.121) | 0.874 (0.853) | 0.892 (0.725) | 0.912 (0.636) |
| | RS | 0.844 | 0.870 | 0.888 | 0.900 |
| $\sigma_1$ | TS | 0.830 (1.421) | 0.866 (1.012) | 0.882 (0.893) | 0.894 (0.678) |
| | RS | 0.846 | 0.878 | 0.892 | 0.908 |
| $\sigma_{r^*}$ | TS | 0.822 (1.120) | 0.860 (0.893) | 0.886 (0.652) | 0.892 (0.503) |
| | RS | 0.840 | 0.874 | 0.890 | 0.906 |
| $\sigma_{\min(p,T)}$ | TS | 0.850 (0.931) | 0.870 (0.802) | 0.896 (0.671) | 0.910 (0.515) |
| | RS | 0.852 | 0.876 | 0.888 | 0.898 |

*Notes:* TS = two-sided percentile interval; RS = right-sided percentile interval.

TABLE 2. Empirical coverages for nominal $90\%$ two-sided and right sided confidence intervals under $(p, T, r^*) = (40, 15, 4)$ along with average widths inside parentheses.

| Functional | Coverage | Sample size $n$ | | | |
|---|---|---|---|---|---|
| | | 100 | 300 | 500 | 800 |
| $A_{1,1}$ | TS | 0.812 (1.412) | 0.854 (1.006) | 0.878 (0.842) | 0.904 (0.736) |
| | RS | 0.826 | 0.858 | 0.884 | 0.902 |
| $A_{p,T}$ | TS | 0.836 (1.198) | 0.868 (0.922) | 0.888 (0.768) | 0.912 (0.642) |
| | RS | 0.824 | 0.860 | 0.882 | 0.904 |
| $\max_{1\le j\le T}\sum_{i=1}^{p}\lvert A_{ij}\rvert$ | TS | 0.828 (1.186) | 0.862 (0.874) | 0.884 (0.732) | 0.906 (0.618) |
| | RS | 0.838 | 0.870 | 0.890 | 0.910 |
| $\min_{1\le j\le T}\sum_{i=1}^{p}\lvert A_{ij}\rvert$ | TS | 0.842 (1.334) | 0.876 (1.012) | 0.890 (0.836) | 0.908 (0.704) |
| | RS | 0.830 | 0.866 | 0.886 | 0.900 |
| $\max_{1\le j\le T}\lVert A_{\cdot,j}\rVert_2$ | TS | 0.818 (1.154) | 0.856 (0.918) | 0.882 (0.744) | 0.898 (0.662) |
| | RS | 0.832 | 0.864 | 0.886 | 0.896 |
| $\min_{1\le j\le T}\lVert A_{\cdot,j}\rVert_2$ | TS | 0.848 (1.172) | 0.872 (0.892) | 0.894 (0.752) | 0.914 (0.644) |
| | RS | 0.840 | 0.874 | 0.892 | 0.904 |
| $\sigma_1$ | TS | 0.824 (1.468) | 0.860 (1.058) | 0.886 (0.918) | 0.902 (0.704) |
| | RS | 0.844 | 0.880 | 0.894 | 0.910 |
| $\sigma_{r^*}$ | TS | 0.816 (1.162) | 0.858 (0.916) | 0.882 (0.688) | 0.896 (0.522) |
| | RS | 0.836 | 0.870 | 0.888 | 0.906 |
| $\sigma_{\min(p,T)}$ | TS | 0.854 (0.968) | 0.874 (0.828) | 0.898 (0.694) | 0.912 (0.532) |
| | RS | 0.850 | 0.878 | 0.892 | 0.900 |

*Notes:* TS = two-sided percentile interval; RS = right-sided percentile interval.

As the sample size increases, the empirical coverages move steadily closer to the target level of $0.90$, demonstrating the improved accuracy of the bootstrap approximation in larger samples. This convergence toward the nominal level is observed uniformly across the different matrix dimensions, signal ranks, and functional classes. At the same time, the average widths of the confidence intervals decrease monotonically with $n$, reflecting the increasing precision of the estimators as more data become available. Importantly, this reduction in interval width does not come at the expense of coverage accuracy, as the empirical coverages remain close to the desired level for moderate and large sample sizes.

Another notable feature is the stability of the results across the nine functionals considered. Whether the functional depends on individual matrix entries, singular values, or more global characteristics of the matrix, the PB intervals exhibit similar coverage behavior and comparable rates of width reduction.

TABLE 3. Empirical coverages for nominal $90\%$ two-sided and right sided confidence intervals under $(p, T, r^*) = (60, 20, 6)$ along with average widths inside parentheses.

| Functional | Coverage | Sample size $n$ | | | |
|---|---|---|---|---|---|
| | | 100 | 300 | 500 | 800 |
| $A_{1,1}$ | TS | 0.836 (1.287) | 0.872 (0.963) | 0.888 (0.811) | 0.902 (0.702) |
| | RS | 0.842 | 0.876 | 0.892 | 0.906 |
| $A_{p,T}$ | TS | 0.820 (1.244) | 0.862 (0.934) | 0.880 (0.776) | 0.892 (0.658) |
| | RS | 0.834 | 0.862 | 0.884 | 0.904 |
| $\max_{1\le j\le T}\sum_{i=1}^{p}\lvert A_{ij}\rvert$ | TS | 0.848 (1.206) | 0.878 (0.884) | 0.888 (0.731) | 0.902 (0.622) |
| | RS | 0.850 | 0.870 | 0.886 | 0.896 |
| $\min_{1\le j\le T}\sum_{i=1}^{p}\lvert A_{ij}\rvert$ | TS | 0.826 (1.352) | 0.868 (1.021) | 0.884 (0.858) | 0.904 (0.726) |
| | RS | 0.836 | 0.872 | 0.890 | 0.908 |
| $\max_{1\le j\le T}\lVert A_{\cdot,j}\rVert_2$ | TS | 0.832 (1.178) | 0.868 (0.946) | 0.880 (0.768) | 0.894 (0.671) |
| | RS | 0.844 | 0.878 | 0.892 | 0.900 |
| $\min_{1\le j\le T}\lVert A_{\cdot,j}\rVert_2$ | TS | 0.854 (1.196) | 0.872 (0.902) | 0.890 (0.744) | 0.906 (0.638) |
| | RS | 0.848 | 0.870 | 0.884 | 0.896 |
| $\sigma_1$ | TS | 0.838 (1.396) | 0.874 (1.034) | 0.892 (0.902) | 0.910 (0.718) |
| | RS | 0.848 | 0.864 | 0.890 | 0.904 |
| $\sigma_{r^*}$ | TS | 0.824 (1.138) | 0.862 (0.902) | 0.880 (0.692) | 0.898 (0.548) |
| | RS | 0.838 | 0.872 | 0.888 | 0.904 |
| $\sigma_{\min(p,T)}$ | TS | 0.852 (0.986) | 0.876 (0.846) | 0.894 (0.708) | 0.906 (0.564) |
| | RS | 0.844 | 0.870 | 0.886 | 0.898 |

*Notes:* TS = two-sided percentile interval; RS = right-sided percentile interval.

The overall pattern is also largely unaffected by changes in the ambient dimension or the underlying rank structure. Taken together, these findings provide strong empirical evidence that the proposed PB method delivers reliable confidence intervals with accurate coverage and increasing efficiency as the sample size grows, thereby supporting the theoretical validity of the procedure.

## APPENDIX

In this section, we address Hadamard differentiability of matrix functionals, provide proofs of requisite lemmas and main results.

## APPENDIX A. REQUISITE LEMMAS

**Lemma A.1** (Convexity of the nuclear norm)**.** *The function*

$$A \;\mapsto\; \|\sigma(A)\|_1 \;=\; \|A\|_*, \quad \textit{is convex on} \quad \mathbf{R}^{p\times T}.$$

Proof of Lemma A.1 The nuclear norm admits the variational representation (cf. Watson (1992))

$$\|A\|_* = \sup_{\|B\|\le 1} \langle A, B\rangle = \sup_{\|B\|\le 1} \operatorname{tr}(A^\top B),$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm. Let $A_1, A_2 \in \mathbf{R}^{p\times T}$ and $t \in [0,1]$. Then

$$\begin{aligned}\|tA_1 + (1-t)A_2\|_* &= \sup_{\|B\|\le 1} \langle tA_1 + (1-t)A_2, B\rangle \\ &\le t \sup_{\|B\|\le 1} \langle A_1, B\rangle + (1-t) \sup_{\|B\|\le 1} \langle A_2, B\rangle \\ &= t\|A_1\|_* + (1-t)\|A_2\|_*.\end{aligned}$$

Hence $\|\cdot\|_*$ is convex, and since $\|\sigma(A)\|_1 = \|A\|_*$, the result follows. ∎

**Lemma A.2** (Hoffman–Wielandt Inequality)**.** *Let $X, Y \in \mathbf{R}^{p\times T}$ and let $\sigma_1(X) \geq \cdots \geq \sigma_m(X)$ and $\sigma_1(Y) \geq \cdots \geq \sigma_m(Y)$ denote their ordered singular values, where $m = \min\{p, T\}$. Then*

$$\sum_{j=1}^{m} \big(\sigma_j(X) - \sigma_j(Y)\big)^2 \;\leq\; \|X - Y\|_F^2.$$

Proof of Lemma A.2 This lemma is proved as Corollary 7.3.5 in Horn and Johnson (2013). ■

**Lemma A.3** (Convergence of matrix functionals)**.** *Assume $\operatorname{rank}(A^{m\times n}) = r$ and $\{B_k : k \geq 1\}$ is a sequence of $n \times n$ matrices such that $B_k \to B$ as $k \to \infty$. Then*

$$\operatorname{tr}(AB_k) \to \operatorname{tr}(AB).$$

Proof of Lemma A.3: Let the singular value decomposition of $A$ be $A = U\Sigma V^\top$, where $\Sigma$ is a diagonal matrix with positive singular values, and $U$ and $V$ have orthonormal columns. Then

$$|\operatorname{tr}(AB_k - AB)| = |\operatorname{tr}(U\Sigma V^\top(B_k - B))| = |\operatorname{tr}(\Sigma V^\top(B_k - B)U)| = |\operatorname{tr}(\Sigma Q_k)|,$$

where we define $Q_k := V^\top(B_k - B)U$. Now,

$$|\operatorname{tr}(\Sigma Q_k)| = \left|\sum_i \sigma_i (Q_k)_{ii}\right| \leq \sum_i |\sigma_i|\,|(Q_k)_{ii}| \leq \Big(\max_i |(Q_k)_{ii}|\Big) \sum_i \sigma_i.$$

Clearly, $|(Q_k)_{ii}| \leq \|Q_k\|_2$, where $\|\cdot\|_2$ denotes the spectral norm. Therefore,

$$|\operatorname{tr}(\Sigma Q_k)| \leq \|Q_k\|_2 \sum_i \sigma_i.$$

Since $B_k \to B$ implies $\|B_k - B\|_2 \to 0$, we have

$$\|Q_k\|_2 = \|V^\top(B_k - B)U\|_2 \leq \|V^\top\|_2\,\|B_k - B\|_2\,\|U\|_2 = \|B_k - B\|_2 \to 0.$$

Thus, $|\operatorname{tr}(AB_k) - \operatorname{tr}(AB)| \to 0$, which completes the proof. ■

**Lemma A.4** (From pointwise to uniform)**.** *Let $C \subset \mathbf{R}^{p\times T}$ be an open convex set, and let $\{f_n\}_{n\geq 1}$ be a sequence of convex functions $f_n : C \to \mathbf{R}$. Assume that for a dense subset $C_0 \subset C$, the pointwise limit $f(A) := \lim_{n\to\infty} f_n(A)$ exists for every $A \in C_0$. Then the following conclusions hold:*

*(i) The sequence $\{f_n\}_{n\geq 1}$ converges pointwise on all of $C$; that is, for every $A \in C$, the limit*

$$f(A) := \lim_{n\to\infty} f_n(A)$$

*exists (as a finite real number).*

*(ii) The limit function $f : C \to \mathbf{R}$ is convex.*

*(iii) The convergence is uniform on every compact subset $K \subset C$, i.e.*

$$\sup_{A\in K} |f_n(A) - f(A)| \;\longrightarrow\; 0, \qquad \text{as } n \to \infty.$$

Proof of Lemma A.4 The proof follows from Theorem 10.8 of Rockafellar (1997). ■

**Lemma A.5** (Strong consistency of minimizers of convex maps)**.** *Let $C \subset \mathbf{R}^{p\times T}$ be an open convex set and $f_n : C \to \mathbf{R}$, $n \geq 1$, is a sequence of random convex maps, and $B_n$ is a minimizer of $f_n$. Also assume $f_n(A) \to f(A)$ almost surely (or in probability) for all*

*$A \in \mathbf{R}^{p\times T}$, where $f$ is a random convex map. Assume $f$ has a unique minimizer at $B$ almost surely. Then*

$$\mathbf{P}(\lim_{n\to\infty} \|B_n - B\|_F = 0) = 1 \ \textit{(or } \|B_n - B\|_F \xrightarrow{P} 0 \textit{ as } n \to \infty).$$

Proof of Lemma A.5 We only prove the result for almost sure case. The probability convergence can be concluded in similar fashion. Fix any $\delta > 0$ and define

$$\Delta_n(\delta) = \sup_{\|A-B\|_F \le \delta} |f_n(A) - f(A)|, \qquad h(\delta) = \inf_{\|A-B\|_F = \delta} f(A) - f(B).$$

Consider any $A = B + cU$, where $\|U\|_F = 1$ and $c > \delta$. So $A$ lies outside a $\delta$-ball centered at $B$. Now consider the following point (exactly at distance $\delta$ from $B$) on the line joining $A$ and $B$:

$$\left(1 - \frac{\delta}{c}\right) B \ + \ \frac{\delta}{c} A = \left(1 - \frac{\delta}{c}\right) B + \frac{\delta}{c}(B + cU) = B + \delta U.$$

By convexity of $f_n$,

$$f_n(B + \delta U) = f_n\left(\left(1 - \frac{\delta}{c}\right) B + \frac{\delta}{c} A\right) \le \left(1 - \frac{\delta}{c}\right) f_n(B) + \frac{\delta}{c} f_n(A),$$

which implies

$$f_n(B + \delta U) - f_n(B) \ \le \ \frac{\delta}{c}\big(f_n(A) - f_n(B)\big).$$

Write $r_n(A) = f_n(A) - f(A)$. Then the above becomes

$$f(B + \delta U) - f(B) + \big(r_n(B + \delta U) - r_n(B)\big) \ \le \ \frac{\delta}{c}\big(f(A) - f(B) + r_n(A) - r_n(B)\big).$$

Consider the left-hand side. By definition of $h(\delta)$,

$$h(\delta) \le f(B + \delta U) - f(B).$$

Also, $r_n(B) \le \sup_{\|A-B\|_F \le \delta} |r_n(A)| = \Delta_n(\delta), \quad |r_n(B + \delta U)| \le \Delta_n(\delta)$. Thus

$$r_n(B) - r_n(B + \delta U) \le 2\Delta_n(\delta), \quad \Rightarrow \quad -2\Delta_n(\delta) \le r_n(B + \delta U) - r_n(B).$$

Combining with the earlier lower bound,

$$h(\delta) - 2\Delta_n(\delta) \ \le \ \frac{\delta}{c}\big(f_n(A) - f_n(B)\big), \quad \text{for any } A \text{ with } \|A - B\|_F > \delta.$$

Hence we have the implication of events:

$$\big\{0 < h(\delta) - 2\Delta_n(\delta)\big\} \ \subseteq \ \{f_n(A) > f_n(B)\}, \quad \forall \ A \text{ outside a } \delta\text{-ball around } B,$$

However, $f_n(B_n) \le f_n(B)$, since $B_n$ is a minimizer of $f_n$. Thus,

$$\limsup_{n\to\infty}\{\|B_n - B\|_F > \delta\} \ \subseteq \ \limsup_{n\to\infty}\{h(\delta) - 2\Delta_n(\delta) \le 0\} = \limsup_{n\to\infty} \Big\{\Delta_n(\delta) \ge \frac{h(\delta)}{2}\Big\}.$$

Therefore to prove our claim, it's enough to prove that,

$$\mathbf{P}\left[\limsup_{n\to\infty} \Big\{\Delta_n(\delta) \ge \frac{h(\delta)}{2}\Big\}\right] = 0.$$

Since $C$ is open and $B \in C$, there exists $r > 0$ such that the closure of the $r$-Frobenius ball, $\overline{B_F(B, r)} \subset C$. Therefore, for any $0 < \delta \le r$,

$$\{A \in C : \|A - B\|_F \le \delta\} = \overline{B_F(B, \delta)},$$

which is closed and bounded and hence is compact by the Heine–Borel theorem. Now since $f$ is the almost sure limit of $\{f_n\}_{n\geq 1}$, from Lemma A.4 we get,

$$\mathbf{P}[\lim_{n\to\infty} \Delta_n(\delta) = 0] = 1.$$

Therefore for any $\varepsilon > 0$, it's true that,

$$\mathbf{P}[\liminf_{n\to\infty} \{\Delta_n(\delta) < \varepsilon\}] = 1.$$

Because $f$ is finite and convex on $C$, it is continuous. Hence the function $g(\cdot) := f(\cdot) - f(B)$ is continuous. Consider the Frobenius sphere

$$S_\delta = \{A : \|A - B\|_F = \delta\}.$$

Since $g$ is continuous and $S_\delta$ is compact, $g$ attains its minimum on $S_\delta$. Hence there exists $A_\delta \in S_\delta$ such that $h(\delta) = g(A_\delta)$. But $A_\delta \neq B$ since $S_\delta$ does not contain $B$. Due to $B$ being the unique minimizer of $f$, we have $g(A_\delta) > 0$ or equivalently $h(\delta) > 0$, almost surely. This completes the proof. ■

**Lemma A.6** (Concentration of $\boldsymbol{W}_n$)**.** *Let $\boldsymbol{W}_n = n^{-1/2} X^\top E \in \mathbf{R}^{p\times T}$. Assume (C.1)-(C.3) hold true . Then, for every $1 \leq k \leq p$ and $1 \leq j \leq T$,*

$$\|\boldsymbol{W}_n\|_F = o(\log n) \text{ almost surely.}$$

Proof of Lemma A.6 Fix a coordinate pair $(k, j)$ and define the scalar sequence

$$Z_i := x_{i,k}\varepsilon_{i,j}, \qquad i \geq 1.$$

Since $\varepsilon_{i,j}$ are i.i.d. with finite variance $\sigma^2$. Now at this point we can directly follow the proof steps of Lemma 4.1 of Chatterjee and Lahiri (2010) to conclude that,

$$n^{-1/2} \sum_{i=1}^{n} x_{i,k}\varepsilon_{i,j} = o(\log n) \quad \text{almost surely}$$

This proves the coordinate-wise concentration for each pair $(k, j)$. This implies that, there exists $\Omega_{k,j}$ with $\mathbf{P}(\Omega_{k,j}) = 1$ such that for every $\omega \in \Omega_{k,j}$,

$$[\mathbf{W}_n]_{k,j}(\omega) = o(\log n).$$

Let $\Omega_0 = \bigcap_{k=1}^{p} \bigcap_{j=1}^{T} \Omega_{k,j}$ (a finite intersection), so $\mathbf{P}(\Omega_0) = 1$. For $\omega \in \Omega_0$, we have $\max_{k,j} |[\boldsymbol{W}_n]_{k,j}(\omega)| = o(\log n)$. Finally

$$\|W_n\|_F = \Big( \sum_{k,j} [\boldsymbol{W}_n]_{k,j}^2 \Big)^{1/2} \leq \sqrt{pT} \max_{k,j} |[\boldsymbol{W}_n]_{k,j}| = o(\log n) \text{ almost surely}$$

which completes the proof. ■

**Lemma A.7** (Directional Derivative of the Nuclear Norm)**.** *Let $A, R \in \mathbf{R}^{m\times n}$ with $m \geq n$ and $\operatorname{rank}(A) = r < n$. Suppose that $A$ admits the singular value decomposition $A = U_1\Sigma_1 V_1^\top$, where $\Sigma_1 \in \mathbf{R}^{r\times r}$ contains the positive singular values of $A$. Complete $U_1$ and $V_1$ to orthogonal matrices $U = \begin{pmatrix} U_1 & U_0 \end{pmatrix}$, $V = \begin{pmatrix} V_1 & V_0 \end{pmatrix}$, so that $A = U \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} V^\top$. Then the directional derivative of the nuclear norm exists and satisfies*

$$\lim_{\gamma\downarrow 0} \frac{\|A + \gamma R\|_* - \|A\|_*}{\gamma} = \operatorname{tr}(U_1^\top R\, V_1) + \big\|U_0^\top R\, V_0\big\|_*.$$

*Equivalently,*

$$\lim_{\gamma\downarrow 0}\frac{\|A+\gamma R\|_*-\|A\|_*}{\gamma}=\sum_{i=1}^{r}\boldsymbol{u}_i^\top R\,\boldsymbol{v}_i+\big\|U_0^\top R\,V_0\big\|_*,$$

*where $\{\boldsymbol{u}_i,\boldsymbol{v}_i\}_{i=1}^r$ are the singular vector pairs associated with the positive singular values of $A$.*

Proof of Lemma A.7 The result follows from Theorem 1 of Watson (1992). Indeed, Watson's characterization of the subdifferential of the nuclear norm gives (see example 2):

$$\partial\|A\|_*=\Big\{U_1V_1^\top+U_0TV_0^\top:\|T\|_{\mathrm{op}}\le 1\Big\}. \tag{A.1}$$

By a standard result from convex analysis (see, e.g., Theorem 23.4 of Rockafellar (1997) for characterization of directional derivatives of convex functions),

$$f'(A;R)=\sup_{G\in\partial f(A)}\langle G,R\rangle,$$

whenever $f$ is convex. Applying this identity to the convex function $f(A)=\|A\|_*$ and using the Frobenius inner product $\langle G,R\rangle=\mathrm{tr}(G^\top R)$ yields

$$\lim_{\gamma\downarrow 0}\frac{\|A+\gamma R\|_*-\|A\|_*}{\gamma}=\sup_{G\in\partial\|A\|_*}\mathrm{tr}(G^\top R).$$

Now using (A.1), this equals

$$\begin{aligned}\sup_{G\in\partial\|A\|_*}\mathrm{tr}(G^\top R)&=\sup_{\|T\|_{\mathrm{op}}\le 1}\mathrm{tr}\Big((U_1V_1^\top+U_0TV_0^\top)^\top R\Big)\\&=\sup_{\|T\|_{\mathrm{op}}\le 1}\Big[\mathrm{tr}(V_1U_1^\top R)+\mathrm{tr}(V_0T^\top U_0^\top R)\Big]\\&=\sup_{\|T\|_{\mathrm{op}}\le 1}\Big[\mathrm{tr}(U_1^\top R\,V_1)+\mathrm{tr}(T^\top U_0^\top R\,V_0)\Big],\end{aligned}$$

where the last equality follows from the cyclic property of the trace. Since the first term does not depend on $T$, it may be taken outside the supremum, yielding

$$\sup_{G\in\partial\|A\|_*}\mathrm{tr}(G^\top R)=\mathrm{tr}(U_1^\top R\,V_1)+\sup_{\|T\|_{\mathrm{op}}\le 1}\mathrm{tr}\big(T^\top U_0^\top R\,V_0\big).$$

Using the duality relation between the spectral and nuclear norms, $\sup_{\|T\|_{\mathrm{op}}\le 1}\mathrm{tr}(T^\top M)=\|M\|_*$, with $M=U_0^\top R\,V_0$, yields the stated formula. Now since

$$U_1=(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_r),\quad V_1=(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_r),$$

the $(i,j)$ entry of $U_1^\top RV_1$ is $(U_1^\top RV_1)_{ij}=\boldsymbol{u}_i^\top R\,\boldsymbol{v}_j$. Therefore, $\mathrm{tr}(U_1^\top RV_1)=\sum_{i=1}^r\boldsymbol{u}_i^\top R\,\boldsymbol{v}_i$. The proof is complete. ■

**Lemma A.8** (Concentration of $\tilde{A}_n$)**.** *Suppose (C.1)-(C.5) hold true. Then we have,*

$$\|\tilde{A}_n-A\|_F=o\Big(n^{-1/2}\log n\Big)\qquad\textit{almost surely.}$$

Proof of Lemma A.8 Set the local parameter $U := (\log n)^{-1} n^{1/2} (\tilde{A}_n - A)$, $\tilde{A}_n = A + n^{-1/2}(\log n)\, U$, and write the rescaled objective for $U$:

$$H_n(U) = (\log n)^{-2} \left[ \|Y - X(A + n^{-1/2}(\log n)\, U)\|_F^2 - \|Y - XA\|_F^2 \right]$$

$$+ (\log n)^{-2} \left[ \lambda_n \Big( \|A + n^{-1/2}(\log n)\, U\|_* - \|A\|_* \Big) \right]. \tag{A.2}$$

A direct expansion of the quadratic part gives

(A.3)

$$H_n(U) = -\frac{2}{\log n} \operatorname{tr}(U^\top W_n) + \operatorname{tr}\big(U^\top C_n U\big) + \frac{\lambda_n}{(\log n)^2} \Big( \|A + n^{-1/2}(\log n) U\|_* - \|A\|_* \Big).$$

Due to Lemma A.6 and for fixed $U$ we have, $\frac{1}{\log n} \operatorname{tr}(U^\top W_n) \leq \frac{1}{\log n} \|U\|_F \|W_n\|_F = o(1)$ almost surely. Next for the penalty term $P_n(U)$:

$$P_n(U) = (\log n)^{-1} \frac{\lambda_n}{n^{1/2}} \frac{\|\boldsymbol{\sigma}(A + n^{-1/2}(\log n) U)\|_1 - \|\boldsymbol{\sigma}(A)\|_1}{n^{-1/2} \log n}$$

Now due to assumption (C.4) and Lemma A.7 we have,

$$\frac{\lambda_n}{\sqrt{n}} \left( \frac{\|\boldsymbol{\sigma}(A + n^{-1/2}(\log n) U)\|_1 - \|\boldsymbol{\sigma}(A)\|_1}{n^{-1/2} \log n} \right) \longrightarrow \lambda_0 \left[ \sum_{j=1}^{r^*} \boldsymbol{u}_j^\top U\, \boldsymbol{v}_j + \big\| U_0^\top U\, V_0 \big\|_* \right],$$

where $\boldsymbol{u}_j, \boldsymbol{v}_j$ are the singular vectors corresponding to positive singular values of $A$. It remains to show this limiting quantity is $O(1)$. Since $\boldsymbol{u}_j \in \mathbb{R}^p$ and $\boldsymbol{v}_j \in \mathbb{R}^T$ are orthonormal vectors, they satisfy $\|\boldsymbol{u}_j\| = \|\boldsymbol{v}_j\| = 1$. Hence, by the definition of operator norm, $|\boldsymbol{u}_j^\top U \boldsymbol{v}_j| \leq \|U\|_{\mathrm{op}}$. Summing over $j = 1, \ldots, r^*$ yields

$$\left| \sum_{j=1}^{r^*} \boldsymbol{u}_j^\top U \boldsymbol{v}_j \right| \leq r^* \|U\|_{\mathrm{op}} = O(1),$$

since $r^*$ is fixed and $U$ is a fixed matrix. We use the standard inequality for the Frobenius norm:

$$\|ABC\|_F \leq \|A\|_{\mathrm{op}}\, \|B\|_{\mathrm{op}}\, \|C\|_F.$$

Applying this, we obtain $\|U_0^\top U V_0\|_F \leq \|U_0^\top\|_{\mathrm{op}}\, \|U\|_{\mathrm{op}}\, \|V_0\|_F$. Since $U_0$ and $V_0$ have orthonormal columns,

$$\|U_0^\top\|_{\mathrm{op}} = 1,\ \|V_0\|_{\mathrm{op}} = 1,\ \|V_0\|_F = \sqrt{T - r^*}.$$

As $p, T, r^*$ are fixed, $\|V_0\|_F = O(1)$. Therefore, $\|U_0^\top U V_0\|_F = O(\|U\|_{\mathrm{op}}) = O(1)$. Using $\|M\|_* \leq \sqrt{\operatorname{rank}(M)}\, \|M\|_F$ and the fact that the rank is bounded by a constant depending only on $p, T, r^*$, we conclude $\|U_0^\top U V_0\|_* = O(1)$. Since, $\lambda_0$ is finite, hence combining both bounds,

$$\lambda_0 \left[ \sum_{j=1}^{r^*} \boldsymbol{u}_j^\top U \boldsymbol{v}_j + \|U_0^\top U V_0\|_* \right] = O(1).$$

Therefore,

$$(\log n)^{-1} n^{1/2} (\tilde{A}_n - A) = \operatorname{argmin}_{U \in \mathbf{R}^{p \times T}} \Big[ tr(U^\top C_n U) + R_n(U) \Big], \tag{A.4}$$

where $R_n(U) = o(1)$ almost surely. Now in the notation of Lemma A.5, let $f_n(U) := tr(U^\top C_n U)$ and $f(U) := tr(U^\top C U)$. Clearly these are convex maps and $f_n(U) \to f(U)$ point wise as $n \to \infty$ due to Lemma A.3. Also since $C$ is positive definite, we must have,

$$\text{argmin}_{U \in \mathbf{R}^{p\times T}} f(U) = \mathbf{0}_{p\times T},$$

which is unique almost surely. Therefore the proof is complete due to Lemma A.5. ■

**Lemma A.9** (Concentration for the thresholded estimator)**.** *Let $\tilde{A}_n^{(\text{thr})}$ be the thresholded estimator defined in* (4.2)*. Then*

$$\big\|\tilde{A}_n^{(\text{thr})} - A\big\|_F = o\big(n^{-1/2}\log n\big) \qquad \textit{almost surely.}$$

Proof of Lemma A.9 Write $\tilde{A}_n^{(\text{thr})} - A = (\tilde{A}_n - A) + (\tilde{A}_n^{(\text{thr})} - \tilde{A}_n) = E_n + R_n$, where

$$E_n := \tilde{A}_n - A, \quad R_n := \tilde{A}_n^{(\text{thr})} - \tilde{A}_n.$$

Let $\tilde{\sigma}_j^{(n)} := \sigma_j(\tilde{A}_n), \quad \mathcal{J}_n := \{j : \tilde{\sigma}_j^{(n)} \le a_n\}$. Then $R_n = -\sum_{j\in\mathcal{J}_n} \tilde{\sigma}_j^{(n)} \tilde{\boldsymbol{u}}_j^{(n)} (\tilde{\boldsymbol{v}}_j^{(n)})^\top$. For $j = 1, \dots, m$, define $M_j := \tilde{\boldsymbol{u}}_j^{(n)} (\tilde{\boldsymbol{v}}_j^{(n)})^\top$. Then $R_n = -\sum_{j\in\mathcal{J}_n} \tilde{\sigma}_j^{(n)} M_j$. We first show that the matrices $\{M_j\}_{j=1}^m$ are orthonormal with respect to the Frobenius inner product. Indeed, for any $j, k$,

$$\langle M_j, M_k\rangle_F = \text{tr}\Big(\tilde{\boldsymbol{v}}_j^{(n)} (\tilde{\boldsymbol{u}}_j^{(n)})^\top \tilde{\boldsymbol{u}}_k^{(n)} (\tilde{\boldsymbol{v}}_k^{(n)})^\top\Big) = \big((\tilde{\boldsymbol{u}}_j^{(n)})^\top \tilde{\boldsymbol{u}}_k^{(n)}\big)\, \text{tr}\Big(\tilde{\boldsymbol{v}}_j^{(n)} (\tilde{\boldsymbol{v}}_k^{(n)})^\top\Big) = \delta_{jk},$$

since the left and right singular vectors form orthonormal systems. Consequently,

$$\|R_n\|_F^2 = \left\langle \sum_{j\in\mathcal{J}_n} \tilde{\sigma}_j^{(n)} M_j, \sum_{k\in\mathcal{J}_n} \tilde{\sigma}_k^{(n)} M_k \right\rangle_F = \sum_{j,k\in\mathcal{J}_n} \tilde{\sigma}_j^{(n)} \tilde{\sigma}_k^{(n)} \langle M_j, M_k\rangle_F = \sum_{j\in\mathcal{J}_n} \big(\tilde{\sigma}_j^{(n)}\big)^2.$$

Hence $\|R_n\|_F \le \sqrt{m}\, \max_{j\in\mathcal{J}_n} \tilde{\sigma}_j^{(n)}$. By Lemma A.8,

$$\|E_n\|_F = o\big(n^{-1/2}\log n\big) \quad \text{almost surely,}$$

and therefore $\|E_n\|_2 \le \|E_n\|_F = o(1)$ almost surely. Fix an outcome in the almost-sure event $\Omega_0$ on which $\|E_n\|_2 = o(1)$. Choose $\omega$ from this set. Since $\sigma_{r^*}(A) > 0$, there exists $N_1(\omega)$ such that $a_n < \frac{\sigma_{r^*}(A)}{2}$ for all $n \ge N_1(\omega)$. Moreover, there exists $N_2(\omega)$ such that $\|E_n\|_2 < \frac{\sigma_{r^*}(A)}{2}$ for all $n \ge N_2(\omega)$. For $n \ge \max\{N_1(\omega), N_2(\omega)\}$, Weyl's inequality yields, for every $j \le r^*$,

$$\tilde{\sigma}_j^{(n)} = \sigma_j(\tilde{A}_n) \ge \sigma_j(A) - \|E_n\|_2 \ge \sigma_{r^*}(A) - \|E_n\|_2 > \frac{\sigma_{r^*}(A)}{2} > a_n.$$

Therefore, $\mathcal{J}_n \subseteq \{r^* + 1, \dots, m\}$ eventually. Now let $j \in \mathcal{J}_n$. Since $j > r^*$, $\sigma_j(A) = 0$, and another application of Weyl's inequality gives

$$\tilde{\sigma}_j^{(n)} = \sigma_j(\tilde{A}_n) \le |\sigma_j(\tilde{A}_n) - \sigma_j(A)| \le \|E_n\|_2 \le \|E_n\|_F.$$

Consequently, $\max_{j\in\mathcal{J}_n} \tilde{\sigma}_j^{(n)} \le \|E_n\|_F$ eventually. Combining this with previous argument, $\|R_n\|_F \le \sqrt{m}\, \|E_n\|_F$ eventually. Finally, by the triangle inequality,

$$\|\tilde{A}_n^{(\text{thr})} - A\|_F \le \|E_n\|_F + \|R_n\|_F \le (1 + \sqrt{m})\, \|E_n\|_F \ \text{ eventually.}$$

Hence by Lemma A.8, $\|\tilde{A}_n^{(\text{thr})} - A\|_F = o\big(n^{-1/2}\log n\big)$ almost surely. Therefore the proof is complete. ■

**Lemma A.10** (Closeness between variances)**.** *Under the assumptions (C.1)-(C.3) we have,*

$$\left\|\tilde{\boldsymbol{S}}^*_{n,G} - I_T \otimes \sigma^2 C\right\|_F = o(1) \qquad \textit{almost surely},$$

*Recall that,* $\tilde{\boldsymbol{S}}^*_{n,G} = \mathrm{diag}(\tilde{\boldsymbol{S}}^*_{n,G,1}, \ldots, \tilde{\boldsymbol{S}}^*_{n,G,T}) = \sum_{t=1}^T \Psi_{tt} \otimes \tilde{\boldsymbol{S}}^*_{n,G,t}$*, where,* $\tilde{\boldsymbol{S}}^*_{n,G,t} = \mathbf{Var}_*(\bar{\boldsymbol{W}}^*_{n,G,t}) = \frac{1}{n}\sum_{i=1}^n \tilde{E}^{(thr)2}_{it}\, \boldsymbol{X}_i^\top \boldsymbol{X}_i$*, with* $\tilde{E}^{(thr)}_{it}$ *denote the* $(i,t)$*-th entry of* $\tilde{E}^{(thr)} = Y - X\tilde{A}^{(thr)}_n$ *and* $\Psi_{tt}$ *denote the* $T \times T$ *matrix with a* $1$ *in position* $(t,t)$ *and* $0$ *elsewhere.*

Proof of Lemma A.10 Note that

$$\left\|\tilde{\boldsymbol{S}}^*_{n,G} - I_T \otimes \sigma^2 C\right\|_F^2 = \sum_{t=1}^T \left\|\tilde{\boldsymbol{S}}^*_{n,G,t} - \sigma^2 C\right\|_F^2.$$

Since $T$ is fixed, it suffices to show, $\left\|\tilde{\boldsymbol{S}}^*_{n,G,t} - \sigma^2 C\right\|_F = o(1)$ almost surely for all $t = 1, \ldots, T$. Now writing, $\left\|\tilde{\boldsymbol{S}}^*_{n,G,t} - \sigma^2 C\right\|_F \le \left\|\tilde{\boldsymbol{S}}^*_{n,G,t} - \sigma^2 C_n\right\|_F + \sigma^2\left\|C_n - C\right\|_F$ The second term is $o(1)$ due to assumption (C.3). Now that the first term is $o(1)$ almost surely, follows from Lemma A.1, A.9 of Choudhury and Das (2026) and Lemma A.8. Hence the proof is complete. ■

**Lemma A.11** (Convergence of leading term to Gaussian)**.** *Under the assumptions (C.1)-(C.3), we have*

$$\boldsymbol{W}_n \xrightarrow{d} \boldsymbol{W} \;\; \textit{in } \mathbf{R}^{p\times T},$$

*where* $\boldsymbol{W} = [\boldsymbol{W}_1, ...., \boldsymbol{W}_T]$ *and* $\boldsymbol{W}_j \sim N_p(\mathbf{0}_{p\times 1}, \sigma^2 C)$ *independently for all* $j = 1, ..., T$*. Equivalently we will say vec* $(\boldsymbol{W}_n) \xrightarrow{d}$ *vec* $(\boldsymbol{W}) \in \mathbf{R}^{pT}$ *with vec* $(\boldsymbol{W}) \sim N_{pT}(\mathbf{0}_{pT\times 1},\, I_T \otimes \sigma^2 C)$*. Alternatively we can conclude the matrix-normal representation as,*

$$\mathbf{W} \sim \mathcal{MN}_{p\times T}\big(\mathbf{0}_{p\times T},\, U = \sigma^2 C,\, V = I_T\big).$$

Proof of Lemma A.11 Write the columns of the $E$ matrix as $z_1, \ldots, z_T$ (each of them is an $n$-dimensional vector), with $z_j = (\varepsilon_{1,j}, \ldots, \varepsilon_{n,j})^\top$. Then,

$$X^\top E = \begin{pmatrix} X^\top z_1 & \cdots & X^\top z_T \end{pmatrix}_{p\times T}, \qquad \text{and} \quad X^\top z_j = \sum_{i=1}^n \boldsymbol{x}_i\, \varepsilon_{i,j},$$

for all $j = 1, \ldots, T$, where $\boldsymbol{x}_i \in \mathbf{R}^p$ denotes the $i$-th row of $X$. Fix an index $j$. Then following the exact proof steps of Lemma A.11 of Choudhury and Das (2026), we can conclude that $\boldsymbol{W}_{n,j} = \frac{1}{\sqrt{n}} X^\top z_j \xrightarrow{d} \boldsymbol{W}_j \sim N_p(\mathbf{0}, \sigma^2 C)$. The columns of $\boldsymbol{W}$ are also independent which follows from the independence of the columns of $E$. Therefore it's safe to conclude that,

$$\text{vec }(\boldsymbol{W}_n) \xrightarrow{d} \text{ vec }(\boldsymbol{W}) \sim N_{pT}(\mathbf{0}_{pT\times 1},\, I_T \otimes \sigma^2 C).$$

The proof is complete. ■

**Lemma A.12** (Convergence of PB-leading term to Gaussian)**.** *Suppose* $\mathcal{L}(\cdot)$ *denote the law of the underlying random quantity and* $\mathcal{E}$ *denote the sigma algebra generated by* $\{\varepsilon_{it} : 1 \le i \le n,\;\; 1 \le t \le T\}$*. Then under assumptions (C.1)-(C.3) we have,*

$$\mathcal{L}(\bar{\boldsymbol{W}}^*_{n,G} \mid \mathcal{E}) \xrightarrow{d_*} \mathcal{MN}_{p\times T}\big(\mathbf{0}_{p\times T},\, U = \sigma^2 C,\, V = I_T\big) \quad \textit{almost surely}$$

$$\textit{Equivalently, } \mathcal{L}\Big(vec(\bar{\boldsymbol{W}}^*_{n,G}) \mid \mathcal{E}\Big) \xrightarrow{d_*} N_{pT\times 1}\big(\mathbf{0}_{pT\times 1},\, I_T \otimes \sigma^2 C\big) \quad \textit{almost surely}$$

*Here,* $d_*$ *denote convergence in distribution conditioned on the* $\{\varepsilon_{it} : 1 \le i \le n,\;\; 1 \le t \le T\}$.

Proof of Lemma A.12 The idea of the proof is similar to Lemma A.11 with application of Lemma A.1, A.11 of Choudhury and Das (2026), Lemma A.6, A.8 and A.10. Hence we skip the details. ■

**Lemma A.13.** *Suppose that $\{\boldsymbol{U}_n(\cdot)\}_{n\geq 1}$ and $\boldsymbol{U}_\infty(\cdot)$ are convex stochastic processes on $\mathbf{R}^p$ such that $\boldsymbol{U}_\infty(\cdot)$ has almost surely unique minimum $\xi_\infty$. Also assume that*
*(a) every finite dimensional distribution of $\boldsymbol{U}_n(\cdot)$ converges to that of $\boldsymbol{U}_\infty(\cdot)$, that is, for any natural number $k$ and for any $\{\boldsymbol{t}_1, .., \boldsymbol{t}_k\} \subset \mathbf{R}^p$, we have $(\boldsymbol{U}_n(\boldsymbol{t}_1), ...\boldsymbol{U}_n(\boldsymbol{t}_k)) \xrightarrow{d} (\boldsymbol{U}_\infty(\boldsymbol{t}_1), ...\boldsymbol{U}_\infty(\boldsymbol{t}_k))$.*
*(b) $\{\boldsymbol{U}_n(\cdot)\}_{n\geq 1}$ is equicontinuous on compact sets, in probability, i.e., for every $\epsilon, \eta, M > 0$, there exists a $\delta > 0$ such that*

$$\limsup_{n\to\infty} \mathbf{P}\left( \sup_{\{\|\boldsymbol{r}-\boldsymbol{s}\|<\delta, \max\{\|\boldsymbol{r}\|, \|\boldsymbol{s}\|\}<M\}} \left|\boldsymbol{U}_n(r) - \boldsymbol{U}_n(s)\right| > \eta \right) < \epsilon.$$

*Then we have* $\operatorname{Argmin}_{\boldsymbol{t}} \boldsymbol{U}_n(\boldsymbol{t}) \xrightarrow{d} \xi_\infty$.

Proof of lemma A.13. This result is proved as Lemma A.14 of Choudhury and Das (2026). ■

**Lemma A.14** (Generalized Wedin's $sin\Theta$ theorem)**.** *Let $A,\ \hat{A} \in \mathbf{R}^{p\times q}$ have singular values $\sigma_1 \geq \cdots \geq \sigma_{\min(p,q)},\ \hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_{\min(p,q)}$ respectively. Fix integers $1 \leq r \leq s \leq \operatorname{rank}(A)$ and assume that $\min\big(\sigma_{r-1}^2 - \sigma_r^2,\ \sigma_s^2 - \sigma_{s+1}^2\big) > 0$, where $\sigma_0^2 := \infty$ and $\sigma_{\operatorname{rank}(A)+1}^2 := -\infty$. Let $d := s - r + 1$, and let*

$$V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbf{R}^{q\times d}, \qquad \hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbf{R}^{q\times d}$$

*have orthonormal columns satisfying*

$$Av_j = \sigma_j u_j, \qquad \hat{A}\,\hat{v}_j = \hat{\sigma}_j \hat{u}_j, \qquad j = r, r+1, \ldots, s.$$

*Then*

$$\|\sin\Theta(\hat{V}, V)\|_F \ \leq\ \frac{2\big(2\sigma_1 + \|\hat{A} - A\|_{\mathrm{op}}\big)\ \min\big(d^{1/2}\|\hat{A} - A\|_{\mathrm{op}},\ \|\hat{A} - A\|_F\big)}{\min\big(\sigma_{r-1}^2 - \sigma_r^2,\ \sigma_s^2 - \sigma_{s+1}^2\big)}.$$

Proof of Lemma A.14 This result is proved as Theorem 4 in Yu et al. (2015). ■

**Corollary A.1** (Corollary from Lemma A.14)**.** *Let $A \in \mathbf{R}^{p\times T}$ have rank $r^*$ and distinct nonzero singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_{r^*} > 0$. Let $A = U\Sigma V^\top$ and $\tilde{A}_n = \tilde{U}_n\tilde{\Sigma}_n\tilde{V}_n^\top$ be the SVD of the perturbed matrix $\tilde{A}_n = A + E_n$. Write $(\boldsymbol{u}_j, \boldsymbol{v}_j)$ and $(\tilde{\boldsymbol{u}}_j^{(n)}, \tilde{\boldsymbol{v}}_j^{(n)})$ for the corresponding left/right singular vectors. Set $r = s = j$ with $d = s - r + 1 = 1$ and*

$$\delta_j := \min\big(\sigma_{j-1}^2 - \sigma_j^2,\ \sigma_j^2 - \sigma_{j+1}^2\big), \quad \sigma_0 := \infty,\ \ \sigma_{r^*+1} := 0.$$

*Then, for each $1 \leq j \leq r^*$, Lemma A.14 implies the per–vector bound*
(A.5)

$$\sin\theta\big(\tilde{\boldsymbol{u}}_j^{(n)}, \boldsymbol{u}_j\big) \ \leq\ \frac{2(2\sigma_1 + \|E_n\|_{\mathrm{op}})\ \|E_n\|_{\mathrm{op}}}{\delta_j} \text{ and } \sin\theta\big(\tilde{\boldsymbol{v}}_j^{(n)}, \boldsymbol{v}_j\big) \ \leq\ \frac{2(2\sigma_1 + \|E_n\|_{\mathrm{op}})\ \|E_n\|_{\mathrm{op}}}{\delta_j}.$$

*Suppose in addition that the spectral gaps satisfy $\delta := \min_{1\leq j\leq r^*} \delta_j > 0$. Since, $\sigma_1$ is fixed and due to Lemma A.8, $\|E_n\|_F = o(n^{-1/2}\log n)$ almost surely, therefore for large enough $n$ and for all $1 \leq j \leq r^*$ we have the modified uniform bound as,*

$$\sin\theta\big(\tilde{\boldsymbol{u}}_j^{(n)}, \boldsymbol{u}_j\big) \ \leq\ \frac{5\sigma_1\|E_n\|_{\mathrm{op}}}{\delta} \text{ and } \sin\theta\big(\tilde{\boldsymbol{v}}_j^{(n)}, \boldsymbol{v}_j\big) \ \leq\ \frac{5\sigma_1\|E_n\|_{\mathrm{op}}}{\delta}. \tag{A.6}$$

**Lemma A.15** (Small singular values determined by the bottom block)**.** *Let*

$$M = \begin{pmatrix} B & C \\ D & E \end{pmatrix}, \qquad B \in \mathbf{R}^{r\times r},\ E \in \mathbf{R}^{(m-r)\times(m-r)}.$$

*Assume $B$ is invertible and $\sigma_{\min}(B) \to \infty$. Assume that $\|C\|, \|D\|, \|E\|$ are uniformly bounded. Then for $k = 1, \dots, m-r$,*

$$\sigma_{r+k}(M) = \sigma_k(E) + o(1).$$

Proof of Lemma A.15 Define $L := \begin{pmatrix} I & 0 \\ DB^{-1} & I \end{pmatrix}$, $\widetilde{M} := \begin{pmatrix} B & C \\ 0 & E - DB^{-1}C \end{pmatrix}$. Thus, $L\widetilde{M} = \begin{pmatrix} B & C \\ D & E \end{pmatrix} = M$. Denote $S = E - DB^{-1}C$. For any matrices $A, X$ it's known that,

$$\sigma_i(AX) \le \|A\|\,\sigma_i(X), \qquad \sigma_i(AX) \ge \sigma_{\min}(A)\,\sigma_i(X).$$

Applying this with $A = L$, $X = \widetilde{M}$, we obtain $\sigma_i(M) \le \|L\|\,\sigma_i(\widetilde{M})$, $\sigma_i(M) \ge \sigma_{\min}(L)\,\sigma_i(\widetilde{M})$. Define

$$K := \begin{pmatrix} 0 & 0 \\ DB^{-1} & 0 \end{pmatrix}, \quad \text{so that} \quad L = I + K.$$

Then it's easy to observe that, $\|L\| \le \|I\| + \|K\| = 1 + \|K\| \le 1 + \frac{\|D\|}{\sigma_{\min}(B)}$ and $\sigma_{\min}(L) \ge 1 - \|K\| \ge 1 - \frac{\|D\|}{\sigma_{\min}(B)}$. This will finally give us that,

$$\left(1 - \frac{\|D\|}{\sigma_{\min}(B)}\right)\sigma_i(\widetilde{M}) \ \le\ \sigma_i(M) \ \le\ \left(1 + \frac{\|D\|}{\sigma_{\min}(B)}\right)\sigma_i(\widetilde{M}). \tag{A.7}$$

Singular values satisfy $\sigma_i(\widetilde{M})^2 = \lambda_i(\widetilde{M}^\top\widetilde{M})$, where $\lambda_i(\cdot)$ denote $i$-th largest eigen value of that underlined matrix. Again we will have $\widetilde{M}^\top\widetilde{M} = H + J$, where,

$$H := \begin{pmatrix} B^\top B & 0 \\ 0 & S^\top S \end{pmatrix}, \quad J := \begin{pmatrix} 0 & B^\top C \\ C^\top B & C^\top C \end{pmatrix}.$$

The eigenvalues of $H$ are $\{\sigma_i(B)^2\}_{i=1}^r \cup \{\sigma_j(S)^2\}_{j=1}^{m-r}$. Now since $\sigma_{\min}(B) \to \infty$, we have a spectral separation: $\lambda_r(H) = \sigma_{\min}(B)^2 \to \infty$, $\lambda_{r+1}(H) = \sigma_1(S)^2$. Using the operator norm bound,

$$\sigma_1(S) = \|S\| = \|E - DB^{-1}C\| \le \|E\| + \|DB^{-1}C\| \le \|E\| + \frac{\|C\|\,\|D\|}{\sigma_{\min}(B)} = O(1).$$

This will give us that, $\lambda_r(H) \gg \lambda_{r+1}(H)$ eventually. Note that $B$ is invertible matrix. Define, $T := \begin{pmatrix} I & 0 \\ -C^\top B(B^\top B)^{-1} & I \end{pmatrix}$ which is invertible. Then it's simple to verify that, $\widetilde{M}^\top\widetilde{M} = T^{-1}H(T^{-1})^\top$. Write

$$T = I + K, \quad K = \begin{pmatrix} 0 & 0 \\ -C^\top B(B^\top B)^{-1} & 0 \end{pmatrix}.$$

A direct computation shows that $K^2 = 0$, hence

$$(I+K)(I-K) = I \text{ and } (I-K)(I+K) = I \implies T^{-1} = (I+K)^{-1} = I - K.$$

Therefore, $\|T^{-1}\| \leq 1 + \|K\| = 1 + \|C^\top (B^{-1})^\top\| \leq 1 + \frac{\|C\|}{\sigma_{\min}(B)}$. Thus we will have, $\sigma_{\min}(T^{-1}) = \sigma_{\min}(I - K) \geq 1 - \frac{\|C\|}{\sigma_{\min}(B)}$. Using the Rayleigh–Ritz characterization for every $i$,

$$[\sigma_{\min}(T^{-1})]^2 \, \lambda_i(H) \;\leq\; \lambda_i(\widetilde{M}^\top \widetilde{M}) \;\leq\; \|T^{-1}\|^2 \, \lambda_i(H).$$

This will in turn imply that,

$$\left(1 - \frac{\|C\|}{\sigma_{\min}(B)}\right)^2 \lambda_i(H) \;\leq\; \lambda_i(\widetilde{M}^\top \widetilde{M}) \;\leq\; \left(1 + \frac{\|C\|}{\sigma_{\min}(B)}\right)^2 \lambda_i(H).$$

Expanding the squares,

$$\left(1 - \frac{\|C\|}{\sigma_{\min}(B)}\right)^2 = 1 - 2\,\frac{\|C\|}{\sigma_{\min}(B)} + \frac{\|C\|^2}{\sigma_{\min}(B)^2}, \left(1 + \frac{\|C\|}{\sigma_{\min}(B)}\right)^2 = 1 + 2\,\frac{\|C\|}{\sigma_{\min}(B)} + \frac{\|C\|^2}{\sigma_{\min}(B)^2}.$$

Since $\|C\|$ is bounded and $\sigma_{\min}(B) \to \infty$, the quadratic term is of smaller order. Hence

$$\left(1 - \frac{\|C\|}{\sigma_{\min}(B)}\right)^2 = 1 + O\left(\frac{\|C\|}{\sigma_{\min}(B)}\right), \left(1 + \frac{\|C\|}{\sigma_{\min}(B)}\right)^2 = 1 + O\left(\frac{\|C\|}{\sigma_{\min}(B)}\right).$$

Therefore,

$$\lambda_i(\widetilde{M}^\top \widetilde{M}) = \lambda_i(H) \, \left(1 + O\left(\frac{\|C\|}{\sigma_{\min}(B)}\right)\right).$$

Since, for $k = 1, \ldots, m - r$, we have established that $\lambda_{r+k}(H) = \sigma_k(S)^2 = O(1)$, we hence write;

$$\lambda_{r+k}(\widetilde{M}^\top \widetilde{M}) = \sigma_{r+k}(\widetilde{M})^2 = \sigma_k(S)^2 \left(1 + O\left(\frac{\|C\|}{\sigma_{\min}(B)}\right)\right).$$

Taking square roots and using for small enough $\varepsilon$ that, $\sqrt{1+\varepsilon} = 1 + O(\varepsilon)$, we finally have;

$$\sigma_{r+k}(\widetilde{M}) = \sigma_k(S) + O\left(\frac{\|C\|}{\sigma_{\min}(B)}\right).$$

Substituting $S = E - DB^{-1}C$ we get, $\|S - E\| \leq \|D\| \, \|B^{-1}\| \, \|C\| = \frac{\|C\| \, \|D\|}{\sigma_{\min}(B)}$. Thus, by Weyl's inequality, $\sigma_k(S) = \sigma_k(E) + O\left(\frac{\|C\| \, \|D\|}{\sigma_{\min}(B)}\right)$ and hence

$$\sigma_{r+k}(\widetilde{M}) = \sigma_k(E) + O\left(\frac{\|C\|[1 + \|D\|]}{\sigma_{\min}(B)}\right).$$

Using this and equation (A.7) we finally write,

$$\sigma_{r+k}(M) = \sigma_k(E) + O\left(\frac{\|D\|}{\sigma_{\min}(B)} \, \sigma_k(E)\right) + O\left(\frac{\|C\| \, [1 + \|D\|]}{\sigma_{\min}(B)}\right) + O\left(\frac{\|C\| \, [1 + \|D\|] \|D\|}{\sigma_{\min}(B)^2}\right).$$

Since $\sigma_k(E) \leq \sigma_1(E) = \|E\| = O(1)$, $\|D\|, \|C\| = O(1)$ and $\sigma_{min}(B) \to \infty$, this will imply for all $k = 1, ..., m - r$ that,

$$\sigma_{r+k}(M) = \sigma_k(E) + o(1).$$

Therefore the proof is complete. ■

## APPENDIX B. HADAMARD DIFFERENTIABILITY AND RELATED LEMMAS

**Definition 1** (Hadamard differentiability, Van der Vaart and Wellner (1996))**.** *A map $\Phi : \mathbb{D} \mapsto \mathbb{E}$ is said to be Hadamard differentiable at $\theta$ if there is a continuous linear map $\Phi'(\theta) : \mathbb{D} \mapsto \mathbb{E}$ such that,*

$$\frac{\Phi(\theta + t_n h_n) - \Phi(\theta)}{t_n} \to \Phi'(\theta; h),$$

*for every sequence of real numbers $t_n \to 0$ and every sequence $h_n \to h$ as $n \to \infty$.*

**Definition 2** (Hadamard directional differentiability, Fang & Santos (2019))**.** *Let $\mathbb{D}$ and $\mathbb{E}$ be Banach spaces and let $\Phi : \mathbb{D}_\Phi \subseteq \mathbb{D} \to \mathbb{E}$. The map $\Phi$ is said to be Hadamard directionally differentiable at $\theta \in \mathbb{D}_\Phi$ tangentially to a set $\mathbb{D}_0 \subseteq \mathbb{D}$ if there exists a continuous map $\Phi'(\theta) : \mathbb{D}_0 \to \mathbb{E}$ such that*

$$\left\| \frac{\Phi(\theta + t_n h_n) - \Phi(\theta)}{t_n} - \Phi'(\theta; h) \right\|_{\mathbb{E}} \longrightarrow 0,$$

*for every sequence $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbf{R}_+$ satisfying $t_n \downarrow 0$, $h_n \to h \in \mathbb{D}_0$, and $\theta + t_n h_n \in \mathbb{D}_\Phi$ for all $n$. Here $\|\cdot\|_{\mathbb{E}}$ denote the appropriate norm of the underlying metric space $\mathbb{E}$.*

**Remark B.1.** *Hadamard differentiability requires the existence of a continuous linear derivative $\Phi'_\theta$, and allows perturbations approaching zero from both signs. Hadamard directional differentiability relaxes this requirement in two ways: (i) the derivative $\Phi'_\theta$ need not be linear, and (ii) convergence is restricted to one-sided sequences $t_n \downarrow 0$ and tangential directions $h_n \to h$.*

*Consequently, every Hadamard differentiable map is Hadamard directionally differentiable, but the converse fails in general. Many functionals arising in modern statistics, such as maxima, minima, norms, and spectral operators, are not classically differentiable but admit well-defined directional derivatives. This weaker notion is therefore essential for non-smooth functionals. As shown in Fang & Santos (2019), Hadamard directional differentiability is sufficient to establish a generalized functional delta method and valid bootstrap procedures for such non-smooth maps.*

Now consider $\mathbb{D} = \mathbf{R}^{p\times T}$, $\mathbb{E} = \mathbf{R}$ or $\mathbf{R}^m$ as needed and $\theta = A^*$. We will consider $H_n \to H$ in appropriate norm of the underlying space. We now provide the following results.

**Lemma B.1** (Hadamard differentiability of entry functionals)**.** *Define*

$$\Phi_1(A) = A_{1,1}, \qquad \Phi_2(A) = A_{p,T}.$$

*Then $\Phi_1$ and $\Phi_2$ are Hadamard differentiable at every $A^* \in \mathbf{R}^{p\times T}$ with derivatives*

$$\Phi'_1(A^*; H) = H_{1,1}, \qquad \Phi'_2(A^*; H) = H_{p,T}.$$

Proof of Lemma B.1: We prove the result for $\Phi_1(A) = A_{1,1}$. The proof for $\Phi_2$ is identical. Let $t_n \to 0$ be any real sequence and let $H_n \to H$ in $\mathbf{R}^{p\times T}$ in Frobenius norm. Consider

$$\frac{\Phi_1(A + t_n H_n) - \Phi_1(A)}{t_n} = \frac{(A_{1,1} + t_n H_{n,1,1}) - A_{1,1}}{t_n} = H_{n,1,1}.$$

Since $H_n \to H$, we have $H_{n,1,1} \to H_{1,1}$. Hence $\frac{\Phi_1(A^* + t_n H_n) - \Phi_1(A^*)}{t_n} \longrightarrow H_{1,1}$. Thus

$$\Phi'_1(A^*; H) = H_{1,1}.$$

The map $H \mapsto H_{1,1}$ is linear. Now if $H_n \to H$, then

$$|\Phi_1'(A^*; H_n) - \Phi_1'(A^*; H)| = |H_{n,1,1} - H_{1,1}| \to 0.$$

Hence the derivative map is continuous. Therefore, by Definition 1, $\Phi_1$ is Hadamard differentiable at $A^*$. The same argument with indices $(p, T)$ proves the result for $\Phi_2$. ■

**Lemma B.2** (Directional differentiability of columnwise $\ell_1$-sum max/min functionals)**.** *Define*

$$\Phi_3(A) = \max_{1\le j\le T} \sum_{i=1}^{p} |A_{ij}|, \qquad \Phi_4(A) = \min_{1\le j\le T} \sum_{i=1}^{p} |A_{ij}|.$$

*For a fixed $A^* \in \mathbb{R}^{p\times T}$, let*

$$S_j(A) = \sum_{i=1}^{p} |A_{ij}|,\ \mathcal{J}_{\max} = \{j :\ S_j(A^*) = \Phi_3(A^*)\},\ \mathcal{J}_{\min} = \{j :\ S_j(A^*) = \Phi_4(A^*)\}.$$

*Then $\Phi_3$ and $\Phi_4$ are Hadamard directionally differentiable at $A^*$ with derivatives*

$$\Phi_3'(A^*; H) = \max_{j\in\mathcal{J}_{\max}} D_j(H), \qquad \Phi_4'(A^*; H) = \min_{j\in\mathcal{J}_{\min}} D_j(H),$$

*where for each column $j$: $D_j(H) = \sum_{i:\, A^*_{ij}\neq 0} \operatorname{sgn}(A^*_{ij})\, H_{ij} + \sum_{i:\, A^*_{ij}=0} |H_{ij}|$. If, additionally, every entry in the maximizing (respectively minimizing) column is nonzero and $\mathcal{J}_{max}$ (or $\mathcal{J}_{min}$) is a singleton set, then $\Phi_3$ (or $\Phi_4$) is fully Hadamard differentiable.*

Proof of Lemma B.2: We prove the result for $\Phi_3$. The proof for $\Phi_4$ is analogous. Fix sequences $t_n \downarrow 0$ and $H_n \to H$. For each entry $(i, j)$, the absolute value satisfies

$$|A^*_{ij} + t_n H_{n,ij}| = |A^*_{ij}| + t_n\, \psi_{ij}(H_{n,ij}) + o(t_n),$$

where

$$\psi_{ij}(h) = \begin{cases} \operatorname{sgn}(A^*_{ij})\, h, & A^*_{ij} \neq 0, \\ |h|, & A^*_{ij} = 0. \end{cases}$$

Summing over $i$ gives for each column $j$,

$$S_j(A^* + t_n H_n) = S_j(A^*) + t_n D_j(H_n) + o(t_n).$$

Since $H_n \to H$, we have $D_j(H_n) \to D_j(H)$ by continuity. Now consider $\Phi_3$. By definition,

$$\Phi_3(A^* + t_n H_n) = \max_{1\le j\le T} S_j(A^* + t_n H_n).$$

Substituting the expansion,

$$\Phi_3(A^* + t_n H_n) = \max_{1\le j\le T} \Big\{ S_j(A^*) + t_n D_j(H_n) + o(t_n) \Big\}.$$

For every $j \notin \mathcal{J}_{\max}$ we have the strict inequality $S_j(A^*) < \Phi_3(A^*)$. Since the number of columns $T$ is finite, we may define

$$\delta_j := \Phi_3(A^*) - \max_{j\notin\mathcal{J}_{\max}} S_j(A^*) > 0, \quad \delta := \min_{j\notin\mathcal{J}_{\max}} \delta_j.$$

Fix $j \notin \mathcal{J}_{\max}$ and $k \in \mathcal{J}_{\max}$. Then using the earlier expansion we can write;

$$S_k(A^* + t_n H_n) = \Phi_3(A^*) + t_n D_k(H_n) + o(t_n),$$
$$S_j(A^* + t_n H_n) = S_j(A^*) + t_n D_j(H_n) + o(t_n).$$

Subtracting,

(B.1)

$$S_k(A^* + t_n H_n) - S_j(A^* + t_n H_n) = \underbrace{\Phi_3(A^*) - S_j(A^*)}_{\geq \delta} + t_n\big(D_k(H_n) - D_j(H_n)\big) + o(t_n).$$

Recall $D_j(H_n) = \sum_{i:\, A^*_{ij} \neq 0} \operatorname{sgn}(A^*_{ij})\, H_{n,ij} \;+\; \sum_{i:\, A^*_{ij}=0} |H_{n,ij}|$. This will imply that,

$$|D_j(H_n)| \leq \sum_{i=1}^{p} |H_{n,ij}| \leq \sqrt{p}\, \|H_{n,\cdot j}\|_2 \leq \sqrt{p}\, \|H_n\|_F.$$

Since $H_n \to H$, the sequence $\{\|H_n\|_F\}$ is bounded. Hence for large enough $n$,

$$t_n\, |D_k(H_n) - D_j(H_n)| \leq 2\sqrt{p}\|H_n\|_F\, t_n \longrightarrow 0.$$

Therefore from (B.1) for large enough $n$ we have,

$$S_k(A^* + t_n H_n) > S_j(A^* + t_n H_n) + \delta/2.$$

Hence the maximum is attained over $\mathcal{J}_{\max}$:

$$\Phi_3(A^* + t_n H_n) = \Phi_3(A^*) + t_n \max_{j \in \mathcal{J}_{\max}} D_j(H_n) + o(t_n).$$

Dividing by $t_n$ and letting $n \to \infty$ yields

$$\frac{\Phi_3(A^* + t_n H_n) - \Phi_3(A^*)}{t_n} \longrightarrow \max_{j \in \mathcal{J}_{\max}} D_j(H),$$

which is the Hadamard directional derivative and continuous map in $H$. The proof for $\Phi_4$ is identical with $\min$ in place of $\max$. ∎

Two key identities used in the proof. Let $J$ be a finite index set.

**(I) Pulling out constants from a maximum.** For any constant $c \in \mathbf{R}$ and any collection $\{x_j\}_{j \in J}$, $\max_{j \in J}(c + x_j) = c + \max_{j \in J} x_j$.

*Proof.* Since $c$ does not depend on $j$, for any $j_1, j_2 \in J$, $c + x_{j_1} \geq c + x_{j_2} \iff x_{j_1} \geq x_{j_2}$. Thus the maximizer of $c + x_j$ coincides with that of $x_j$, hence $\max_{j \in J}(c + x_j) = c + \max_{j \in J} x_j$. ∎

**(II) Stability of the maximum under $o(t_n)$ perturbations.**

Let $t_n \downarrow 0$, $J$ be finite, and assume $r_{j,n} = o(t_n)$ uniformly in $j \in J$. Then

$$\max_{j \in J} \big(t_n D_j(H_n) + r_{j,n}\big) = t_n \max_{j \in J} D_j(H_n) + o(t_n).$$

*Proof.* Write $r_{j,n} = t_n \epsilon_{j,n}$ where $\epsilon_{j,n} \to 0$ uniformly in $j$. Then

$$t_n D_j(H_n) + r_{j,n} = t_n\big(D_j(H_n) + \epsilon_{j,n}\big),$$

so $\max_{j \in J} \big(t_n D_j(H_n) + r_{j,n}\big) = t_n \max_{j \in J} \big(D_j(H_n) + \epsilon_{j,n}\big)$. Let $M_n := \max_{j \in J} D_j(H_n)$. Then

$$D_j(H_n) + \epsilon_{j,n} \leq M_n + \max_{k \in J} |\epsilon_{k,n}|,$$

implying $\max_{j \in J} \big(D_j(H_n) + \epsilon_{j,n}\big) \leq M_n + \max_{k \in J} |\epsilon_{k,n}|$. Let $j_n \in \arg\max_{j \in J} D_j(H_n)$. Then

$$\max_{j \in J} \big(D_j(H_n) + \epsilon_{j,n}\big) \geq D_{j_n}(H_n) + \varepsilon_{j_n,n} = M_n + \epsilon_{j_n,n} \geq M_n - \max_{k \in J} |\epsilon_{k,n}|.$$

Thus, $\big|\max_{j \in J} \big(D_j(H_n) + \epsilon_{j,n}\big) - M_n\big| \leq \max_{k \in J} |\epsilon_{k,n}| \to 0$, so

$$\max_{j \in J} \big(D_j(H_n) + \epsilon_{j,n}\big) = \max_{j \in J} D_j(H_n) + o(1).$$

Multiplying by $t_n$ yields $\max_{j\in J}\left(t_n D_j(H_n)+r_{j,n}\right)=t_n\max_{j\in J}D_j(H_n)+o(t_n)$. ■

**Lemma B.3** (Directional differentiability of $\ell_2$-norm max/min functionals)**.** *Define*

$$\Phi_5(A)=\max_{1\le j\le T}\|A_{\cdot j}\|_2,\qquad \Phi_6(A)=\min_{1\le j\le T}\|A_{\cdot j}\|_2.$$

*For a fixed $A^*\in\mathbb{R}^{p\times T}$, let*

$$R_j(A)=\|A_{\cdot j}\|_2,\ \mathcal{J}_{\max}=\{j:\ R_j(A^*)=\Phi_5(A^*)\},\ \mathcal{J}_{\min}=\{j:\ R_j(A^*)=\Phi_6(A^*)\}.$$

*Then $\Phi_5$ and $\Phi_6$ are Hadamard directionally differentiable at $A^*$ with derivatives*

$$\Phi_5'(A^*;H)=\max_{j\in\mathcal{J}_{\max}}E_j(H),\qquad \Phi_6'(A^*;H)=\min_{j\in\mathcal{J}_{\min}}E_j(H),$$

*where for each column $j$,*

$$E_j(H)=\begin{cases}\dfrac{\langle A^*_{\cdot j},\,H_{\cdot j}\rangle}{\|A^*_{\cdot j}\|_2}, & \text{if } A^*_{\cdot j}\neq 0,\\[2ex] \|H_{\cdot j}\|_2, & \text{if } A^*_{\cdot j}=0.\end{cases}$$

*If, additionally, every entry in the maximizing (respectively minimizing) column is nonzero and $\mathcal{J}_{max}$ (or $\mathcal{J}_{min}$) is a singleton set, then $\Phi_5$ (or $\Phi_6$) is fully Hadamard differentiable.*

Proof of Lemma B.3: We prove the result for $\Phi_5$. The proof for $\Phi_6$ is analogous with $\max$ replaced by $\min$. Fix a column $j$ and define $R_j(A)=\|A_{\cdot j}\|_2$. Consider sequences $t_n\downarrow 0$ and $H_n\to H$ in Frobenius norm. We study $R_j(A^*+t_nH_n)$.
Case 1: $A^*_{\cdot j}\neq 0$. Write $a=A^*_{\cdot j}$ and $h_n=H_{n,\cdot j}$. Then

$$\|a+t_nh_n\|_2^2=\|a\|_2^2+\underbrace{2t_n\langle a,h_n\rangle+t_n^2\|h_n\|_2^2}_{:=u(t_n)}.$$

Note that, $u(t_n)=O(t_n)$ and $t_n^2\|h_n\|_2^2=o(t_n)$ for large enough $n$. Now taking square roots and using a Taylor expansion of $x\mapsto\sqrt{x}$ at $x=\|a\|_2^2$ gives

$$\begin{aligned}\|a+t_nh_n\|_2&=\|a\|_2+\frac{1}{2\|a\|_2}\Big[2t_n\langle a,h_n\rangle+t_n^2\|h_n\|_2^2\Big]+o(u(t_n))\\&=\|a\|_2+t_n\,\frac{\langle a,h_n\rangle}{\|a\|_2}+o(t_n).\end{aligned}\tag{B.2}$$

Since $h_n\to h$, we obtain

$$R_j(A^*+t_nH_n)=R_j(A^*)+t_n\,\frac{\langle A^*_{\cdot j},H_{\cdot j}\rangle}{\|A^*_{\cdot j}\|_2}+o(t_n).$$

Case 2: $A^*_{\cdot j}=0$. Then we have,

$$R_j(A^*+t_nH_n)=\|t_nH_{n,\cdot j}\|_2=t_n\|H_{n,\cdot j}\|_2=t_n\|H_{\cdot j}\|_2+o(t_n).$$

Thus for every $j$,

$$R_j(A^*+t_nH_n)=R_j(A^*)+t_nE_j(H)+o(t_n),$$

with $E_j(H)$ defined in the statement. Let $\mathcal{J}_{\max}=\{j:\ R_j(A^*)=\Phi_5(A^*)\}$. For any $k\notin\mathcal{J}_{\max}$, there exists $\delta>0$ such that $R_k(A^*)\le\Phi_5(A^*)-\delta$. Hence

$$\begin{aligned}R_k(A^*+t_nH_n)&=R_k(A^*)+t_nE_k(H)+o(t_n)\\&\le\Phi_5(A^*)-\delta+t_nE_k(H)+o(t_n).\end{aligned}$$

Since $t_n \to 0$, for sufficiently large $n$ this is strictly smaller than $\Phi_5(A^*)+t_nE_j(H)+o(t_n)$ for any $j \in \mathcal{J}_{\max}$. Thus, for large $n$, the maximum is attained over $\mathcal{J}_{\max}$. Therefore,

$$\begin{aligned}\Phi_5(A^* + t_nH_n) &= \max_{j\in\mathcal{J}_{\max}} R_j(A^* + t_nH_n)\\ &= \Phi_5(A^*) + t_n \max_{j\in\mathcal{J}_{\max}} E_j(H) + o(t_n).\end{aligned}$$

Dividing by $t_n$ and letting $n \to \infty$ yields

$$\frac{\Phi_5(A^* + t_nH_n) - \Phi_5(A^*)}{t_n} \longrightarrow \max_{j\in\mathcal{J}_{\max}} E_j(H).$$

Each map $H \mapsto E_j(H)$ is continuous on $\mathbf{R}^{p\times T}$. The maximum of finitely many continuous functions is continuous. Hence the directional derivative is continuous in $H$.

**Lemma B.4** (Directional differentiability of the singular value map)**.** *Let $m = \min(p, T)$, and define*

$$\Phi : \mathbf{R}^{p\times T} \to \mathbf{R}^m, \qquad \Phi(A) = (\sigma_1(A), \ldots, \sigma_m(A))^\top,$$

*where $\sigma_1(A) > \cdots > \sigma_{r^*}(A) > 0$, and $\sigma_j(A) = 0,\ j = r^* + 1, \ldots, m$. Let*

$$U_0 \in \mathbf{R}^{p\times(p-r^*)}, V_0 \in \mathbf{R}^{T\times(T-r^*)}$$

*denote the left and right null singular subspaces. Then $\Phi$ is Hadamard directionally differentiable at $A^*$ with derivative*

$$\Phi'_{A^*}(H) = \begin{pmatrix} (u_1^*)^\top H v_1^* \\ \vdots \\ (u_{r^*}^*)^\top H v_{r^*}^* \\ \sigma_1(U_0^\top H V_0) \\ \vdots \\ \sigma_{m-r^*}(U_0^\top H V_0) \end{pmatrix}.$$

*Proof of Lemma B.4:* Fix sequences $t_n \downarrow 0$ and $H_n \to H$ in $\mathbf{R}^{p\times T}$. Define $A_n = A + t_nH_n$. We show that

$$\frac{\Phi(A^* + t_nH_n) - \Phi(A^*)}{t_n} \to \Phi'_{A^*}(H).$$

By definition,

$$\Phi(A_n) - \Phi(A^*) = \begin{pmatrix} \sigma_1(A_n) - \sigma_1(A^*) \\ \vdots \\ \sigma_{r^*}(A_n) - \sigma_{r^*}(A^*) \\ \sigma_{r^*+1}(A_n) \\ \vdots \\ \sigma_m(A_n) \end{pmatrix}.$$

Since $\sigma_1(A^*), \ldots, \sigma_{r^*}(A^*)$ are simple singular values, Stewart's perturbation expansion (cf. Stewart (1990)) yields

$$\sigma_j(A^* + t_nH_n) = \sigma_j(A^*) + t_n(u_j^*)^\top H_n v_j^* + O(t_n\|H_n\|^2), \qquad j = 1, \ldots, r^*.$$

Since $H_n \to H$ as $n \to \infty$ and $t_n \downarrow 0$,we get,

$$\frac{\sigma_j(A^* + t_nH_n) - \sigma_j(A^*)}{t_n} \to (u_j^*)^\top H v_j^*, \qquad j = 1, \ldots, r^*.$$

Now consider the remaining singular values jointly. Since

$$\sigma_{r^*+1}(A^*) = \cdots = \sigma_m(A^*) = 0,$$

the zero singular values form one repeated block. Perturbation theory for clustered singular values implies that

$$\big(\sigma_{r^*+1}(A^* + t_n H_n), \ldots, \sigma_m(A^* + t_n H_n)\big) = t_n\, \sigma(U_0^\top H_n V_0) + o(t_n),$$

where $\sigma(B) = (\sigma_1(B), \ldots, \sigma_{m-r^*}(B))$ denotes the ordered singular value vector of $B$ (see Lemma A.15).

Therefore,

$$\frac{1}{t_n}\begin{pmatrix} \sigma_{r^*+1}(A^* + t_n H_n) \\ \vdots \\ \sigma_m(A^* + t_n H_n) \end{pmatrix} = \sigma(U_0^\top H_n V_0) + o(1).$$

Since $H_n \to H$ and using continuity of singular values,

$$\sigma(U_0^\top H_n V_0) \to \sigma(U_0^\top H V_0).$$

Combining all coordinates jointly,

$$\frac{\Phi(A^* + t_n H_n) - \Phi(A^*)}{t_n} \to \begin{pmatrix} (u_1^*)^\top H v_1^* \\ \vdots \\ (u_{r^*}^*)^\top H v_{r^*}^* \\ \sigma_1(U_0^\top H V_0) \\ \vdots \\ \sigma_{m-r^*}(U_0^\top H V_0) \end{pmatrix} := \Phi'_{A^*}(H).$$

Hence, $\Phi'_{A^*}(H)$ is the Hadamard directional derivative. Finally, the map $H \mapsto \Phi'_{A^*}(H)$ is continuous because matrix multiplication and singular values are continuous mappings. Therefore $\Phi$ is Hadamard directionally differentiable at $A^*$. ■

**Lemma B.5** (Existence of First Order Perturbation for singular vectors)**.** *Let $A \in \mathbf{R}^{p\times T}$ and fix $j$ such that its singular value $\sigma_j := \sigma_j(A) > 0$ is simple for all $j = 1, .., r^*$ and $\sigma_j = 0$ for all $j > r^*$. Let $\boldsymbol{u}_j \in \mathbf{R}^p$ and $\boldsymbol{v}_j \in \mathbf{R}^T$ be corresponding left and right unit singular vectors satisfying*

$$A\boldsymbol{v}_j = \sigma_j \boldsymbol{u}_j, \qquad A^\top \boldsymbol{u}_j = \sigma_j \boldsymbol{v}_j.$$

*Let $A_n = A + t_n H_n$ with $t_n \downarrow 0$ and $H_n \to H$. Suppose $(\boldsymbol{u}_j^{(n)}, \boldsymbol{v}_j^{(n)}, \sigma_j^{(n)})$ denote the $j$-th singular triplet of $A_n$, chosen with a consistent sign normalization. Then there exist vectors $\dot{\boldsymbol{u}}_j \in \mathbf{R}^p$, $\dot{\boldsymbol{v}}_j \in \mathbf{R}^T$ such that*

$$\boldsymbol{u}_j^{(n)} = \boldsymbol{u}_j^* + t_n \dot{\boldsymbol{u}}_j + o(t_n), \quad \boldsymbol{v}_j^{(n)} = \boldsymbol{v}_j^* + t_n \dot{\boldsymbol{v}}_j + o(t_n). \tag{B.3}$$

**Proof:** We proceed directly from the defining singular vector equations. For each $n$,

$$A_n \boldsymbol{v}_j^{(n)} = \sigma_j^{(n)} \boldsymbol{u}_j^{(n)}, \qquad A_n^\top \boldsymbol{u}_j^{(n)} = \sigma_j^{(n)} \boldsymbol{v}_j^{(n)}.$$

Subtract the population relations and divide by $t_n$:

$$\begin{aligned} A^* \frac{\boldsymbol{v}_j^{(n)} - \boldsymbol{v}_j^*}{t_n} + H_n \boldsymbol{v}_j^{(n)} &= \sigma_j^* \frac{\boldsymbol{u}_j^{(n)} - \boldsymbol{u}_j^*}{t_n} + \frac{\sigma_j^{(n)} - \sigma_j^*}{t_n} \boldsymbol{u}_j^{(n)}, \\ (A^*)^\top \frac{\boldsymbol{u}_j^{(n)} - \boldsymbol{u}_j^*}{t_n} + H_n^\top \boldsymbol{u}_j^{(n)} &= \sigma_j^* \frac{\boldsymbol{v}_j^{(n)} - \boldsymbol{v}_j^*}{t_n} + \frac{\sigma_j^{(n)} - \sigma_j^*}{t_n} \boldsymbol{v}_j^{(n)}. \end{aligned} \tag{B.4}$$

Towards contradiction, we assume that the representation (B.3) does not hold for the sequence $\{(\boldsymbol{u}_j^{(n)}, \boldsymbol{v}_j^{(n)}, \sigma_j^{(n)})\}_{n\geq 1}$. More precisely,

$$\Delta u_n := \frac{\boldsymbol{u}_j^{(n)} - \boldsymbol{u}_j^*}{t_n} \to \dot{\boldsymbol{u}}_j, \quad \Delta v_n := \frac{\boldsymbol{v}_j^{(n)} - \boldsymbol{v}_j^*}{t_n} \to \dot{\boldsymbol{v}}_j, \quad \Delta\sigma_n := \frac{\sigma_j^{(n)} - \sigma_j^*}{t_n} \to \dot{\sigma}_j,$$

doesn't hold as $t_n \downarrow 0$. Therefore, there exist at least two subsequences $n_{k_s}$ for $s = 1, 2$ with two different pair of triplets $(\dot{\boldsymbol{u}}_{j1}, \dot{\boldsymbol{v}}_{j1}, \dot{\sigma}_{j1})$ and $(\dot{\boldsymbol{u}}_{j2}, \dot{\boldsymbol{v}}_{j2}, \dot{\sigma}_{j2})$, such that,

$$\Delta u_{n_{k_s}} \to \dot{\boldsymbol{u}}_{js}, \quad \Delta v_{n_{k_s}} \to \dot{\boldsymbol{v}}_{js}, \quad \Delta\sigma_{n_{k_s}} \to \dot{\sigma}_{js} \quad \text{for all } \ s = 1, 2. \tag{B.5}$$

Now following (B.4), passing to the limit along $n_{k_s}$ and using $\boldsymbol{u}_j^{(n_{k_s})} \to \boldsymbol{u}_j^*$, $\boldsymbol{v}_j^{(n_{k_s})} \to \boldsymbol{v}_j^*$, and $H_{n_{k_s}} \to H$, we obtain

$$\begin{aligned} A^*\dot{\boldsymbol{v}}_{js} - \sigma_j^*\dot{\boldsymbol{u}}_{js} &= \dot{\sigma}_{js}\,\boldsymbol{u}_j^* - H\boldsymbol{v}_j^*, \\ (A^*)^\top\dot{\boldsymbol{u}}_{js} - \sigma_j^*\dot{\boldsymbol{v}}_{js} &= \dot{\sigma}_{js}\,\boldsymbol{v}_j^* - H^\top\boldsymbol{u}_j^*. \end{aligned} \tag{B.6}$$

From $(\boldsymbol{u}_j^{(n_{k_s})})^\top\boldsymbol{u}_j^{(n_{k_s})} = 1$ and substituting $\boldsymbol{u}_j^{(n_{k_s})} = \boldsymbol{u}_j^* + t_{n_{k_s}}\Delta u_{n_{k_s}}$ gives $(\boldsymbol{u}_j^*)^\top\Delta u_{n_{k_s}} = -\frac{t_{n_{k_s}}}{2}\|\Delta u_{n_{k_s}}\|^2$. Since $\Delta u_{n_{k_s}}$ is bounded and $t_{n_{k_s}} \downarrow 0$ this will imply for some $C > 0$,

$$\left|(\boldsymbol{u}_j^*)^\top\Delta u_{n_{k_s}}\right| \leq \frac{t_{n_{k_s}}}{2}C^2 \to 0.$$

Along the subsequence $\Delta u_{n_{k_s}} \to \dot{\boldsymbol{u}}_{js}$, continuity of the inner product implies

$$(\boldsymbol{u}_j^*)^\top\Delta u_{n_{k_s}} \to (\boldsymbol{u}_j^*)^\top\dot{\boldsymbol{u}}_{js}.$$

Combining the two limits gives $(\boldsymbol{u}_j^*)^\top\dot{\boldsymbol{u}}_{js} = 0$ Similar argument will give us that $(\boldsymbol{v}_j^*)^\top\dot{\boldsymbol{v}}_{js} = 0$ for all $s = 1, 2$. So any subsequential limit will be orthogonal to $\dot{\boldsymbol{u}}_{js}$ and $\dot{\boldsymbol{v}}_{js}$ respectively. Define

$$\delta u := \dot{\boldsymbol{u}}_{j1} - \dot{\boldsymbol{u}}_{j2}, \quad \delta v := \dot{\boldsymbol{v}}_{j1} - \dot{\boldsymbol{v}}_{j2}, \quad \delta\sigma := \dot{\sigma}_{j1} - \dot{\sigma}_{j2}.$$

Subtracting the limit equations (B.6) yields

$$A^*\delta v - \sigma_j^*\delta u = \delta\sigma\,\boldsymbol{u}_j^*, \ \ (A^*)^\top\delta u - \sigma_j^*\delta v = \delta\sigma\,\boldsymbol{v}_j^*,$$

with $(\boldsymbol{u}_j^*)^\top\delta u = 0, \ (\boldsymbol{v}_j^*)^\top\delta v = 0$. Multiplying the first equation by $(\boldsymbol{u}_j^*)^\top$ and using $(\boldsymbol{u}_j^*)^\top A^* = \sigma_j^*(\boldsymbol{v}_j^*)^\top$ gives, $\delta\sigma = \sigma_j^*(\boldsymbol{v}_j^*)^\top\delta v$. Similarly, multiplying the second equation by $(\boldsymbol{v}_j^*)^\top$ yields $\delta\sigma = \sigma_j^*(\boldsymbol{u}_j^*)^\top\delta u = 0$. Hence $\delta\sigma = 0$. Therefore,

$$A^*\delta v = \sigma_j^*\delta u, \ (A^*)^\top\delta u = \sigma_j^*\delta v.$$

Thus $(\delta u, \delta v)$ is a singular vector pair of $A^*$ associated with $\sigma_j^*$. Since $\sigma_j^*$ is simple, its singular subspace is one-dimensional, spanned by $(\boldsymbol{u}_j^*, \boldsymbol{v}_j^*)$. Together with the orthogonality conditions, this implies $\delta u = 0, \ \delta v = 0$. Hence all subsequential limits must coincide. This will imply that,

$$\frac{\boldsymbol{u}_j(A^* + t_nH_n) - \boldsymbol{u}_j^*}{t_n} \to \dot{\boldsymbol{u}}, \qquad \frac{\boldsymbol{v}_j(A^* + t_nH_n) - \boldsymbol{v}_j^*}{t_n} \to \dot{\boldsymbol{v}}.$$

Rewriting the convergence of difference quotients gives

$$\boldsymbol{u}_j^{(n)} = \boldsymbol{u}_j^* + t_n\dot{\boldsymbol{u}}_j + o(t_n), \quad \boldsymbol{v}_j^{(n)} = \boldsymbol{v}_j^* + t_n\dot{\boldsymbol{v}}_j + o(t_n).$$

The proof is complete. ■

**Lemma B.6** (First-order perturbation expansion for singular vectors)**.** *Let*

$$A = U\Sigma V^\top = \begin{bmatrix} U_1 & U_0 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_0 \end{bmatrix}^\top,$$

*where* $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_{r^*})$*, with* $\sigma_1 > \cdots > \sigma_{r^*} > 0$*. Suppose* $A_n = A + t_n H_n,\ t_n \downarrow 0,\ H_n \to H$*. For each* $j = 1, \ldots, r^*$*, let* $\tilde{u}_j^{(n)},\ \tilde{v}_j^{(n)}$ *denote the left and right singular vectors corresponding to* $\sigma_j(A_n)$*. Define,*

$$s_{j,n} = \begin{cases} 1, & \tilde{\boldsymbol{u}}_j^{(n)\top} \boldsymbol{u}_j \geq 0, \\ -1, & \tilde{\boldsymbol{u}}_j^{(n)\top} \boldsymbol{u}_j < 0. \end{cases}$$

*and define the aligned singular vectors:* $\boldsymbol{u}_j^{(n)} = s_{j,n}\tilde{\boldsymbol{u}}_j^{(n)},\quad \boldsymbol{v}_j^{(n)} = s_{j,n}\tilde{\boldsymbol{v}}_j^{(n)}$*. Assume* $\|A_n - A\|_{op} = O(t_n)$*. Then for every* $j = 1, \ldots, r^*$*,*

$$\boldsymbol{u}_j^{(n)} = \boldsymbol{u}_j^* + t_n\, \Gamma_j^{(u)}(H_n) + o(t_n),\ \boldsymbol{v}_j^{(n)} = \boldsymbol{v}_j^* + t_n\, \Gamma_j^{(v)}(H_n) + o(t_n),$$

*where*

$$\Gamma_j^{(u)}(H) = \sum_{k \neq j}^{r^*} \frac{\sigma_j^* (u_k^*)^\top H v_j^* + \sigma_k^* (u_j^*)^\top H v_k^*}{(\sigma_j^*)^2 - (\sigma_k^*)^2}\, u_k^* + \sum_{k=r^*+1}^{p} \frac{(u_k^*)^\top H v_j^*}{\sigma_j^*}\, u_k^*,$$

$$\Gamma_j^{(v)}(H) = \sum_{k \neq j}^{r^*} \frac{\sigma_k^* (u_k^*)^\top H v_j^* + \sigma_j^* (u_j^*)^\top H v_k^*}{(\sigma_j^*)^2 - (\sigma_k^*)^2}\, v_k^* + \sum_{k=r^*+1}^{T} \frac{(u_j^*)^\top H v_k^*}{\sigma_j^*}\, v_k^*.$$

*Therefore, the maps* $A \mapsto \boldsymbol{u}_j(A)$ *and* $A \mapsto \boldsymbol{v}_j(A)$ *are Hadamard directionally differentiable at* $A^*$ *for all* $j = 1, .., r^*$*. In fact, the maps are Hadamard differentiable since corrsponding singular values are simple.*

**Proof**: Fix $j \in \{1, \ldots, r^*\}$. Let $(\tilde{\boldsymbol{u}}_j^{(n)}, \tilde{\boldsymbol{v}}_j^{(n)}, \sigma_j^{(n)} := \sigma_j(A_n))$ denote the singular vectors and singular value triplet corresponding to the $j$-th singular component of $A_n$. Recall the random sign $s_{j,n}$ and the aligned singular vectors $\boldsymbol{u}_j^{(n)}, \boldsymbol{v}_j^{(n)}$. We have the singular vector relations:

$$A_n \boldsymbol{v}_j^{(n)} = \sigma_j^{(n)} \boldsymbol{u}_j^{(n)},\quad A_n^\top \boldsymbol{u}_j^{(n)} = \sigma_j^{(n)} \boldsymbol{v}_j^{(n)}.$$

Since $\sigma_j^*$ is simple and strictly positive, standard singular-vector perturbation theory together with Wedin-type $\sin\Theta$ bounds (see proof of Theorem 4.1) implies

$$\|\boldsymbol{u}_j^{(n)} - \boldsymbol{u}_j^*\|_2 = O(\|A_n - A^*\|_{op}),\ \ \|\boldsymbol{v}_j^{(n)} - \boldsymbol{v}_j^*\|_2 = O(\|A_n - A^*\|_{op}).$$

Because $\|A_n - A^*\|_{op} = O(t_n)$, we apply Lemma B.5 to obtain that,

$$\boldsymbol{u}_j^{(n)} = \boldsymbol{u}_j^* + t_n \dot{\boldsymbol{u}}_j + o(t_n),\ \boldsymbol{v}_j^{(n)} = \boldsymbol{v}_j^* + t_n \dot{\boldsymbol{v}}_j + o(t_n),\ \sigma_j^{(n)} = \sigma_j^* + t_n \dot{\sigma}_j + o(t_n).$$

Now substituting $A_n = A^* + t_n H_n$ into $A_n \boldsymbol{v}_j^{(n)} = \sigma_j^{(n)} \boldsymbol{u}_j^{(n)}$, gives

$$(A^* + t_n H_n)(\boldsymbol{v}_j^* + t_n \dot{\boldsymbol{v}}_j + o(t_n)) = (\sigma_j^* + t_n \dot{\sigma}_j + o(t_n))(\boldsymbol{u}_j^* + t_n \dot{\boldsymbol{u}}_j + o(t_n)).$$

Using the identity $A^* \boldsymbol{v}_j^* = \sigma_j^* \boldsymbol{u}_j^*$, dividing both side by $t_n$ and collecting first-order terms ,

$$H_n \boldsymbol{v}_j^* + A^* \dot{\boldsymbol{v}}_j = \dot{\sigma}_j\, \boldsymbol{u}_j^* + \sigma_j^* \dot{\boldsymbol{u}}_j + o(1).$$

Passing to the limit using $H_n \to H$, we get $H \boldsymbol{v}_j^* + A^* \dot{\boldsymbol{v}}_j = \dot{\sigma}_j\, \boldsymbol{u}_j^* + \sigma_j^* \dot{\boldsymbol{u}}_j$. Similarly, substituting into $A_n^\top \boldsymbol{u}_j^{(n)} = \sigma_j^{(n)} \boldsymbol{v}_j^{(n)}$, yields $H^\top \boldsymbol{u}_j^* + A^{*\top} \dot{\boldsymbol{u}}_j = \dot{\sigma}_j\, \boldsymbol{v}_j^* + \sigma_j^* \dot{\boldsymbol{v}}_j$. Next

we use the normalization constraints $\|\boldsymbol{u}_j^{(n)}\|_2 = 1$, $\|\boldsymbol{v}_j^{(n)}\|_2 = 1$ and using the same arguments of Lemma B.5, we obtain

$$(\boldsymbol{u}_j^*)^\top \dot{\boldsymbol{u}}_j = 0, \quad (\boldsymbol{v}_j^*)^\top \dot{\boldsymbol{v}}_j = 0.$$

Since $\{\boldsymbol{u}_k^*\}_{k=1}^p$ and $\{\boldsymbol{v}_k^*\}_{k=1}^T$ form orthonormal bases, the orthogonality relations imply

$$\dot{\boldsymbol{u}}_j = \sum_{k\neq j} \alpha_{kj} \boldsymbol{u}_k^*, \quad \dot{\boldsymbol{v}}_j = \sum_{k\neq j} \beta_{kj} \boldsymbol{v}_k^*.$$

Project the first perturbation equation onto $\boldsymbol{u}_k^*$, for $k \neq j$:

$$(\boldsymbol{u}_k^*)^\top H \boldsymbol{v}_j^* + (\boldsymbol{u}_k^*)^\top A^* \dot{\boldsymbol{v}}_j = \sigma_j^* \alpha_{kj}.$$

Since $A^* \boldsymbol{v}_k^* = \sigma_k^* \boldsymbol{u}_k^*$, we obtain $(\boldsymbol{u}_k^*)^\top A^* \dot{\boldsymbol{v}}_j = \sigma_k^* \beta_{kj}$. Hence

$$(\boldsymbol{u}_k^*)^\top H \boldsymbol{v}_j^* + \sigma_k^* \beta_{kj} = \sigma_j^* \alpha_{kj}.$$

Similarly, projecting the second perturbation equation onto $\boldsymbol{v}_k^*$ we get,

$$(\boldsymbol{u}_j^*)^\top H \boldsymbol{v}_k^* + \sigma_k^* \alpha_{kj} = \sigma_j^* \beta_{kj}.$$

Therefore we have,

$$\begin{aligned} \sigma_j^* \alpha_{kj} - \sigma_k^* \beta_{kj} &= (\boldsymbol{u}_k^*)^\top H \boldsymbol{v}_j^*, \\ -\sigma_k^* \alpha_{kj} + \sigma_j^* \beta_{kj} &= (\boldsymbol{u}_j^*)^\top H \boldsymbol{v}_k^*. \end{aligned}$$

Solving these resulting linear system yields

$$\alpha_{kj} = \frac{\sigma_j^* (\boldsymbol{u}_k^*)^\top H \boldsymbol{v}_j^* + \sigma_k^* (\boldsymbol{u}_j^*)^\top H \boldsymbol{v}_k^*}{(\sigma_j^*)^2 - (\sigma_k^*)^2}, \quad \beta_{kj} = \frac{\sigma_k^* (\boldsymbol{u}_k^*)^\top H \boldsymbol{v}_j^* + \sigma_j^* (\boldsymbol{u}_j^*)^\top H \boldsymbol{v}_k^*}{(\sigma_j^*)^2 - (\sigma_k^*)^2}.$$

For $k > r^*$, we have $\sigma_k^* = 0$. Hence

$$\alpha_{kj} = \frac{(\boldsymbol{u}_k^*)^\top H \boldsymbol{v}_j^*}{\sigma_j^*}, \qquad \beta_{kj} = \frac{(\boldsymbol{u}_j^*)^\top H \boldsymbol{v}_k^*}{\sigma_j^*}.$$

Substituting these coefficients into the basis expansions for $\dot{\boldsymbol{u}}_j$ and $\dot{\boldsymbol{v}}_j$ gives

$$\dot{\boldsymbol{u}}_j = \Gamma_j^{(u)}(H), \quad \dot{\boldsymbol{v}}_j = \Gamma_j^{(v)}(H).$$

Therefore,

$$\boldsymbol{u}_j^{(n)} = \boldsymbol{u}_j^* + t_n \Gamma_j^{(u)}(H_n) + o(t_n), \quad \boldsymbol{v}_j^{(n)} = \boldsymbol{v}_j^* + t_n \Gamma_j^{(v)}(H_n) + o(t_n).$$

Since $H \longmapsto a^\top H b$ is a bounded linear functional on $(\mathbb{R}^{p\times T}, \|\cdot\|_F)$ for every fixed vectors $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^T$, and the coefficients appearing in the definitions of $\Gamma_j^{(u)}$ and $\Gamma_j^{(v)}$ are fixed finite constants (under the assumption that the nonzero singular values are distinct), it follows that both

$$\Gamma_j^{(u)} : \mathbb{R}^{p\times T} \to \mathbb{R}^p, \qquad \Gamma_j^{(v)} : \mathbb{R}^{p\times T} \to \mathbb{R}^T$$

are continuous linear maps. This completes the proof. ■

**Remark B.2.** *Singular vectors are only identifiable up to a simultaneous sign change: if $(u_j, v_j)$ is a singular vector pair corresponding to a singular value $\sigma_j$, then $(-u_j, -v_j)$ is also a valid singular vector pair. Consequently, the raw singular vectors $\tilde{u}_j^{(n)}$ and $\tilde{v}_j^{(n)}$ need not converge even when $A_n \to A^*$, since numerical or analytical constructions may flip their signs arbitrarily with $n$. The random sign $s_{j,n}$ is therefore introduced to align the*

*orientations of the estimated singular vectors with the population singular vectors. This guarantees*

$$u_j^{(n)\top} u_j^* \geq 0, \qquad v_j^{(n)\top} v_j^* \geq 0,$$

*and makes the perturbation expansions well-defined. Without this alignment, first-order expansions of singular vectors are not meaningful because the sequence of singular vectors may oscillate between two opposite directions.*

## APPENDIX C. PROOF OF MAIN RESULTS

C.1. **Proof of Theorem 3.1.** Define the map $Z_n(B) = \frac{1}{n}\|Y - XB\|_F^2 + \frac{\lambda_n}{n}\|\sigma(B)\|_1$. It is clear that $\tilde{A}_n$ is the minimizer of $Z_n(\cdot)$.

**Convexity.** The map $B \mapsto Z_n(B)$ is convex. Indeed, write $g_1 : [0,\infty) \to [0,\infty)$ as $g_1(x) = x^2$, and write $g_2(B) = \|Y - XB\|_F$. Then $g_1$ is convex and non-decreasing on $[0,\infty)$, and $g_2$ is a norm and therefore convex. Hence the composition $g_1 \circ g_2$ is convex (see Section 3.2.4 of Boyd and Vandenberghe (2004)). Moreover,it's easy to verify that $B \mapsto \|\sigma(B)\|_1$ is a norm and therefore convex. Hence $Z_n$ is a sum of convex maps (multiplied by nonnegative constants), and is therefore convex.

**Pointwise convergence of $Z_n(B)$.** We write

$$\begin{aligned}&\frac{1}{n}\|Y - XB\|_F^2 \\ &= \operatorname{tr}\left((B-A)^\top \frac{X^\top X}{n}(B-A)\right) - 2\operatorname{tr}\left((B-A)^\top \frac{X^\top E}{n}\right) + \operatorname{tr}\left(\frac{E^\top E}{n}\right). \end{aligned} \tag{C.1}$$

Assumption (C.3) and Lemma A.3 imply that, for every fixed matrix $B$,

$$\operatorname{tr}\left((B-A)^\top \frac{X^\top X}{n}(B-A)\right) \xrightarrow{P} \operatorname{tr}\big((B-A)^\top C(B-A)\big),$$

and Lemma A.11 will imply that

$$\operatorname{tr}\left((B-A)^\top \frac{X^\top E}{n}\right) \xrightarrow{P} 0.$$

Also, by the WLLN, $\operatorname{tr}\left(\frac{E^\top E}{n}\right) \xrightarrow{P} T\sigma^2$. Using $n^{-1}\lambda_n \to \lambda_0$, we obtain for each fixed $B$,

$$Z_n(B) \xrightarrow{P} Z(B) + T\sigma^2.$$

**Limit objective is strictly convex.** Assumption (C.3) implies that

$$B \mapsto \operatorname{tr}\big((B-A)^\top C(B-A)\big)$$

is strictly convex, since $C$ is positive definite. Note that $B \mapsto \lambda_0\|\sigma(B)\|_1$ is convex due to Lemma A.1. Hence $B \mapsto Z(B)$ is strictly convex and since $\arg\min_{B\in\mathbf{R}^{p\times T}}[Z(B) + T\sigma^2] = \arg\min_{B\in\mathbf{R}^{p\times T}} Z(B)$, therefore the limiting minimizer is unique. We now apply Lemma A.5, which guarantees convergence of the minimizers of $Z_n$ to the minimizer of $Z$, under convexity and pointwise convergence on a dense set. Therefore,

$$\tilde{A}_n \xrightarrow{P} \arg\min_B Z(B).$$

The proof is complete. ∎

C.2. **Proof of Theorem 3.2.** Recall that, $\sqrt{n}\,(\tilde{A}_n - A) = \arg\min_{U \in \mathbf{R}^{p\times T}} H_n(U)$, where,

(C.2)
$$H_n(U) = -2\,\mathrm{tr}(U^\top W_n) + \mathrm{tr}(U^\top C_n U) + \frac{\lambda_n}{\sqrt{n}}\left(\frac{\|\sigma(A + n^{-1/2}U)\|_1 - \|\sigma(A)\|_1}{n^{-1/2}}\right).$$

As seen above from the expression of $H(U)$ in (3.1), the minimizer of $H(U)$ will be unique since $U \mapsto \big(2\,\mathrm{tr}(U^\top W) + \mathrm{tr}(U^\top C U)\big)$ will be strictly convex due to assumption (C.3). The third term of $H(U)$ is linear in $U$ and the last term of $H(U)$ is convex. It must be noted that $U \mapsto \left\|\boldsymbol{\sigma}\Big(U_0^\top U V_0\Big)\right\|_1$ may not be a linear map in $U$. Again due to Lemma A.11, we have that $\boldsymbol{W}_n \xrightarrow{d} \boldsymbol{W}$ in $\mathbf{R}^{p\times T}$, where $W$ is defined there. As a result, using the continuous mapping theorem and Slutsky's theorem it follows that

$$-2\,\mathrm{tr}\big(U^\top \boldsymbol{W}_n\big) \;+\; \mathrm{tr}\big(U^\top C_n U\big) \;\xrightarrow{d}\; -2\,\mathrm{tr}(U^\top \boldsymbol{W}) + \mathrm{tr}(U^\top C U),$$

for any fixed matrix $U$. Using similar arguments, one can show that for any fixed integer $k \geq 1$ and fixed matrices $U_1, \ldots, U_k \in \mathbf{R}^{p\times T}$, the above distributional convergence holds jointly. Consider the last term of $H_n(U)$, which involves $\lambda_n$. Using Lemma A.7, we can show that,

$$\frac{\lambda_n}{\sqrt{n}}\left(\frac{\|\sigma(A + n^{-1/2}U)\|_1 - \|\sigma(A)\|_1}{n^{-1/2}}\right) \longrightarrow\; \lambda_0 \sum_{j=1}^{r^*} \big(\boldsymbol{u}_j{}^\top U\, \boldsymbol{v}_j\big) + \;\lambda_0 \left\|\boldsymbol{\sigma}\Big(U_0^\top U V_0\Big)\right\|_1,$$

where $(\boldsymbol{u}_j,\ \boldsymbol{v}_j)$ is a left-right singular vector pair of $A$ corresponding to $\sigma_j$. This implies that $H_n(U) \xrightarrow{d} H(U)$ for each fixed $U$, and also for finitely many choices of $U_1, \ldots, U_k$, for any $k \geq 1$. In order to complete the proof we need to verify stochastic equicontinuity of $\{H_n(U)\}_{n\geq 1}$ on compact sets. That is, for every $\varepsilon, \eta, M > 0$, there exists a $\delta(\varepsilon, \eta, M) > 0$ such that for large enough $n$,

(C.3)
$$\mathbf{P}\left(\sup_{\substack{\|\boldsymbol{U}' - \boldsymbol{U}\|_F < \delta \\ \max\{\|\boldsymbol{U}'\|_F, \|\boldsymbol{U}\|_F\} < M}} \big|H_n(\boldsymbol{U}') - H_n(\boldsymbol{U})\big| > \eta\right) < \varepsilon.$$

Now we see that,

$$\begin{aligned} &H_n(\boldsymbol{U}') - H_n(\boldsymbol{U}) \\ &= -2\,\mathrm{tr}\Big[(\boldsymbol{U}' - \boldsymbol{U})^\top W_n\Big] + \Big[\mathrm{tr}\Big\{(\boldsymbol{U}' - \boldsymbol{U})^\top C_n (\boldsymbol{U}' - \boldsymbol{U})\Big\} + 2\mathrm{tr}\Big\{(\boldsymbol{U}' - \boldsymbol{U})^\top C_n \boldsymbol{U}\Big\}\Big] \\ &\quad + \frac{\lambda_n}{\sqrt{n}}\left(\frac{\|\sigma(A + n^{-1/2}\boldsymbol{U}')\|_1 - \|\sigma(A + n^{-1/2}\boldsymbol{U})\|_1}{n^{-1/2}}\right) \end{aligned} \tag{C.4}$$

Now for any $X, Y \in \mathbf{R}^{p\times T}$, applying Lemma A.2 and Cauchy-Schwarz inequality, $\big|\|X\|_* - \|Y\|_*\big| \leq \sum_{j=1}^m |\sigma_j(X) - \sigma_j(Y)| \leq \sqrt{m}\,\|X - Y\|_F$, where $m = \min\{p, T\}$. Now set $X = A + n^{-1/2}\boldsymbol{U}'$ and $Y = A + n^{-1/2}\boldsymbol{U}$. Then

$$\left|\left\|\boldsymbol{\sigma}\Big(A + n^{-1/2}\boldsymbol{U}'\Big)\right\|_1 - \left\|\boldsymbol{\sigma}\Big(A + n^{-1/2}\boldsymbol{U}\Big)\right\|_1\right| \;\leq\; n^{-1/2}\sqrt{m}\,\|\boldsymbol{U}' - \boldsymbol{U}\|_F. \tag{C.5}$$

Now denote $\gamma_1$ as the largest eigen value of the matrix $C$. Then due to assumption (C.3) and basic operator norm bound, we have;

$$\left|\mathrm{tr}\Big\{(\boldsymbol{U}'-\boldsymbol{U})^\top C_n(\boldsymbol{U}'-\boldsymbol{U})\Big\}\right| \le 2\gamma_1\|\boldsymbol{U}'-\boldsymbol{U}\|_F^2, \tag{C.6}$$

$$\left|\mathrm{tr}\Big\{(\boldsymbol{U}'-\boldsymbol{U})^\top C_n\boldsymbol{U}\Big\}\right| \le 2\gamma_1\|\boldsymbol{U}'-\boldsymbol{U}\|_F\|\boldsymbol{U}\|_F \tag{C.7}$$

Lastly, $\big|\mathrm{tr}\big[(\boldsymbol{U}'-\boldsymbol{U})^\top W_n\big]\big| = \big|\langle \boldsymbol{U}'-\boldsymbol{U},\, W_n\rangle_F\big| \le \|\boldsymbol{U}'-\boldsymbol{U}\|_F\,\|W_n\|_F$. Now due to Lemma A.11, for every $\varepsilon>0$, there exists $0<B':=B'(\varepsilon)<\infty$ such that,

$$\mathbf{P}\big(\|\boldsymbol{W}_n\|_F > B'(\varepsilon)\big) < \varepsilon. \tag{C.8}$$

Hence combining (C.5), (C.6) we write,

(C.9)
$$\sup_{\substack{\|\boldsymbol{U}'-\boldsymbol{U}\|_F<\delta\\ \max\{\|\boldsymbol{U}'\|_F,\|\boldsymbol{U}\|_F\}<M}} \big|H_n(\boldsymbol{U}')-H_n(\boldsymbol{U})\big| \le 2\delta\|W_n\|_F + 2\gamma_1\delta^2 + 2\gamma_1\delta M + 2\lambda_0\delta m^{1/2}$$

Now considering $\delta = \frac{1}{2}\min\left\{\sqrt{\frac{\eta}{8\gamma_1}}, \frac{\eta}{8[M\gamma_1+\lambda_0\sqrt{m}]}, \frac{\eta}{4B'(\varepsilon)}\right\}$, and from (C.8) we can conclude that for large enough $n$, (C.3) is true. Since every matrix valued map can be equivalently expressed in terms of vectorization operator, we can apply Lemma A.13 for $vec(\boldsymbol{U}^{p\times T}) \in \mathbf{R}^{pT}$ and the proof is complete. ■

C.3. **Proof of Theorem 3.3.** Let $S$ be a perturbation matrix and consider $A+S$. If $\sigma_j$ is a simple singular value of $A$ with left and right singular vectors $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$, then

$$\sigma_j(A+S) = \sigma_j + (\boldsymbol{u}_j)^\top S\,\boldsymbol{v}_j + O(\|S\|^2). \tag{C.10}$$

Next we verify, Hadamard differentiability of the map $\phi_j : \mathbf{R}^{p\times T} \to \mathbf{R}$, $\phi_j(A) = \sigma_j(A)$, at $A$ tangentially to $\mathbf{R}^{p\times T}$ with derivative $\phi_j'(A)[H] = (\boldsymbol{u}_j)^\top H\boldsymbol{v}_j$, in the sense of definition 1. Let $t_n \to 0$ and $H_n \to H$ in $\mathbf{R}^{p\times T}$. Define, $S_n = t_nH_n$. Substituting this into (C.10) gives

$$\frac{\sigma_j(A+t_nH_n)-\sigma_j}{t_n} = (\boldsymbol{u}_j)^\top H_n\boldsymbol{v}_j + O(t_n\|H_n\|^2).$$

Since $H_n \to H$ in the Frobenius norm and $t_n \to 0$, the remainder term vanishes and we obtain

$$\frac{\sigma_j(A+t_nH_n)-\sigma_j}{t_n} \longrightarrow (\boldsymbol{u}_j)^\top H\boldsymbol{v}_j.$$

It remains to verify that the derivative $H \mapsto (\boldsymbol{u}_j)^\top H\boldsymbol{v}_j$ is a continuous linear map on $\mathbf{R}^{p\times T}$ under the Frobenius norm. Linearity is immediate. For continuity, note that

$$(\boldsymbol{u}_j)^\top H\boldsymbol{v}_j = \mathrm{tr}\big(\boldsymbol{v}_j(\boldsymbol{u}_j)^\top H\big) \implies \big|(\boldsymbol{u}_j)^\top H\boldsymbol{v}_j\big| \le \|\boldsymbol{v}_j(\boldsymbol{u}_j)^\top\|_F\,\|H\|_F \le \|H\|_F,$$

Since $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are unit vectors, $\|\boldsymbol{v}_j(\boldsymbol{u}_j)^\top\|_F = 1$. Thus the derivative is a bounded linear functional and hence continuous. This verifies the Hadamard differentiability of $\phi_j$ at $A$.

In our setting, $S_n = \tilde{A}_n - A$. Since Theorem 3.2 shows that $\sqrt{n}(\tilde{A}_n - A) \xrightarrow{d} U^*$, we can write $S_n = \frac{1}{\sqrt{n}}\sqrt{n}(\tilde{A}_n - A)$, $t_n := n^{-1/2}$, $H_n := \sqrt{n}(\tilde{A}_n - A)$. Substituting $S_n$

into (C.10) gives $\sigma_j(\tilde{A}_n) = \sigma_j + \frac{1}{\sqrt{n}}(\boldsymbol{u}_j)^\top \big[\sqrt{n}(\tilde{A}_n - A)\big]\boldsymbol{v}_j + O_p(n^{-1})$. Then we can apply functional delta method for all $1 \le j \le r^*$ such that we have,

$$\sqrt{n}\big(\sigma_j(\tilde{A}_n) - \sigma_j^*\big) \xrightarrow{d} (\boldsymbol{u}_j)^\top U^* \boldsymbol{v}_j,$$

where $U^* = \arg\min_{U \in \mathbf{R}^{p\times T}} H(U)$ is the limit in Theorem 3.2. Next we will address asymptotic behavior of singular values corresponding to the zero block. Let the singular value decomposition of $A$ be written in block form as

$$A = U\,\Sigma\,V^\top = \begin{bmatrix} U_1 & U_0 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_0 \end{bmatrix}^\top,$$

where, $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_{r^*})$, $U_0 \in \mathbf{R}^{p\times(p-r^*)}$, $V_0 \in \mathbf{R}^{T\times(T-r^*)}$, $m = \min(p, T)$, $\sigma_{\min}(\Sigma_1) = \sigma_{r^*} > 0$. Define the rotated matrix $Z_n := U^\top \tilde{A}_n V$ and $\Delta_n := U^\top[n^{1/2}(\tilde{A}_n - A)]V$. Then

$$\begin{aligned} n^{1/2} Z_n = n^{1/2} U^\top \tilde{A}_n V &= n^{1/2} U^\top A V + n^{1/2} U^\top (\tilde{A}_n - A) V \\ &= n^{1/2}\Sigma + \Delta_n = \underbrace{\begin{pmatrix} n^{1/2}\Sigma_1 + \Delta_{n,11} & \Delta_{n,10} \\ \Delta_{n,01} & \Delta_{n,00} \end{pmatrix}}_{=:M_n \ (say)}, \end{aligned}$$

with, $\Delta_{n,11} \in \mathbf{R}^{r^*\times r^*}, \Delta_{n,10} \in \mathbf{R}^{r^*\times(T-r^*)}, \Delta_{n,01} \in \mathbf{R}^{(p-r^*)\times r^*}, \Delta_{n,00} \in \mathbf{R}^{(p-r^*)\times(T-r^*)}$. Set $B_n := n^{1/2}\Sigma_1 + \Delta_{n,11}, C_n := \Delta_{n,10}, D_n := \Delta_{n,01}, E_n := \Delta_{n,00}$. Now due to Theorem 3.2 and since orthogonal transformations preserve the Frobenius norm, therefore $\|\Delta_n\|_F = O_p(1)$. More importantly, $\|\Delta_{n,11}\|_F$, $\|\Delta_{n,10}\|_F$, $\|\Delta_{n,01}\|_F$, $\|\Delta_{n,00}\|_F = O_p(1)$. Since $B_n = n^{1/2}\Sigma_1 + \Delta_{n,11}$, $\sigma_{r^*} = \sigma_{min}(\Sigma_1) > 0$, therefore for large enough $n$ and for some constant $c_1 > 0$, $\sigma_{\min}(B_n) \ge \sigma_{\min}(n^{1/2}\Sigma_1) - \|\Delta_{n,11}\| = n^{1/2}\sigma_{r^*} - c_1 \xrightarrow{p} \infty$. Hence $B_n$ is invertible with probability tending to one. Now applying Lemma A.15 to $M_n$, for $k = 1, \ldots, m - r^*$,

$$\sigma_{r^*+k}(M_n) = \sigma_{r^*+k}(n^{1/2}Z_n) = n^{1/2}\sigma_{r^*+k}(Z_n) = \sigma_k(E_n) + o_p(1).$$

Recall that, $E_n = \Delta_{n,00} = U_0^\top[n^{1/2}(\tilde{A}_n - A)]V_0$. From Theorem 3.2, we have $E_n = U_0^\top\big[\sqrt{n}(\tilde{A}_n - A)\big]V_0 \xrightarrow{d} U_0^\top U^* V_0$. Since singular values are continuous functions of the underlying matrices, we then have for all $k = 1, ..., m - r^*$,

$$\sigma_k(E_n) = \sigma_k\Big(U_0^\top\big[\sqrt{n}(\tilde{A}_n - A)\big]V_0\Big) \xrightarrow{d} \sigma_k(U_0^\top U^* V_0).$$

Now orthogonal rotations preserve singular values, so $n^{1/2}\sigma_j(\tilde{A}_n) = \sigma_j(M_n)$. Combining the above results and applying Slutsky's theorem, for $k = 1, \ldots, m - r^*$,

$$\sqrt{n}\,\sigma_{r^*+k}(\tilde{A}_n) \xrightarrow{d} \sigma_k(U_0^\top U^* V_0).$$

Therefore the proof is complete. ■

C.4. **Proof of Proposition 4.1.** Define the event $\Omega_0 := \Big\{\|\tilde{A}_n - A\|_F = o\big(n^{-1/2}\log n\big)\Big\}$. Note that $\mathbf{P}(\Omega_0) = 1$ due to Lemma A.8. Fix $\omega \in \Omega_0$ throughout the argument. All limits below are therefore deterministic limits along this fixed sample path. By Weyl's inequality for singular values,

$$|\tilde{\sigma}_j^{(n)}(\omega) - \sigma_j| \le \|\tilde{A}_n(\omega) - A\|_{op} \le \|\tilde{A}_n(\omega) - A\|_F.$$

Hence, $\max_{1\le j\le m}|\tilde{\sigma}_j^{(n)}(\omega)-\sigma_j| = o\big(n^{-1/2}\log n\big)$. Since $\dfrac{n^{-1/2}\log n}{a_n}\to 0$, it follows that

$$\max_{1\le j\le m}|\tilde{\sigma}_j^{(n)}(\omega)-\sigma_j| = o(a_n). \tag{C.11}$$

Since $\sigma_1 > \cdots > \sigma_{r^*} > 0$ and $a_n \to 0$, there exists $N_1(\omega)$ such that $a_n < \frac{\sigma_{r^*}}{2}$ for all $n \ge N_1(\omega)$. From (C.11), there exists $N_2(\omega)$ such that for all $n \ge N_2(\omega)$, $\max_{1\le j\le r^*}|\tilde{\sigma}_j^{(n)}(\omega)-\sigma_j| < \frac{\sigma^*_{r^*}}{2}$. Let $N_{12}(\omega) := \max\{N_1(\omega), N_2(\omega)\}$. Then for all $n \ge N_{12}(\omega)$ and $1 \le j \le r^*$, $\tilde{\sigma}_j^{(n)}(\omega) \ge \sigma_j - \frac{\sigma_{r^*}}{2} \ge \frac{\sigma_{r^*}}{2} > a_n$. Hence, $\tilde{\sigma}_j^{(n)}(\omega) > a_n \quad \forall j \le r^*, \quad \forall n \ge N_{12}(\omega)$. Thus no non-zero singular value of $\tilde{A}_n(\omega)$ is removed by thresholding for sufficiently large $n$. On the other hand, For $j > r^*$, $\sigma_j = 0$. From (C.11), there exists $N_3(\omega)$ such that for all $n \ge N_3(\omega)$, $\max_{j>r^*}|\tilde{\sigma}_j^{(n)}(\omega)| < \frac{a_n}{2}$. Hence, for all $n \ge N_3(\omega)$ and all $j > r^*$, $\tilde{\sigma}_j^{(n)}(\omega) < a_n$. Thus no spurious singular value of $\tilde{A}_n(\omega)$ exceeds the threshold eventually. Let $N(\omega) := \max\{N_{12}(\omega), N_3(\omega)\}$. Then for all $n \ge N(\omega)$,

$$\tilde{\sigma}_j^{(n)}(\omega) > a_n \quad \text{for } j \le r^*, \text{ and } \tilde{\sigma}_j^{(n)}(\omega) \le a_n \quad \text{for } j > r^*.$$

Therefore, $\operatorname{rank}\big(\tilde{A}_n^{(\mathrm{thr})}(\omega)\big) = r^*\ \forall n \ge N(\omega)$. Hence for every $\omega \in \Omega_0$,

$$\operatorname{rank}(\tilde{A}_n^{(\mathrm{thr})}(\omega)) \ne r^* \quad \text{only finitely many times.}$$

Since $\mathbf{P}(\Omega_0) = 1$, we conclude

$$\mathbf{P}\big(\operatorname{rank}(\tilde{A}_n^{(\mathrm{thr})}) \ne r^* \text{ infinitely often}\big) = 0.$$

The proof is complete. ∎

C.5. **Proof of Theorem 4.1.** First we define the set,

$$\begin{aligned}\mathbf{B} = \Big\{\ \|\tilde{\boldsymbol{A}}_n - \boldsymbol{A}\|_F = o\big(n^{-1/2}\log n\big)\ \Big\} \cap \Big\{\operatorname{rank}(\tilde{A}_n^{(thr)}) = r^*\Big\} \\ \cap\ \Big\{\ \mathcal{L}(\bar{\boldsymbol{W}}^*_{n,G}\mid \mathcal{E}) \xrightarrow{d^*} \mathcal{MN}_{p\times T}\big(\mathbf{0}_{p\times T},\ U=\sigma^2\boldsymbol{C},\ V=\boldsymbol{I}_T\big)\ \Big\}.\end{aligned}$$

Note that $\mathbf{P}(\mathbf{B}) = 1$ due to Lemma A.8, A.12 and Proposition 4.1. Therefore it's enough to show for every $\omega \in \mathbf{B}$,

$$\hat{F}_n^{(\mathrm{thr})}(\omega,\cdot) \xrightarrow{\mathrm{d}_*} F_\infty(\omega,\cdot). \tag{C.12}$$

Fix an arbitrary $\omega \in \mathbf{B}$ and suppress the dependence on $\omega$ in the notation. Recall that,

$$\sqrt{n}\,(\tilde{A}_n^{*(mod)} - \tilde{A}_n^{(\mathrm{thr})}) = \arg\min_{V\in\mathbf{R}^{p\times T}} \check{H}_n^*(V,\cdot),$$

where

$$\begin{aligned}\check{H}_n^*(V,\cdot) = \Bigg[ &-2\operatorname{tr}\big(V^\top \bar{\boldsymbol{W}}^*_{n,G}(\omega,\cdot)\big) + \operatorname{tr}\big(V^\top C_n V\big) \\ &+ \frac{\lambda_n}{\sqrt{n}}\left(\frac{\|\sigma(\tilde{A}_n^{(thr)} + n^{-1/2}V)\|_1 - \|\sigma(\tilde{A}_n^{(thr)})\|_1}{n^{-1/2}}\right)\Bigg]\end{aligned} \tag{C.13}$$

For that every $\omega \in \mathbf{B}$, we will first show that;

$$\mathcal{L}\Big(\check{H}_n^*(\omega,V,\cdot)\mid\mathcal{E}\Big) \xrightarrow{d} \mathcal{L}\Big(F_\infty(\omega,V)\Big) \tag{C.14}$$

and consequently the argmin laws converge. Denote by $\tilde{U}_n = (\tilde{U}_{1,n}, \tilde{U}_{0,n})$ and $\tilde{V}_n = (\tilde{V}_{1,n}, \tilde{V}_{0,n})$ the left/right singular vector matrices of $\tilde{A}_n^{(\mathrm{thr})}$ with block sizes $r^*$ and $m - r^*$ (on the rank-identification event which has probability tending to one).

**Step 1: convergence of the smooth part.**

Consider the smooth portion of $\tilde{H}_n^*$ as: $S_n(V) := -2\,\mathrm{tr}(V^\top \bar{W}_{n,G}^*) + \mathrm{tr}(V^\top C_n V)$. On the set $\mathbf{B}$ we have (pathwise), $\bar{\boldsymbol{W}}_{n,G}^* \xrightarrow{d^*} \boldsymbol{W}$. Concerning the linear term, the bootstrap law of $-2\,\mathrm{tr}(V^\top \bar{\boldsymbol{W}}_{n,G}^*)$ converges (conditionally) to the law of $-2\,\mathrm{tr}(V^\top \boldsymbol{W})$ by Lemma A.12 on $\mathbf{B}$. Because $\mathrm{tr}(V^\top \cdot)$ is continuous in the matrix argument and $C_n \to C$ in Frobenius norm on $\mathbf{B}$, we have pathwise (for this fixed $\omega$)

$$S_n(V, \omega, \cdot) \xrightarrow{d_*} S_\infty(\cdot) \quad \text{where, } S_\infty(V) := -2\,\mathrm{tr}(V^\top \boldsymbol{W}) + \mathrm{tr}(V^\top C V)$$

More precisely, the finite-dimensional distributions of $(S_n(V_1, \omega), \ldots, S_n(V_k, \omega))$ converge to those of $(S_\infty(V_1), \ldots, S_\infty(V_k))$ for any fixed $V_1, \ldots, V_k$.

**Step 2: Convergence of the top $r^*$ singular values of the thresholded estimator.**

Recall that, for some fixed $\omega \in \mathbf{B}$, there exists $N(\omega)$ such that, $\tilde{\sigma}_{j,n}^{(thr)}(\omega) = \tilde{\sigma}_j^{(n)}(\omega)$ for all $n \geq N(\omega)$ whenever $1 \leq j \leq r^*$. Since $\sigma_j > 0$ and $\tilde{\sigma}_j^{(n)}(\omega) = \sigma_j + o(1)$ due to Lemma A.8 on the set $\mathbf{B}$ for all $j \leq r^*$, consequently these top $r^*$ singular values survive thresholding and $\tilde{A}_n^{(\mathrm{thr})}$ has exactly $r^*$ positive singular values for large $n$ (on $\mathbf{B}$).

**Step 3: Convergence of the top $r^*$ singular vectors of the thresholded estimator.** Note that, $\{\tilde{\boldsymbol{u}}_j^{(n)}, \tilde{\boldsymbol{v}}_j^{(n)}\}_{j=1}^m$ be the left-right singular vector pairs for $\tilde{A}_n^{(thr)}$. Therefore by Corollary A.1 and for all $n \geq N(\omega)$ and on the set $\mathbf{B}$ we have,

$$\sin\theta\big(\tilde{\boldsymbol{u}}_j^{(n)}, \boldsymbol{u}_j\big) \;\leq\; \frac{5\sigma_1 \|E_n\|_{\mathrm{op}}}{\delta} \qquad \text{for all} \quad j = 1, \ldots, r^* \qquad \text{with } E_n = \tilde{A}_n^{(thr)} - A.$$

Let $\boldsymbol{u}, \tilde{\boldsymbol{u}} \in \mathbf{R}^d$ be unit vectors and let $\theta$ be the principal angle between them, so that $\cos\theta = \tilde{\boldsymbol{u}}^\top \boldsymbol{u}$. Decompose $\tilde{\boldsymbol{u}}$ into its projection onto $\boldsymbol{u}$ and its orthogonal residual: $\tilde{\boldsymbol{u}} = (\tilde{\boldsymbol{u}}^\top \boldsymbol{u})\boldsymbol{u} + \boldsymbol{w}$, $\boldsymbol{w} := \tilde{\boldsymbol{u}} - (\tilde{\boldsymbol{u}}^\top \boldsymbol{u})\boldsymbol{u}$, $\boldsymbol{w} \perp \boldsymbol{u}$. Since both vectors have unit norm, $1 = \|\tilde{\boldsymbol{u}}\|_2^2 = (\tilde{\boldsymbol{u}}^\top \boldsymbol{u})^2 + \|\boldsymbol{w}\|_2^2$. Hence $\|\boldsymbol{w}\|_2 = \sqrt{1 - (\tilde{\boldsymbol{u}}^\top \boldsymbol{u})^2} = \sqrt{1 - \cos^2\theta} = \sin\theta$. Because $\boldsymbol{w} = \tilde{u} - (\tilde{\boldsymbol{u}}^\top \boldsymbol{u})\boldsymbol{u}$, we obtain the identity $\sin\theta(\tilde{\boldsymbol{u}}, \boldsymbol{u}) = \|\tilde{\boldsymbol{u}} - (\tilde{\boldsymbol{u}}^\top \boldsymbol{u})\boldsymbol{u}\|_2$. Since $\sin\theta(\tilde{\boldsymbol{u}}, \boldsymbol{u}) = \|\tilde{\boldsymbol{u}} - (\tilde{\boldsymbol{u}}^\top \boldsymbol{u})\boldsymbol{u}\|_2$, the earlier Wedin type bound implies for any $\omega \in \mathbf{B}$,

$$\big\|\tilde{\boldsymbol{u}}_j^{(n)} - (\tilde{\boldsymbol{u}}_j^{(n)\top} \boldsymbol{u}_j)\,\boldsymbol{u}_j\big\|_2 = O\Big(\|E_n\|_{op}\Big).$$

Because both vectors are unit-norm, the scalar inner product satisfies $1 - \big(\tilde{\boldsymbol{u}}_j^{(n)\top} \boldsymbol{u}_j\big)^2 = O(\|E_n\|_{op}^2)$, which implies $|\tilde{\boldsymbol{u}}_j^{(n)\top} \boldsymbol{u}_j| = \sqrt{1 - O(\|E_n\|_{op}^2)} = \Big[1 - \frac{1}{2}O(\|E_n\|_{op}^2) + O(\|E_n\|_{op}^4)\Big]$. Define,

$$s_{j,n} = \begin{cases} 1, & \tilde{\boldsymbol{u}}_j^{(n)\top} \boldsymbol{u}_j \geq 0, \\ -1, & \tilde{\boldsymbol{u}}_j^{(n)\top} \boldsymbol{u}_j < 0. \end{cases}$$

so that $s_{j,n}\,\tilde{\boldsymbol{u}}_j^{(n)\top} \boldsymbol{u}_j \;\geq\; 0$. Then $\|\tilde{\boldsymbol{u}}_j^{(n)} - s_{j,n}\boldsymbol{u}_j\|_2^2 = \|s_{j,n}\tilde{\boldsymbol{u}}_j^{(n)} - \boldsymbol{u}_j\|_2^2$. Expanding the square and using $\|\boldsymbol{u}_j\|_2 = \|\tilde{\boldsymbol{u}}_j^{(n)}\|_2 = 1$ and $s_{j,n}^2 = 1$,

$$\|s_{j,n}\tilde{\boldsymbol{u}}_j^{(n)} - \boldsymbol{u}_j\|_2^2 = 2 - 2\,s_{j,n}\tilde{\boldsymbol{u}}_j^{(n)\top} \boldsymbol{u}_j.$$

By the sign choice, $s_{j,n}\,\tilde{\boldsymbol{u}}_j^{(n)\top}\boldsymbol{u}_j = \big|\tilde{\boldsymbol{u}}_j^{(n)\top}\boldsymbol{u}_j\big|$. Thus we have, $\|\tilde{\boldsymbol{u}}_j^{(n)} - s_{j,n}\boldsymbol{u}_j\|_2^2 = 2 - 2\big|\tilde{\boldsymbol{u}}_j^{(n)\top}\boldsymbol{u}_j\big|$. Using the expansion above on the set $\mathbf{B}$,

$$\|\tilde{\boldsymbol{u}}_j^{(n)} - s_{j,n}\boldsymbol{u}_j\|_2 = O(\|E_n\|_{op}). \tag{C.15}$$

Applying the same argument to $\tilde{A}_n^\top = A^\top + E_n^\top$ gives

$$\big\|\tilde{\boldsymbol{v}}_j^{(n)} - s_{j,n}\boldsymbol{v}_j\big\|_2 = O\Big(\|E_n\|_{op}\Big). \tag{C.16}$$

Now fix any matrix $V \in \mathbf{R}^{p\times T}$. Writing $\tilde{\boldsymbol{u}}_j^{(n)} = s_{j,n}\boldsymbol{u}_j + \boldsymbol{r}_{u,n}$ and $\tilde{\boldsymbol{v}}_j^{(n)} = s_{j,n}\boldsymbol{v}_j + \boldsymbol{r}_{v,n}$ with $\boldsymbol{r}_{u,n}, \boldsymbol{r}_{v,n}$ satisfying (C.15) and (C.16), we obtain

$$\tilde{\boldsymbol{u}}_j^{(n)\top} V\,\tilde{\boldsymbol{v}}_j^{(n)} = (\boldsymbol{u}_j)^\top V\boldsymbol{v}_j + (\boldsymbol{u}_j)^\top V\boldsymbol{r}_{v,n} + \boldsymbol{r}_{u,n}^\top V\boldsymbol{v}_j + \boldsymbol{r}_{u,n}^\top V\boldsymbol{r}_{v,n}.$$

Each error term is bounded by $\|V\|_{op}$ times the corresponding vector norms, giving

$$\big|\tilde{\boldsymbol{u}}_j^{(n)\top} V\,\tilde{\boldsymbol{v}}_j^{(n)} - (\boldsymbol{u}_j)^\top V\boldsymbol{v}_j\big| \le 2\|V\|_{op}\|E_n\|_{op} + \|V\|_{op}\|E_n\|_{op}^2.$$

Since due to Lemma A.9, $\|E_n\|_{op} = o(1)$ on the set $\mathbf{B}$ then for any fixed matrix $V$, the right-hand side is $o(1)$ pathwise (for $\omega \in \mathbf{B}$) for large enough $n$.

**Step 4: Convergence of the lower $m - r^*$ adjusted singular values.** Recall that, for some fixed $\omega \in \mathbf{B}$, there exists $N(\omega)$ such that, $\tilde{\sigma}_{j,n}^{(thr)}(\omega) = 0$ for all $n \ge N(\omega)$ whenever $r^* < j \le m$. Since $\sigma_j = 0$ and $\tilde{\sigma}_j^{(n)}(\omega) = O(n^{-1/2})$ on the set $\mathbf{B}$ for all $j > r^*$, consequently these lower $m - r^*$ singular values do not survive thresholding and $\tilde{A}_n^{(\mathrm{thr})}$ has exactly $m - r^*$ zero singular values for large $n$ (on $\mathbf{B}$).

**Step 5: Convergence of the lower $m - r^*$ associated singular vectors.** Recall that, $m := \min(p, T)$. Define the signal and null singular-vector blocks by $U = \begin{pmatrix} U_1 & U_0 \end{pmatrix}$, $V = \begin{pmatrix} V_1 & V_0 \end{pmatrix}$, where

$$U_1 = \big(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{r^*}\big), \quad U_0 = \big(\boldsymbol{u}_{r^*+1}, \ldots, \boldsymbol{u}_p\big),$$

and similarly $V_1 = \big(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{r^*}\big)$, $V_0 = \big(\boldsymbol{v}_{r^*+1}, \ldots, \boldsymbol{v}_T\big)$. Likewise let $\widetilde{U}_n = \begin{pmatrix} \widetilde{U}_{1,n} & \widetilde{U}_{0,n} \end{pmatrix}$, $\widetilde{V}_n = \begin{pmatrix} \widetilde{V}_{1,n} & \widetilde{V}_{0,n} \end{pmatrix}$, where $\widetilde{U}_{0,n} = \big(\tilde{\boldsymbol{u}}_{r^*+1}^{(n)}, \ldots, \tilde{\boldsymbol{u}}_p^{(n)}\big)$, and similarly for $\widetilde{V}_{0,n}$. For the associated orthogonal projectors define:

$$P_1 = U_1U_1^\top, \;\; \widetilde{P}_{1,n} = \widetilde{U}_{1,n}\widetilde{U}_{1,n}^\top, \;\; P_0 = U_0U_0^\top, \;\; \widetilde{P}_{0,n} = \widetilde{U}_{0,n}\widetilde{U}_{0,n}^\top.$$

Since $P_0 = I - P_1$, $\widetilde{P}_{0,n} = I - \widetilde{P}_{1,n}$, it follows that $\|\widetilde{P}_{0,n} - P_0\|_F = \|\widetilde{P}_{1,n} - P_1\|_F$. By Step 3 on $\mathbf{B}$, for each $j = 1, \ldots, r^*$,

$$\begin{aligned}
\|\tilde{P}_{1,j,n} - P_{1,j}\|_F^2 &= \big\|\tilde{\boldsymbol{u}}_j^{(n)}\,\tilde{\boldsymbol{u}}_j^{(n)\top} - \boldsymbol{u}_j(\boldsymbol{u}_j)^\top\big\|_F^2 \\
&= \mathrm{tr}\Big[\big(\tilde{\boldsymbol{u}}_j^{(n)}\,\tilde{\boldsymbol{u}}_j^{(n)\top} - \boldsymbol{u}_j(\boldsymbol{u}_j)^\top\big)^\top\big(\tilde{\boldsymbol{u}}_j^{(n)}\,\tilde{\boldsymbol{u}}_j^{(n)\top} - \boldsymbol{u}_j(\boldsymbol{u}_j)^\top\big)\Big] \\
&= \mathrm{tr}\big(\tilde{\boldsymbol{u}}_j^{(n)}\,\tilde{\boldsymbol{u}}_j^{(n)\top}\tilde{\boldsymbol{u}}_j^{(n)}\,\tilde{\boldsymbol{u}}_j^{(n)\top}\big) + \mathrm{tr}\big(\boldsymbol{u}_j\boldsymbol{u}_j^\top\boldsymbol{u}_j\boldsymbol{u}_j^\top\big) - 2\,\mathrm{tr}\big(\tilde{\boldsymbol{u}}_j^{(n)}\,\tilde{\boldsymbol{u}}_j^{(n)\top}\boldsymbol{u}_j\boldsymbol{u}_j^\top\big) \\
&= 2[1 - (\tilde{\boldsymbol{u}}_j^{(n)\top}\boldsymbol{u}_j)^2] = O(\|E_n\|_{op}^2),
\end{aligned} \tag{C.17}$$

using $\boldsymbol{u}_j^\top\boldsymbol{u}_j = \tilde{\boldsymbol{u}}_j^{(n)\top}\tilde{\boldsymbol{u}}_j^{(n)} = 1$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. Therefore, $\|\widetilde{P}_{1,j,n} - P_{1,j}\|_F = O(\|E_n\|_{\mathrm{op}})$, where $E_n = \tilde{A}_n^{(\mathrm{thr})} - A$. Since $P_1 = \sum_{j=1}^{r^*} P_{1,j}$, $\widetilde{P}_{1,n} = \sum_{j=1}^{r^*} \tilde{P}_{1,j,n}$, the triangle inequality yields

$$\|\widetilde{P}_{1,n} - P_1\|_F = \left\|\sum_{j=1}^{r^*}(\tilde{P}_{1,j,n} - P_{1,j})\right\|_F \le \sum_{j=1}^{r^*}\|\tilde{P}_{1,j,n} - P_{1,j}\|_F = O(\|E_n\|_{op}).$$

Hence on $\mathbf{B}$, we have $\|\widetilde{P}_{0,n} - P_0^*\|_F = O(\|E_n\|_{\mathrm{op}})$. Applying the same argument to the right singular subspaces yields $\|\widetilde{V}_{0,n}\widetilde{V}_{0,n}^\top - V_0V_0^\top\|_F = O(\|E_n\|_{\mathrm{op}})$. Let $k := p - r^*$. Since $U_0$ and $\widetilde{U}_{0,n}$ both have orthonormal columns, $U_0^\top U_0 = I_k,\ \widetilde{U}_{0,n}^\top\widetilde{U}_{0,n} = I_k$. Consider the matrix

$$M_n := U_0^\top \widetilde{U}_{0,n} \text{ with its singular value decomposition } M_n = L_n\Sigma_n K_n^\top.$$

Define, $Q_n := L_nK_n^\top$. Then $Q_n$ is orthogonal. We claim that $\|\widetilde{U}_{0,n} - U_0Q_n\|_F^2 = 2\ \mathrm{tr}(I_k - \Sigma_n)$. Indeed,

$$\begin{aligned}\|\widetilde{U}_{0,n} - U_0Q_n\|_F^2 &= \mathrm{tr}\Big[(\widetilde{U}_{0,n} - U_0Q_n)^\top(\widetilde{U}_{0,n} - U_0Q_n)\Big] \\ &= \mathrm{tr}(I_k) + \mathrm{tr}(I_k) - 2\ \mathrm{tr}\big(Q_n^\top U_0^\top \widetilde{U}_{0,n}\big) \\ &= 2k - 2\ \mathrm{tr}\big(K_nL_n^\top L_n\Sigma_nK_n^\top\big) = 2\ \mathrm{tr}(I_k - \Sigma_n).\end{aligned}$$

Now we see that,

$$\begin{aligned}&\|P_0 - \widetilde{P}_{0,n}\|_F^2 = \|U_0U_0^\top - \widetilde{U}_{0,n}\widetilde{U}_{0,n}^\top\|_F^2 \\ &= 2k - 2\ \mathrm{tr}\Big(U_0U_0^\top\widetilde{U}_{0,n}\widetilde{U}_{0,n}^\top\Big) = 2k - 2\ \mathrm{tr}\Big(M_nM_n^\top\Big) = 2\sum_{i=1}^{k}(1 - \sigma_i(M_n)^2).\end{aligned}$$

Since both $U_0$ and $\widetilde{U}_{0,n}$ have orthonormal columns, Hence, for any vector $x$,

$$\|M_nx\|_2 = \|U_0^\top\widetilde{U}_{0,n}x\|_2 \le \|\widetilde{U}_{0,n}x\|_2 = \|x\|_2.$$

Therefore $\|M_n\|_{\mathrm{op}} \le 1$. Since every singular value is bounded above by the operator norm, $0 \le \sigma_i(M_n) \le 1$. Consequently,

$$1 - \sigma_i(M_n) = \frac{1 - \sigma_i(M_n)^2}{1 + \sigma_i(M_n)} \le 1 - \sigma_i(M_n)^2,$$

because $1 + \sigma_i(M_n) \ge 1$. Therefore we obtain,

$$\mathrm{tr}(I_k - \Sigma_n) = \sum_{i=1}^{k}\Big[1 - \sigma_i(M_n)\Big] \le \frac{1}{2}\|P_0^* - \widetilde{P}_{0,n}\|_F^2.$$

Therefore $\|\widetilde{U}_{0,n} - U_0Q_n\|_F^2 \le \|P_0 - \widetilde{P}_{0,n}\|_F^2$, and hence

$$\|\widetilde{U}_{0,n} - U_0Q_n\|_F \le \|P_0 - \widetilde{P}_{0,n}\|_F = O(\|E_n\|_{\mathrm{op}}).$$

Exactly the same argument applied to the right singular subspaces yields an orthogonal matrix $R_n$ such that $\|\widetilde{V}_{0,n} - V_0R_n\|_F = O(\|E_n\|_{\mathrm{op}})$. Therefore we can write, $\widetilde{U}_{0,n} = U_0Q_n + Z_n,\ \widetilde{V}_{0,n} = V_0R_n + J_n$, such that $\|Z_n\|_F,\ \|J_n\|_F = O(\|E_n\|_{\mathrm{op}})$. Now fix any matrix $V \in \mathbf{R}^{p\times T}$. Using these arguments,

$$\widetilde{U}_{0,n}^\top V\widetilde{V}_{0,n} = Q_n^\top U_0^\top VV_0R_n + \underbrace{Q_n^\top U_0^\top VJ_n + Z_n^\top VV_0R_n + Z_n^\top VJ_n}_{:=\Delta_n \text{ (say)}}.$$

To bound $\Delta_n$, note that $\|Q_n\|_{\mathrm{op}} = \|R_n\|_{\mathrm{op}} = \|U_0\|_{\mathrm{op}} = \|V_0\|_{\mathrm{op}} = 1$. Using $\|ABC\|_F \le \|A\|_{\mathrm{op}}\|B\|_F\|C\|_{\mathrm{op}}$, we have,

$$\begin{aligned}\|\Delta_n\|_F &\le \|V\|_F\|J_n\|_F + \|V\|_F\|Z_n\|_F + \|V\|_F\|Z_n\|_F\|J_n\|_F \\ &= O\big[\|V\|_F\|E_n\|_{\mathrm{op}} + \|V\|_F\|E_n\|_{\mathrm{op}}^2\big].\end{aligned}$$

The nuclear norm is continuous and invariant under left and right orthogonal transformations. Denote the nuclear norm of a matrix $D$ as $\|D\|_* = \|\boldsymbol{\sigma}(D)\|_1$. Therefore for any $\|V\|_F \leq R$ using reverse triangle inequality for norms,

$$\begin{aligned}\Big|\|\widetilde{U}_{0,n}^\top V \widetilde{V}_{0,n}\|_* - \|(U_0)^\top V V_0\|_*\Big| &= \Big|\|Q_n^\top (U_0)^\top V V_0 R_n + \Delta_n\|_* - \|Q_n^\top (U_0)^\top V V_0 R_n\|_*\Big| \\ &\leq \|\Delta_n\|_* \leq \sqrt{m}\|\Delta_n\|_F = O\big[\|V\|_F \|E_n\|_{\rm op} + \|V\|_F \|E_n\|_{\rm op}^2\big] \\ &= o(1), \text{ on the set } \mathbf{B}.\end{aligned}$$

Thus we obtain that,

$$\|\widetilde{U}_{0,n}^\top U \widetilde{V}_{0,n}\|_* = \|(U_0)^\top U V_0\|_* + o(1). \tag{C.18}$$

**Step 6: Matching the original limit.** Recall that penalty part from (C.13) and due to step (2) and (4); apply Lemma A.7 with $\gamma = n^{-1/2} \downarrow 0$ to have,

$$\lim_{\gamma \downarrow 0} \frac{\|\sigma(\tilde{A}_n^{(\rm thr)} + \gamma V)\|_1 - \|\sigma(\tilde{A}_n^{(\rm thr)})\|_1}{\gamma} = \sum_{j=1}^{r^*} \tilde{\boldsymbol{u}}_j^{(n)\top} V\, \tilde{\boldsymbol{v}}_j^{(n)} + \big\|\tilde{U}_{0,n}^\top V\, \tilde{V}_{0,n}\big\|_*. \tag{C.19}$$

We now bound the point wise difference:

$$\begin{aligned}&\left|\sum_{j=1}^{r^*} \tilde{\boldsymbol{u}}_j^{(n)\top} V\, \tilde{\boldsymbol{v}}_j^{(n)} + \big\|\tilde{U}_{0,n}^\top V\, \tilde{V}_{0,n}\big\|_* - \sum_{j=1}^{r^*} \big(\boldsymbol{u}_j{}^\top V\, \boldsymbol{v}_j\big) - \Big\|U_0^\top U V_0\Big\|_*\right| \\ &= O\big[\|V\|_F \|E_n\|_{\rm op} + \|V\|_F \|E_n\|_{\rm op}^2\big] = o(1), \text{ on the set } \mathbf{B}.\end{aligned}$$

Combining step 1-6, this completes the proof for (C.14). The argmin convergence follows similarly as the steps of Theorem 3.2 and Lemma A.13. Therefore the proof is complete. ■

## REFERENCES


1. ANDERSON, T. W. (2003) An Introduction to Multivariate Statistical Analysis. 3rd ed. Wiley.
2. BACH, F. R. (2008) Consistency of trace norm minimization. The Journal of Machine Learning Research. 9. 1019–1048.
3. BECK, A. & TEBOULLE, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences. 2(1). 183–202.
4. BECK, A. (2017) First-Order Methods in Optimization. SIAM.
5. BOYD, S. & VANDENBERGHE, L. (2004) Convex Optimization. Cambridge University Press.
6. BREIMAN, L. & FRIEDMAN, J. H. (1997) Predicting multivariate responses in multiple linear regression. Journal of the American Statistical Association. 92(437). 131–142.
7. BUNEA, F., SHE, Y. & WEGKAMP, M. H. (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. The Annals of Statistics. 39(2). 1282–1309.
8. CAI, J.-F., CANDÈS, E. J. & SHEN, Z. (2010) A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization. 20(4). 1956–1982.
9. CANDÈS, E. & RECHT, B. (2012) Exact matrix completion via convex optimization. Communications of the ACM. 55(6). 111–119.
10. CHATTERJEE, A. & LAHIRI, S. N. (2010). *Asymptotic properties of the residual bootstrap for lasso estimators*. *Proceedings of the American Mathematical Society*. **138(12)**, 4497–4509.

11. CHATTERJEE, A. & LAHIRI, S. N. (2011) Bootstrapping lasso estimators. Journal of the American Statistical Association. 106(494). 608–625.
12. CHATTERJEE, S. (2015) Matrix estimation by universal singular value thresholding. The Annals of Statistics. 43(1). 177–214.
13. CHEN, L. & HUANG, J. Z. (2012) Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. Journal of the American Statistical Association. 107(500). 1533–1545.
14. CHEN, K., DONG, H. & CHAN, K.-S. (2013) Reduced rank regression via adaptive nuclear norm penalization. Biometrika. 100(4). 901–920.
15. CHOUDHURY, M. & DAS, D. (2026) Bootstrapping lasso in generalized linear models. Journal of Statistical Planning and Inference. 246. 106428.
16. DAVIS, C. & KAHAN, W. M. (1970) The rotation of eigenvectors by a perturbation. III. SIAM Journal on Numerical Analysis. 7(1). 1–46.
17. DAVIES, P. T. & TSO, M. K. S. (1982) Procedures for reduced-rank regression. Journal of the Royal Statistical Society: Series C (Applied Statistics). 31(3). 244–255.
18. DÜMBGEN, L. (1993) On nondifferentiable functions and the bootstrap. Probability Theory and Related Fields. 95(1). 125–140.
19. ECK, D. J. (2018) Bootstrapping for multivariate linear regression models. Statistics & Probability Letters. 134. 141–149.
20. FANG, Z. & SANTOS, A. (2019) Inference on directionally differentiable functions. The Review of Economic Studies. 86(1). 377–412.
21. FREEDMAN, D. A. (1981) Bootstrapping regression models. The Annals of Statistics. 9(6). 1218–1228.
22. FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 33(1). 1–22.
23. GIRAUD, C. (2011) Low rank multivariate regression. Electronic Journal of Statistics. 5. 775–799.
24. GIRAUD, C. (2021) Introduction to High-Dimensional Statistics. Chapman and Hall/CRC, New York.
25. GOLDBERG, A., RECHT, B., XU, J., NOWAK, R. & ZHU, J. (2010) Transduction with matrix completion: Three birds with one stone. Advances in Neural Information Processing Systems. 23.
26. HASTIE, T., TIBSHIRANI, R. & WAINWRIGHT, M. (2015) Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton.
27. HORN, R. A. & JOHNSON, C. R. (2013) Matrix Analysis. 2nd ed. Cambridge University Press.
28. IZENMAN, A. J. (1975) Reduced-rank regression for the multivariate linear model. Journal of Multivariate Analysis. 5(2). 248–264.
29. JOHNSON, R. A. & WICHERN, D. W. (2007) Applied Multivariate Statistical Analysis. 6th ed. Pearson.
30. JOSSE, J. & WAGER, S. (2016) Bootstrap-based regularization for low-rank matrix estimation. Journal of Machine Learning Research. 17(124). 1–29.
31. KNIGHT, K. & FU, W. (2000) Asymptotics for Lasso-type estimators. The Annals of Statistics. 28(5). 1356–1378.
32. KOLTCHINSKII, V., LOUNICI, K. & TSYBAKOV, A. B. (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. The Annals of Statistics. 39(5). 2302–2329.

33. MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1979) Multivariate Analysis. Academic Press.
34. MILAN, L. & WHITTAKER, J. (1995) Application of the parametric bootstrap to models that incorporate a singular value decomposition. Journal of the Royal Statistical Society: Series C (Applied Statistics). 44(1). 31–49.
35. MUIRHEAD, R. J. (1982) Aspects of Multivariate Statistical Theory. Wiley.
36. NEGAHBAN, S. & WAINWRIGHT, M. J. (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics. 39(2). 1069–1097.
37. NESTEROV, Y. (2004) Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Boston, MA.
38. PORTIER, F. & DELYON, B. (2014) Bootstrap testing of the rank of a matrix via least-squared constrained estimation. Journal of the American Statistical Association. 109(505). 160–172.
39. REINSEL, G. C. & VELU, R. P. (1998) Multivariate Reduced-Rank Regression: Theory and Applications. Springer.
40. RENCHER, A. C. & CHRISTENSEN, W. F. (2012) Methods of Multivariate Analysis. 3rd ed. Wiley.
41. ROCKAFELLAR, R. T. (1997). *Convex analysis*. vol. **11** Princeton university press.
42. ROHDE, A. & TSYBAKOV, A. B. (2011) Estimation of high-dimensional low-rank matrices. Electronic Journal of Statistics. 5. 887–930.
43. SAEGUSA, T., SUGASAWA, S. & LAHIRI, P. (2022) Parametric bootstrap confidence intervals for the multivariate Fay–Herriot model. Journal of Survey Statistics and Methodology. 10(1). 115–130.
44. SHAPIRO, A. (1991) Asymptotic analysis of stochastic programs. Annals of Operations Research. 30(1). 169–186.
45. STEWART, G. W. (1990) Perturbation theory for the singular value decomposition. UMIACS Technical Report UMIACS-TR-90-124, University of Maryland.
46. TIBSHIRANI, R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B. 58(1). 267–288.
47. VAN DER VAART, A. W. & WELLNER, J. A. (1996) Weak Convergence and Empirical Processes. Springer, New York.
48. WAINWRIGHT, M. J. (2019) High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press.
49. WATSON, G. A. (1992) Characterization of the subdifferential of some matrix norms. Linear Algebra and its Applications. 170. 33–45.
50. WILKS, D. S. (2019) Statistical Methods in the Atmospheric Sciences. 4th ed. Academic Press.
51. YU, Y., WANG, T. & SAMWORTH, R. J. (2015) A useful variant of the Davis–Kahan theorem for statisticians. Biometrika. 102(2). 315–323.
52. ZELLNER, A. (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association. 57(298). 348–368.

DEPARTMENT OF MATHEMATICS, INDIAN INSTITUTE OF TECHNOLOGY BOMBAY, MUMBAI 400076, INDIA
*Email address*: 214090002@iitb.ac.in

DEPARTMENT OF MATHEMATICS, INDIAN INSTITUTE OF TECHNOLOGY BOMBAY, MUMBAI 400076, INDIA
*Email address*: debrajdas@math.iitb.ac.in

STATISTICAL SCIENCES UNIT, INDIAN STATISTICAL INSTITUTE, NEW DELHI 110016, INDIA.
*Email address*: cha@isid.ac.in