Journal of Machine Learning Research 25 (2024) 1-66 Submitted 6/23; Revised 5/24; Published 8/24

A Note on Entrywise Consistency for Mixed-data Matrix Completion

Yunxiao Chen Y.CHEN186@LSE.AC.UK
Department of Statistics
London School of Economics and Political Science
London WC2A 2AE, UK

Xiaoou Li LIXX1766@UMN.EDU
School of Statistics
University of Minnesota
Minneapolis, MN 55455

Editor: Ali Shojaie

Abstract

This note studies matrix completion for a partially observed n by p data matrix involving mixed types of variables (e.g., continuous, binary, ordinal). A general family of non-linear factor models is considered, under which the matrix completion problem becomes the estimation of an n by p low-rank matrix $\mathbf{M}$. For existing methods in the literature, estimation consistency is established by showing $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F/\sqrt{np}$, the scaled Frobenius norm of the difference between the estimated and true $\mathbf{M}$ matrices, converges to zero in probability as n and p grow to infinity. However, this notion of consistency does not guarantee the convergence of each individual entry and, thus, may not be sufficient when specific data entries or the worst-case scenario is of interest. To address this issue, we consider the notion of entrywise consistency based on $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_{\max}$, the max norm of the estimation error matrix. We propose refinement procedures that turn estimators, which are consistent in the Frobenius norm sense, into entrywise estimators through a one-step refinement. Tight probabilistic error bounds are derived for the proposed estimators. The proposed methods are evaluated by simulation studies and real-data applications for collaborative filtering and large-scale educational assessment.

Keywords: Matrix completion; generalized latent factor model; mixed data; entrywise consistency; max norm

1. Introduction

Missing data are commonly encountered in machine learning, especially for large-scale data involving many observations and variables. Matrix completion concerns the prediction of missing entries in a partially observed matrix, which has received wide applications, such as collaborative filtering (Goldberg et al., 1992; Feuerverger et al., 2012), social network recovery (Jayasumana et al., 2019), sensor localization (Biswas et al., 2006), and educational and psychological measurement (Bergner et al., 2022; Chen et al., 2023).

Many matrix completion methods consider real-valued matrices (Candès and Recht, 2009; Candès and Tao, 2010; Keshavan et al., 2010; Klopp, 2014; Koltchinskii et al., 2011; Negahban and Wainwright, 2012; Chen et al., 2020c; Xia and Yuan, 2021). Their theoretical guarantees are typically established under a linear factor model (e.g. Bartholomew et al., 2008), which says the underlying complete data matrix can be decomposed as the sum of a low-rank signal matrix $\mathbf{M}$ and a mean-zero noise matrix. Under this statistical model, the matrix completion task becomes to estimate the signal matrix $\mathbf{M}$ based on the observed data entries. However, many real applications of matrix completion involve mixed types of variables (e.g., continuous, count, binary, ordinal), for which the linear factor model may not be suitable. For example, in survey studies, different questionnaire items may be of different measurement scales – some items may be binary (e.g., yes/no), some may be ordinal (e.g., disagree/neutral/agree), while others may be count variables (e.g., the number of times that one skipped school). Mixed data also appear in multimodal biomedical data, where different types of variables are collected with different technologies (e.g., gene expression, genotype, protein activity). Methods have been developed for matrix completion with specific variable types, such as binary (Cai and Zhou, 2013; Davenport et al., 2014; Han et al., 2020, 2023), categorical (Bhaskar, 2016; Klopp et al., 2015), count (Cao and Xie, 2015; McRae and Davenport, 2021; Robin et al., 2019), and mixed data (Robin et al., 2020). Non-linear factor models, which are extensions of the linear factor model, are typically assumed in these works.

A matrix completion method is typically evaluated by a mean squared error (MSE), defined as $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F^2/(np) = \sum_{i=1}^n \sum_{j=1}^p (\hat{m}_{ij} - m_{ij}^*)^2/(np)$, where $\|\cdot\|_F$ denotes the matrix Frobenius norm, $n \times p$ is the size of the data matrix, and $\hat{\mathbf{M}} = (\hat{m}_{ij})_{n\times p}$ and $\mathbf{M}^* = (m_{ij}^*)_{n\times p}$ are the estimated and true signal matrices, respectively. Probabilistic error bounds have been established for the MSE in the literature (see Chen et al., 2020c; Chen and Li, 2022; Cai and Zhou, 2016, and references therein). Under suitable conditions, these error bounds imply that the MSE decays to zero when both n and p grow to infinity, which is viewed as a notion of statistical consistency for matrix completion. However, this notion of consistency slightly differs from that in our traditional sense; that is, the MSE converging to zero does not imply the convergence of each individual entry, which, however, may be important in some applications which concern the prediction of individual data entries. Entrywise results for matrix completion have been established under linear factor models (Abbe et al., 2020; Chen et al., 2019b, 2020c; Chernozhukov et al., 2023). However, such results are not available for non-linear factor models, and extending these entrywise results to non-linear factor models is non-trivial.

This note considers a general matrix completion problem that allows the variables to be of mixed types. The generalized latent factor model (GLFM; Bartholomew et al., 2008; Skrondal and Rabe-Hesketh, 2004) is a general family of latent variable models that combine factor analysis with generalized linear modelling. By allowing for variable-specific link functions, the GLFM is suitable for modelling multivariate data with mixed types. Under the GLFM framework, we propose two methods that ensure entrywise consistency under dense and sparse missingness settings. Both methods apply to an initial estimate whose MSE converges to zero. They obtain refined estimates by solving some estimating equations constructed based on the initial estimate. The difference between the two methods is that one involves data splitting while the other does not. The two methods have the same asymptotic behavior under a dense setting where the proportion of observed entries does not decay to zero. In that case, their entrywise error rate matches the MSE of the initial estimate up to a logarithm factor, suggesting that there is virtually no loss when performing refinement. However, under a sparse setting where the proportion of observed entries converges to zero, the procedure with data splitting achieves a smaller error rate than the one without data splitting, and the error rate of the data splitting procedure matches the MSE of the initial estimate up to a logarithm factor. To our best knowledge, the current work is the first one obtaining an entrywise consistent estimator for counts and binary data, assuming that the counts and binary data follow the Poisson factor and the multidimensional two-parameter logistic model, respectively. Moreover, it is also the first one for the more general GLFM model for mixed data. Our theoretical analysis further shows that the refined estimator based on a constrained joint maximum likelihood estimator (Chen et al., 2020a) for the GLFM is minimax optimal in an entrywise sense under a suitable asymptotic regime. The proposed methods are evaluated by simulation studies and real-data applications for collaborative filtering and large-scale educational assessment.

The rest of the note is organized as follows. In Section 2, we introduce a generalized latent factor model for matrix completion with mixed data. Section 3 introduces two methods for achieving entrywise consistency. Theoretical guarantees on the proposed methods are established in Section 4. A simulation study is given in Section 5, and two real data examples are given in Section 6. Finally, we conclude with some discussions in Section 7. Additional simulation results and theoretical results, and proofs of the theorems are given in the appendix. The computation code used in Sections 5 and 6 can be found at `https://github.com/yunxiaochen/MatrixCompletion_MixedData`.

2. Mixed-data Matrix Completion

2.1 Notation

For a positive integer n, let $[n] := \{1, \cdots, n\}$ be the set containing all the integers 1, ..., n. Let $\|\mathbf{x}\|$ denote the standard Euclidean norm for a vector $\mathbf{x} = (x_1, ..., x_n)^T$ and $\|\mathbf{x}\|_\infty = \max_i |x_i|$ be the infinity norm (also called the maximum norm) of a vector. For a matrix $\mathbf{X} = (x_{ij})_{n\times m}$, let $\|\mathbf{X}\|_F$, $\|\mathbf{X}\|_*$ and $\|\mathbf{X}\|_2$ denote its Frobenius, nuclear and spectral norms, respectively. We use $\|\mathbf{X}\|_{\max} := \max_{i\in[n],j\in[m]} |x_{ij}|$ to denote the matrix maximum norm, and use $\|\mathbf{X}\|_{2\to\infty} := \sup_{\|\mathbf{u}\|=1} \|\mathbf{X}\mathbf{u}\|_\infty$ to denote the two-to-infinity norm. According to Proposition 6.1, Cape et al. (2019), the two-to-infinity norm is the same as the maximum matrix row norm $\|\mathbf{X}\|_{2\to\infty} = \max_{i\in[n]}(\sum_{j\in[p]} x_{ij}^2)^{1/2}$. For two sequences of real numbers, we write $a_{n,p} \ll b_{n,p}$ (or $a_{n,p} = o(b_{n,p})$) if $\lim_{n,p\to\infty} a_{n,p}/b_{n,p} = 0$, $a_{n,p} \gg b_{n,p}$ if $\lim_{n,p\to\infty} a_{n,p}/b_{n,p} = \infty$, $a_{n,p} \lesssim b_{n,p}$ (or $a_{n,p} = O(b_{n,p})$) if there is a positive constant M independent with n and p, such that $|a_{n,p}| \leq M|b_{n,p}|$, $a_{n,p} \gtrsim b_{n,p}$ if there is a positive constant c independent with n and p, such that $|a_{n,p}| \geq c|b_{n,p}|$, and $a_{n,p} \sim b_{n,p}$ if $b_{n,p} \lesssim a_{n,p} \lesssim b_{n,p}$. For two real numbers x and y, we denote their maximum and minimum as $x \vee y = \max(x, y)$ and $x \wedge y = \min(x, y)$, respectively. We use the standard $O_p(\cdot)$ and $o_p(\cdot)$ notation for stochastic boundedness and convergence in probability, respectively. We use "$\circ$" for the matrix Hadamard (entrywise) product.

2.2 Problem Setup

Consider an $n \times p$ data matrix $\mathbf{Y}$, with the (i, j)th entry denoted by Y_{ij}, for $i \in [n]$ and $j \in [p]$. In the rest, we refer to the rows and columns as the observations and variables, respectively. We do not observe the full matrix due to data missingness. The missing pattern is indicated by an $n \times p$ binary matrix $\mathbf{\Omega} = (\omega_{ij})_{i\in[n],j\in[p]}$, where $\omega_{ij} = 1$ if Y_{ij} is observed and $\omega_{ij} = 0$ if Y_{ij} is missing. Matrix completion concerns inferring the value of Y_{ij} for the missing entries, i.e., entries with $\omega_{ij} = 0$. We consider variables of mixed types, which occurs in many real-world applications; that is, we allow Y_{ij} in different columns to be of mixed types, such as continuous, binary, ordinal, and count variables. Throughout the paper, we make the following assumption on the missing pattern matrix $\mathbf{\Omega}$.

Assumption 1. *The missing indicators, $\omega_{ij}, i \in [n], j \in [p]$, are jointly independent. In addition, $\mathbf{\Omega}$ and $\mathbf{Y}$ are independent.*

2.3 Generalized Latent Factor Model

Additional assumptions are needed for matrix completion, as otherwise, the missing entries can take any feasible values. A typical assumption for matrix completion is a low-rank assumption, i.e., $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, where $\mathbf{M}$ is a low-rank signal matrix, and $\mathbf{E}$ is the noise matrix whose entries are independent and mean-zero. Let the rank of $\mathbf{M}$ be r. Then we can write $\mathbf{Y} = \boldsymbol{\Theta}\mathbf{A}^T + \mathbf{E}$, where $\boldsymbol{\Theta}$ and $\mathbf{A}$ are $n \times r$ and $p \times r$ matrices, respectively. This model is typically known as a linear factor model (e.g. Bartholomew et al., 2008), where $\boldsymbol{\Theta}$ and $\mathbf{A}$ are referred to as the factor-score and loading matrices, respectively. The matrix completion task then becomes an estimation problem, i.e., estimating the signal matrix $\mathbf{M} = \boldsymbol{\Theta}\mathbf{A}^T$ based on the observed data entries.

However, the linear factor model may be restricted when not all variables are continuous. The GLFM is an extension of the linear factor model (Bartholomew et al., 2008; Skrondal and Rabe-Hesketh, 2004). It assumes that entries Y_{ij} are independent, and the probability density function of Y_{ij} (with respect to some baseline measure) takes an exponential family form $f_j(y_{ij}|m_{ij}, \phi_j) = \exp[\phi_j^{-1}\{y_{ij}m_{ij} - b_j(m_{ij})\} + c_j(y_{ij}, \phi_j)]$, where b_j and c_j are pre-specified functions, m_{ij} is the (i, j)th entry of a low-rank signal matrix $\mathbf{M} = \boldsymbol{\Theta}\mathbf{A}^T$ and ϕ_j is a dispersion parameter. The density function depends on variable j so that the variables can be of different types. We give some examples below.

Example 1. *For a continuous variable j, we may assume f_j to be a normal density function, where ϕ_j is the variance, $b_j(m_{ij}) = m_{ij}^2/2$ and $c_j(y_{ij}, \phi_j) = -y_{ij}^2/(2\phi_j) - (\log(2\pi\phi_j))/2$. When all the variables follow this normal model, the data matrix follows a linear factor model.*

Example 2. *Consider a binary or ordinal variable j such that Y_{ij} in $\{0, 1, ..., k_j\}$ for some given $k_j \geq 1$, where $k_j = 1$ and $k_j > 1$ correspond to binary and ordinal variables, respectively. We can assume f_j to follow a Binomial logistic model, for which $\phi_j = 1$, $b_j(m_{ij}) = k_j \log(1 + \exp(m_{ij}))$ and $c_j(y_{ij}, \phi_j) = \log(k_j!) - \log(y_{ij}!) - \log((k_j - y_{ij})!)$. This model has been considered in Masters and Wright (1984) with psychometric applications. When all the variables are binary and follow this logistic model, the data matrix is said to follow a multidimensional two-parameter logistic (M2PL) item response theory model (Reckase, 2009). This model has been considered in Davenport et al. (2014) and Cai and Zhou (2013) for the completion of binary matrices.*

Example 3. *A Poisson model may be assumed for count variables j, for which $\phi_j = 1$, $b_j(m_{ij}) = \exp(m_{ij})$ and $c_j(y_{ij}, \phi_j) = -\log(y_{ij}!)$. When all the variables follow this Poisson model, the joint model for the data matrix is known as a Poisson factor model (Wedel et al., 2003). This Poisson model has been considered in Robin et al. (2019) and Robin et al. (2020) for count data with missing values.*

Under the GLFM, $\mathbb{E}\mathbf{Y} = (b_j'(m_{ij}))_{n\times p}$, where $b_j'(\cdot)$ denotes the derivative of the known function $b_j(\cdot)$. Thus, matrix completion under the GLFM again boils down to estimating the signal matrix $\mathbf{M} = \boldsymbol{\Theta}\mathbf{A}^T$. This estimation problem will be investigated in the rest. We note that a similar GLFM framework has been considered in Robin et al. (2020) for analyzing mixed data with missing values. However, they focused on evaluating the estimation accuracy by the MSE, while our main focus is the entrywise loss.

3. Refined Estimation for Entrywise Consistency

The accuracy in estimating $\mathbf{M}$ is typically measured by the MSE, or equivalently, a scaled Frobenius norm $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F/\sqrt{np}$, where $\mathbf{M}^*$ is the underlying true signal matrix. We say an estimator is F-consistent, if $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F/\sqrt{np} = o_p(1)$. As discussed in Section 3.3 below, a few F-consistent estimators are available under general or specific GLFMs. However, the F-consistency only guarantees consistency in an average sense – the proportion of inconsistently estimated entries decays to zero. It cannot guarantee entrywise consistency, i.e., the consistency of $\hat{m}_{ij}$ for each individual data entry, which may be important in some applications concerning the prediction of individual data entries. Entrywise results for matrix completion, which focus on the loss $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_{\max}$, have been established under linear factor models (Abbe et al., 2020; Chen et al., 2019b, 2020c; Chernozhukov et al., 2023) but not under the GLFM. Establishing entrywise consistency is more challenging under the GLFM due to the involvement of non-linear link functions of the exponential family. In what follows, we propose methods that can improve an F-consistent estimator to an entrywise consistent (E-consistent) estimator under the GLFM.

3.1 Refinement without Data Splitting

Let $\hat{\mathbf{M}}$ be given by an F-consistent estimator based on observed data $(\mathbf{Y} \circ \mathbf{\Omega}, \mathbf{\Omega})$; see Section 3.3 for examples of such estimators. We propose the following refinement procedure that inputs $\hat{\mathbf{M}}$ and outputs an E-consistent estimator.

Algorithm 1: Refinement Procedure without Data Splitting

Input: Observed data $(\mathbf{Y} \circ \mathbf{\Omega}, \mathbf{\Omega})$, an initial estimate $\hat{\mathbf{M}}$ and a pre-specified constant C_2.
Step 1. Perform singular value decomposition (SVD) to $\hat{\mathbf{M}}$ and obtain $\hat{\mathbf{V}}_r \in \mathbb{R}^{p\times r}$ which contains the top-r right singular vectors of $\hat{\mathbf{M}}$.
Step 2. Calculate $\hat{\mathbf{A}} = \mathbf{proj}_{\{\mathbf{A}\in\mathbb{R}^{p\times r}:\|\mathbf{A}\|_{2\to\infty}\le C_2\}}(\hat{\mathbf{V}}_r)$, where $\mathbf{proj}_{\{\mathbf{A}\in\mathbb{R}^{p\times r}:\|\mathbf{A}\|_{2\to\infty}\le C_2\}}(\cdot)$ denotes a projection operator that projects a $p \times r$ matrix to satisfy the two-to-infinity norm constraint.
Step 3. For each $i \in [n]$, calculate $\tilde{\boldsymbol{\theta}}_i$ by solving an equation:

$$\sum_{j=1}^{p} \omega_{ij}\{y_{ij} - b_j'((\hat{\mathbf{a}}_j)^T\tilde{\boldsymbol{\theta}}_i)\}\hat{\mathbf{a}}_j = \mathbf{0}_r. \tag{1}$$

Step 4. For each $j \in [p]$, obtain $\tilde{\mathbf{a}}_j$ by solving the following equation:

$$\sum_{i=1}^{n} \omega_{ij}\{y_{ij} - b_j'((\tilde{\mathbf{a}}_j)^T\tilde{\boldsymbol{\theta}}_i)\}\tilde{\boldsymbol{\theta}}_i = \mathbf{0}_r. \tag{2}$$

Output: $\tilde{\mathbf{M}} = \tilde{\mathbf{\Theta}}(\tilde{\mathbf{A}})^T$, where $\tilde{\mathbf{\Theta}} = (\tilde{\boldsymbol{\theta}}_1, \cdots, \tilde{\boldsymbol{\theta}}_n)^T \in \mathbb{R}^{n\times r}$ and $\tilde{\mathbf{A}} = (\tilde{\mathbf{a}}_1, \cdots, \tilde{\mathbf{a}}_p)^T \in \mathbb{R}^{p\times r}$ are obtained from Steps 3 and 4, respectively.

We comment on the implementation. First, the constant C_2 depends on the true signal matrix $\mathbf{M}^*$. Recall that we assume $\mathbf{M}^*$ to be of rank r under the GLFM. Thus, $\mathbf{M}^*$ can be decomposed as $\mathbf{M}^* = \mathbf{U}_r^*\mathbf{D}_r^*(\mathbf{V}_r^*)^T$, where $\mathbf{U}_r^* \in \mathbb{R}^{n\times r}$ and $\mathbf{V}_r^* \in \mathbb{R}^{p\times r}$ are the left and right singular matrices corresponding to the non-zero singular values, and $\mathbf{D}_r^* \in \mathbb{R}^{r\times r}$ is a diagonal matrix whose diagonal elements are the singular values $\sigma_1(\mathbf{M}^*) \geq \cdots \geq \sigma_r(\mathbf{M}^*) > 0$. We require C_2 to satisfy $C_2 \geq \|\mathbf{V}_r^*\|_{2\to\infty}$. On the other hand, C_2 should not be chosen too large. As will be shown in Section 4.2, it is assumed that C_2 has the same asymptotic order as $\|\mathbf{V}_r^*\|_{2\to\infty}$; otherwise, the error bound for $\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max}$ needs additional modification. Second, we note that the projection in Step 2 is very easy to perform. Let $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_p)^T$ be a $p \times r$ matrix. Then $\mathbf{proj}_{\{\mathbf{A}\in\mathbb{R}^{p\times r}:\|\mathbf{A}\|_{2\to\infty}\leq C_2\}}(\mathbf{V}) = (\tilde{\mathbf{v}}_1, ..., \tilde{\mathbf{v}}_p)^T$, where $\tilde{\mathbf{v}}_i = \mathbf{v}_i$ if $\|\mathbf{v}_i\| \leq C_2$ and $\tilde{\mathbf{v}}_i = (C_2/\|\mathbf{v}_i\|)\mathbf{v}_i$ otherwise. Third, the algorithm requires knowing the number of factors r. Under the GLFM and suitable conditions, this quantity can be consistently selected based on information criteria (Chen and Li, 2022) or by identifying a singular value gap using a SVD-based approach (Zhang et al., 2020). Finally, we provide a remark on solving the equations in Steps 3 and 4.

Remark 1. *In Steps 3 and 4, we propose to solve some estimating equations. As will be shown in Section 4, these equations have a unique solution with probability converging to* 1 *under a suitable asymptotic regime. These steps are equivalent to performing optimization to certain log-likelihood functions. Let* $\ell(\mathbf{M}) = \sum_{i,j:\omega_{ij}=1}\{y_{ij}m_{ij} - b_j(m_{ij})\}$ *be a weighted log-likelihood function based on observed data* $(\mathbf{Y}\circ\mathbf{\Omega}, \mathbf{\Omega})$*, where the individual log-likelihood terms are weighted by the dispersion parameters*[1]*. Then, solving the estimating equations* (1) *is equivalent to solving* $\tilde{\mathbf{\Theta}} \in \arg\max_{\mathbf{\Theta}} \ell(\mathbf{\Theta}\hat{\mathbf{A}}^T)$*, and solving the estimating equations* (2) *is equivalent to solving* $\tilde{\mathbf{A}} \in \arg\max_{\mathbf{A}} \ell(\tilde{\mathbf{\Theta}}\mathbf{A}^T)$*. This is due to that the estimating equations* (1) *and* (2) *are obtained by taking the partial derivatives of* $\ell(\mathbf{\Theta}\mathbf{A}^T)$*with respect to* $\mathbf{\Theta}$ *and* $\mathbf{A}$*, respectively, and that the objective function* $\ell(\mathbf{\Theta}\mathbf{A}^T)$ *is convex with respect to* $\mathbf{\Theta}$ *and* $\mathbf{A}$ *given the other.*

We provide an informal theorem under a simplified setting to shed some light on the asymptotic behavior of Algorithm 1. Its formal version is Theorem 5 in Section 4.2, which is established under a more general setting. For the missing pattern $\mathbf{\Omega} = (\omega_{ij})_{i\in[n],j\in[p]}$, let $\pi_{ij} = \mathbb{P}(\omega_{ij} = 1)$ be the sampling probabilities and $\pi_{\min} = \min_{i\in[n],j\in[p]}\pi_{ij}$ and $\pi_{\max} = \max_{i\in[n],j\in[p]}\pi_{ij}$ be the minimal and maximal sampling probabilities, respectively. The notation π for the sampling probabilities should be distinguished from the Roman (upright font) notation $\mathrm{\pi}$ for the mathematical constant of circumference ratio in Example 1.

Theorem 2 (An informal and simplified version of Theorem 5)**.** *Assume that* $\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F \leq e_{\mathbf{M},F}) = 1$ *and let* $\tilde{\mathbf{M}}$ *be obtained by Algorithm 1. Then, under suitable assumptions on* $\mathbf{M}^*$ *and the asymptotic regime* $\pi_{\min} = \pi_{\max} = \pi$*,* r *is fixed,* $p\pi, n\pi \gg (\log(np))^3$*, and* $\{(n \wedge p)\pi\}^{-1/2} \lesssim (np)^{-1/2}e_{\mathbf{M},F} \ll \pi^{1/2}(\log(np))^{-2}$*, with probability tending to 1, we have* $\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \lesssim (\log(np))^2\pi^{-1/2}(np)^{-1/2}e_{\mathbf{M},F}$.

We clarify that $e_{\mathbf{M},F}$ in the above theorem is a non-random number that depends on n and p. We consider the asymptotic regime $\{(n\wedge p)\pi\}^{-1/2} \lesssim (np)^{-1/2}e_{\mathbf{M},F}$ above because $\{(n\wedge p)\pi\}^{-1/2}$ is the minimax error rate of $(np)^{-1/2}\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F$; see Chen and Li (2022).

Remark 3. *We provide intuitions on the result of Theorem 2 under the linear factor model setting. Using Wedin's sine angle theorem (Wedin, 1972) and under suitable assumptions, one can show that there exist* $\mathbf{\Theta}^* = (\theta_{ij}^*)_{n\times r}$ *and* $\mathbf{A}^* = (a_{ij}^*)_{p\times r}$*, such that* $\mathbf{M}^* = \mathbf{\Theta}^*(\mathbf{A}^*)^T$*,* $\|\hat{\mathbf{A}} - \mathbf{A}^*\|_F \lesssim$

1. The weighted likelihood is used so that the nuisance parameters ϕ_j do not involve in estimating $\mathbf{M}$, which simplifies the theoretical analysis. We believe that the current analysis can be extended to the unweighted log-likelihood function for the joint estimation of $\mathbf{M}$ and dispersion parameters ϕ_j.

$(np)^{-1/2}e_{\mathbf{M},F}$ *with probability tending to 1 as* n *and* p *grow to infinity,* $\|\mathbf{A}^*\|_{2\to\infty}\lesssim p^{-1/2}$*, and* $\|\boldsymbol{\Theta}^*\|_{2\to\infty}\lesssim p^{1/2}$.

Then solving for $\tilde{\boldsymbol{\theta}}_i$ *in Step 3 can be viewed as a linear regression problem with a small measurement error in the covariates, where* $\mathbf{a}_j^*$*s are the true covariates and* $\hat{\mathbf{a}}_j$ *are the covariates with measurement error. Under the linear factor model,* $b_j(m_{ij}) = m_{ij}^2/2$ *for all* j*. Thus, one can write down the analytic form for* $\tilde{\boldsymbol{\theta}}_i$ *that solves Equation* (1)*. From these analytic forms, one can show that with probability tending to 1,* $\|\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{2\to\infty}\lesssim \log(np)\pi^{-1/2}p^{1/2}\|\mathbf{A}^* - \hat{\mathbf{A}}\|_F\lesssim \log(np)\pi^{-1/2}n^{-1/2}e_{\mathbf{M},F}$*, which also implies that* $\|\tilde{\boldsymbol{\Theta}}\|_{2\to\infty}\lesssim p^{1/2}$*. Here, the* $\log(np)$ *term comes from a tail bound of* $\max_{i=1,\ldots,n,j=1,\ldots,p}|Y_{ij} - b'(m_{ij}^*)|$*. Similarly, one can obtain the analytical expression for* $\tilde{\mathbf{a}}_j$ *that solves Equation* (2)*, which now involves* $\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*$*,* $i = 1, \ldots, n$*. From these expressions, one can show that* $\|\tilde{\mathbf{A}} - \mathbf{A}^*\|_{2\to\infty}\lesssim \log(np)p^{-1}\|\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{2\to\infty}\lesssim (\log(np))^2\pi^{-1/2}n^{-1/2}p^{-1}e_{\mathbf{M},F}$ *holds with probability tending to 1. Combining the above results, it holds that, with probability tending to 1,*

$$\begin{aligned}\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} &\leq \|\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{2\to\infty}\|\mathbf{A}^*\|_{2\to\infty}+\|\tilde{\mathbf{A}} - \mathbf{A}^*\|_{2\to\infty}\|\tilde{\boldsymbol{\Theta}}\|_{2\to\infty}\\ &\lesssim (\log(np))^2\pi^{-1/2}(np)^{-1/2}e_{\mathbf{M},F}.\end{aligned}$$

3.2 Refinement with Data Splitting

From Theorem 2 above, we see that $\|\tilde{\mathbf{M}}-\mathbf{M}^*\|_{\max}$ achieves the same error rate as $\|\hat{\mathbf{M}}-\mathbf{M}^*\|_F/\sqrt{np}$ (up to a logarithm factor) when $\pi \sim 1$. However, when $\pi = o(1)$, the rate of $\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max}$ becomes worse than that of $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F/\sqrt{np}$, due to the factor $\pi^{-1/2}$ in the upper bound. This term comes from the worst-case scenario where $\hat{\mathbf{A}} - \mathbf{A}^*$ is highly dependent with $(\omega_{ij})_{j\in[p]}$ for some i (e.g., $\hat{\mathbf{a}}_j - \mathbf{a}_j^* \approx \omega_{ij}\mathbf{b}$ for all $j \in [p]$, some $i \in [n]$, and some random vector $\mathbf{b} \in \mathbb{R}^r$). To obtain a better error rate under the max norm, we propose a new procedure that uses a data splitting step to break the dependence between $\hat{\mathbf{A}}$ and $\boldsymbol{\Omega}$ in the following Algorithm 2. The proposed data splitting method is similar to the one proposed in Chernozhukov et al. (2023) for linear factor models, where a similar dependence issue exists. However, due to the non-linear link functions involved in the GLFM, the development of our method and its theory faces unique challenges.

Let $\mathcal{N}_1 \subset [n]$ be a random subset independent of $(\mathbf{Y}, \boldsymbol{\Omega})$. In particular, we let $I(i \in \mathcal{N}_1)$ be i.i.d. Bernoulli random variables with $\mathbb{P}(i \in \mathcal{N}_1) = 1/2$ for $i \in [n]$, where $I(\cdot)$ denotes the indicator function. By the law of large numbers, $\mathcal{N}_1$ is a subset of $[n]$ with size around $n/2$. We further let $\mathcal{N}_2 = [n] \setminus \mathcal{N}_1$.

The comments on Algorithm 1 regarding the choice of C_2, the number of factors r, the projection operator, and the solutions to the estimating equations apply similarly to Algorithm 2. As the rows and columns of the data matrix play a similar role, Algorithm 2 can be modified to split the columns instead of the rows. As summarized in Theorem 4, which is an informal and simplified version of Theorem 10 in Section 4.3, Algorithm 2 improves the error rate of Algorithm 1. In fact, $\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max}$ now achieves the same error rate as $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F/\sqrt{np}$ up to a logarithm factor, regardless of the missing rate π.

Theorem 4 (An informal and simplified version of Theorem 10)**.** *Assume that* $\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{M}}_{\mathcal{N}_k\cdot}-\mathbf{M}^*_{\mathcal{N}_k\cdot}\|_F\leq e_{\mathbf{M},F}) = 1$ *for* $e_{\mathbf{M},F}$ $(k = 1, 2)$ *and* $\tilde{\mathbf{M}}$ *is obtained by Algorithm 2. Then, under suitable assumptions on* $\mathbf{M}^*$ *and the asymptotic regime* $\pi_{\min} = \pi_{\max} = \pi$*,* r *is fixed,* $p\pi, n\pi \gg (\log(np))^3$*,*

Algorithm 2: Refinement Procedure with Data Splitting

Input: Observed data $(\mathbf{Y} \circ \mathbf{\Omega}, \mathbf{\Omega})$, a constraint parameter C_2, and initial estimates $\hat{\mathbf{M}}_{\mathcal{N}_k,\cdot}$ for $\mathbf{M}_{\mathcal{N}_k,\cdot} = (m_{ij})_{i\in\mathcal{N}_k, j\in[p]}$ obtained based on $(\mathbf{Y} \circ \mathbf{\Omega}, \mathbf{\Omega})_{\mathcal{N}_k,\cdot} = (y_{ij}\omega_{ij}, \omega_{ij})_{i\in\mathcal{N}_k, j\in[p]}$ for $k = 1, 2$.

Step 1. Perform SVD to $\hat{\mathbf{M}}_{\mathcal{N}_1,\cdot}$ and calculate $\hat{\mathbf{V}}_r^{(1)} \in \mathbb{R}^{p\times r}$ which contains the top-r right singular vectors of $\hat{\mathbf{M}}_{\mathcal{N}_1,\cdot}$.

Step 2. Calculate $\hat{\mathbf{A}}^{(1)} = (\hat{\mathbf{a}}_j^{(1)})_{j\in[p]}^T = \mathbf{proj}_{\{\mathbf{A}\in\mathbb{R}^{p\times r}:\|\mathbf{A}\|_{2\to\infty}\le C_2\}}(\hat{\mathbf{V}}_r^{(1)})$.

Step 3. Calculate $\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} = (\tilde{\boldsymbol{\theta}}_i)_{i\in\mathcal{N}_2}^T$, where for each $i \in \mathcal{N}_2$, $\tilde{\boldsymbol{\theta}}_i$ is obtained by solving the equation $\sum_{j=1}^p \omega_{ij}\{y_{ij} - b_j'((\hat{\mathbf{a}}_j^{(1)})^T\tilde{\boldsymbol{\theta}}_i)\}\hat{\mathbf{a}}_j^{(1)} = \mathbf{0}_r$.

Step 4. Calculate $\tilde{\mathbf{A}}^{(1)} = (\tilde{\mathbf{a}}_j^{(1)})_{j\in[p]}^T$, where for each $j \in [p]$, $\tilde{\mathbf{a}}_j^{(1)}$ is obtained by solving the equation $\sum_{i\in\mathcal{N}_2} \omega_{ij}\{y_{ij} - b_j'((\tilde{\mathbf{a}}_j^{(1)})^T\tilde{\boldsymbol{\theta}}_i)\}\tilde{\boldsymbol{\theta}}_i = 0_r$.

Step 5. Swap $\mathcal{N}_1$ and $\mathcal{N}_2$ in Steps 1 – 4, and obtain $\tilde{\mathbf{\Theta}}_{\mathcal{N}_1}$ and $\tilde{\mathbf{A}}^{(2)}$ accordingly.

Output: $\tilde{\mathbf{M}} = (\tilde{m}_{ij})_{i\in[n], j\in[p]}$, where $(\tilde{m}_{ij})_{i\in\mathcal{N}_1, j\in[p]} = \tilde{\mathbf{\Theta}}_{\mathcal{N}_1}(\tilde{\mathbf{A}}^{(2)})^T$ and $(\tilde{m}_{ij})_{i\in\mathcal{N}_2, j\in[p]} = \tilde{\mathbf{\Theta}}_{\mathcal{N}_2}(\tilde{\mathbf{A}}^{(1)})^T$.

---

and $\{(n \wedge p)\pi\}^{-1/2} \lesssim (np)^{-1/2}e_{\mathbf{M},F} \ll (\log(np))^{-2}$*, with probability tending to 1, we have* $\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \lesssim (\log(np))^2(np)^{-1/2}e_{\mathbf{M},F}$.

As the data splitting in Algorithm 2 is random, it may be beneficial to run it multiple times and then aggregate the resulting estimates. We describe this variation of Algorithm 2 below. For a fixed number of random splittings, the asymptotic behavior of Algorithm 3 is the same as that of Algorithm 2.

Algorithm 3: Refinement Procedure with Multiple Data Splittings

---

Input: Observed data $(\mathbf{Y} \circ \mathbf{\Omega}, \mathbf{\Omega})$ a constraint C_2 and the number of data splittings tot.

Step 1. Independently generate index sets $\mathcal{N}_1^{(k)} and \mathcal{N}_2^{(k)}$ and obtain initial estimates $\hat{\mathbf{M}}_{\mathcal{N}_1}^{(k)}$ and $\hat{\mathbf{M}}_{\mathcal{N}_2}^{(k)}$ based on $(\mathbf{Y} \circ \mathbf{\Omega}, \mathbf{\Omega})_{\mathcal{N}_1^{(k)},\cdot} = (y_{ij}\omega_{ij}, \omega_{ij})_{i\in\mathcal{N}_1^{(k)}, j\in[p]}$ and $(\mathbf{Y} \circ \mathbf{\Omega}, \mathbf{\Omega})_{\mathcal{N}_2^{(k)},\cdot} = (y_{ij}\omega_{ij}, \omega_{ij})_{i\in\mathcal{N}_2^{(k)}, j\in[p]}$, respectively, for $k = 1, 2, ..., \text{tot}$.

Step 2. For $k = 1, ..., \text{tot}$, run Algorithm 2 with data $(\mathbf{Y} \circ \mathbf{\Omega}, \mathbf{\Omega})$, initial estimates $\hat{\mathbf{M}}_{\mathcal{N}_1}^{(k)}$ and $\hat{\mathbf{M}}_{\mathcal{N}_2}^{(k)}$, index sets $\mathcal{N}_1^{(k)}, \mathcal{N}_2^{(k)}$ and a constraint parameter C_2. Obtain outputs $\tilde{\mathbf{M}}^{(k)}$, $k = 1, ..., \text{tot}$.

Output: $\tilde{\mathbf{M}} = (\sum_{k=1}^{\text{tot}} \tilde{\mathbf{M}}^{(k)})/\text{tot}$.

3.3 F-consistent Estimators

Our refinement methods require input from an F-consistent estimator. We give examples of F-consistent estimators.

CJMLE. The constrained joint maximum likelihood estimator (CJMLE) solves the following optimization problem

$$(\hat{\mathbf{\Theta}}, \hat{\mathbf{A}}) \in \arg\max_{\mathbf{\Theta},\mathbf{A}} \ \ell(\mathbf{\Theta}\mathbf{A}^T), \text{s.t. } \mathbf{\Theta} \in \mathbb{R}^{n\times r}, \mathbf{A} \in \mathbb{R}^{p\times r}, \|\mathbf{\Theta}\|_{2\to\infty} \leq C, \|\mathbf{A}\|_{2\to\infty} \leq C. \tag{3}$$

The estimate of $\mathbf{M}$ is then given by $\hat{\mathbf{M}} = \hat{\mathbf{\Theta}}\hat{\mathbf{A}}^T$. The terminology "joint likelihood" comes from the latent variable model literature (Chapter 6, Skrondal and Rabe-Hesketh, 2004). This literature distinguishes the joint likelihood from the marginal likelihood, depending on whether entries of $\mathbf{\Theta}$ are treated as fixed parameters or random variables, where the marginal likelihood is more commonly adopted in the statistical inference of traditional latent variable models. This estimator was first proposed in Chen et al. (2019a) and Chen et al. (2020b) for the estimation of high-dimensional GLFMs, and an error bound on $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F$ under a general matrix completion setting can be found in Theorem 2 of Chen and Li (2022). The computation of (3) can be done by an alternating maximization algorithm as given in Chen et al. (2020b). This algorithm is theoretically guaranteed to converge to a critical point and has good convergence performance according to numerical experiments (Chen et al., 2020b), though (3) is a nonconvex optimization problem.

More specifically, suppose that the true signal matrix has a decomposition $\mathbf{M}^* = \mathbf{\Theta}^*(\mathbf{A}^*)^T$, such that $\|\mathbf{\Theta}^*\|_{2\to\infty} \leq C$ and $\|\mathbf{A}^*\|_{2\to\infty} \leq C$. Then, under a similar setting as in Theorems 2 and 4, we have $\lim_{n,p\to\infty} \mathbb{P}(\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F/\sqrt{np} \leq \kappa^\dagger\{(p \wedge n)\pi\}^{-1/2}) = 1$, for some finite positive constant $\kappa^\dagger$. As shown in Proposition 1 of Chen and Li (2022), $\{(p \wedge n)\pi\}^{-1/2}$ is also the minimax lower bound for estimating $\mathbf{M}$ in the scaled Frobenius norm, which is why this lower bound is assumed for $(np)^{-1/2}e_{\mathbf{M},F}$ in Theorems 2 and 4.

NBE. The CJMLE requires solving a non-convex optimization problem for which convergence to the global optimum is not always guaranteed. The nuclear-norm-constrained-based estimator (NBE) is a convex approximation to CJMLE. It solves the following optimization problem

$$\hat{\mathbf{M}} \in \arg\max_{\mathbf{M}} \ \ell(\mathbf{M}), \ \text{ s.t. } \|\mathbf{M}\|_{\max} \leq \rho', \|\mathbf{M}\|_* \leq \rho'\sqrt{rnp}. \tag{4}$$

The nuclear norm constraint is introduced, since $\{\mathbf{M} \in \mathbb{R}^{n\times p} : \|\mathbf{M}\|_{\max} \leq \rho', \|\mathbf{M}\|_* \leq \rho'\sqrt{rnp}\}$ is a convex relaxation of $\{\mathbf{M} \in \mathbb{R}^{n\times p} : \|\mathbf{M}\|_{\max} \leq \rho', \text{rank}(\mathbf{M}) \leq r\}$. This estimator has been considered in Davenport et al. (2014) for the completion of binary matrices. When the true model follows the M2PL model and the true signal matrix $\mathbf{M}^*$ satisfies $\|\mathbf{M}^*\|_{\max} \leq \rho'$, then Theorem 1 of Davenport et al. (2014) implies that under the same setting of Theorems 2 and 4, $\lim_{n,p\to\infty} \mathbb{P}(\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F/\sqrt{np} \leq \kappa^\ddagger\{(p \wedge n)\pi\}^{-1/4}) = 1$, where $\kappa^\ddagger$ is a finite positive constant which depends on the true model parameters. We believe that the same rate holds for other GLFMs under the simplified setting of Theorems 2 and 4.

Other estimators. Note that other F-consistent estimators may be available for GLFMs, such as SVD-based methods (Chatterjee, 2015; Zhang et al., 2020), nuclear-norm-regularized estimators (Klopp, 2014; Koltchinskii et al., 2011; Negahban and Wainwright, 2012; Robin et al., 2020; Alaya and Klopp, 2019) and methods based on a matrix factorization norm (Cai and Zhou, 2013, 2016).

4. Theoretical Results

4.1 Assumptions and Useful Quantities

We make the following Assumptions 2 and 3 throughout Section 4.

Assumption 2. *$b_1(x) = \cdots = b_p(x) = b(x)$ for all $x \in \mathbb{R}$. In addition, $b(x) < \infty$ and $b''(x) > 0$ for all $x \in \mathbb{R}$.*

We note that this assumption is made for ease of presentation. It can be relaxed to allowing functions b_j to be variable-specific, and similar theoretical results hold following a similar proof. For each $\alpha > 0$, define functions $\kappa_2(\alpha) = \sup_{|x|\leq\alpha} b''(x), \kappa_3(\alpha) = \sup_{|x|\leq\alpha}|b^{(3)}(x)|$, and $\delta_2(\alpha) = \inf_{|x|\leq\alpha} b''(x)$. Let $\mathbf{M}^*$ have the SVD $\mathbf{M}^* = \mathbf{U}_r^*\mathbf{D}_r^*(\mathbf{V}_r^*)^T$ where r is the rank of $\mathbf{M}^*$, $\mathbf{U}_r^* \in \mathbb{R}^{n\times r}$ and $\mathbf{V}_r^* \in \mathbb{R}^{p\times r}$ are the left and right singular matrices corresponding to the top-r singular values, respectively, and $\mathbf{D}_r^* \in \mathbb{R}^{r\times r}$ is a diagonal matrix whose diagonal elements are the singular values $\sigma_1(\mathbf{M}^*) \geq \cdots \geq \sigma_r(\mathbf{M}^*) > 0$. In order to apply the proposed methods, we need to input C_2.

Assumption 3. *We choose C_2 such that $C_2 \geq \|\mathbf{V}_r^*\|_{2\to\infty}$.*

Define the following quantities that depend on $\mathbf{M}^*$. Let $\rho = \max_{i\in[n],j\in[p]}|m_{ij}^*|, C_1 = \{\|\mathbf{U}_r^*\|_{2\to\infty}\vee(r/n)^{1/2}\} \cdot \sigma_1(\mathbf{M}^*)$, $\kappa_2^* = \kappa_2(2\rho+1), \delta_2^* = \delta_2(2\rho+1),$ and $\kappa_3^* = \kappa_3(6C_1C_2)$.

4.2 Error Analysis without Data Splitting

Theorem 5. *Assume that $\lim_{n,p\to\infty} \mathbb{P}(\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F \leq e_{\mathbf{M},F}) = 1$, $\tilde{\mathbf{M}}$ is obtained by Algorithm 1, and the following asymptotic regime holds:*

R1: $\phi_1 = \cdots = \phi_p = \phi \sim 1$;

R2: $\pi_{\min} \sim \pi_{\max} \sim \pi$;

R3: $\|\mathbf{U}_r^*\|_{2\to\infty} \lesssim (r/n)^{1/2}$, $\|\mathbf{V}_r^*\|_{2\to\infty} \lesssim (r/p)^{1/2}$, $C_2 \sim (r/p)^{1/2}$;

R4: $\sigma_r(\mathbf{M}^*) \sim \sigma_1(\mathbf{M}^*) \sim (np)^{1/2}r^\eta$ for some constants $\eta \geq -1$;

R5: $p\pi \gg (\kappa_2^*)^4(\delta_2^*)^{-6}(\log(np))^3 \max\left[r^{(1+2\eta)\vee 5}, (\kappa_3^*)^2 r^{(3+4\eta)\vee 7}\right]$;

R6: $n\pi \gg (\kappa_2^*)^2(\delta_2^*)^{-4}(\log(np))^2 \max\{r^3, (\kappa_3^*)^2 r^5\}$;

R7: $(np)^{-1/2}e_{\mathbf{M},F} \ll (\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-2} \min[r^{-5/2}, (\kappa_3^*)^{-1}r^{-7/2}]\pi^{1/2}$.

Then, with probability converging to 1, estimating equations in steps 3 and 4 of Algorithm 1 have a unique solution and

$$\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \lesssim (\delta_2^*)^{-2}(\kappa_2^*)^2(\log(np))^2 r^{5/2}\left[\{(n\wedge p)\pi\}^{-1/2} + (np\pi)^{-1/2}e_{\mathbf{M},F}\right]. \qquad (5)$$

In particular, if we further assume that $r \sim 1$, then, the asymptotic regime requirements R5 – R7 *can be simplified as $p\pi \gg (\log(np))^3$, $n\pi \gg (\log(np))^2$ and $(np)^{-1/2}e_{\mathbf{M},F} \ll (\log(np))^{-2}\pi^{1/2}$, and we have that with probability converging to 1,*

$$\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \lesssim (\log(np))^2[\{(n\wedge p)\pi\}^{-1/2} + (np\pi)^{-1/2}e_{\mathbf{M},F}].$$

Remark 6. *We comment on the asymptotic requirements* R1–R7. R1 *requires the dispersion parameters to be the same for different $j \in [p]$. This assumption is made for ease of presentation, and it can be easily relaxed to allowing varying values of dispersion parameters. It further requires that the dispersion parameter is bounded as n and p grow large.* R2 *requires $\pi_{\max}$ and $\pi_{\min}$ to* *be of the same asymptotic order. That is, the missing pattern is not too far from the commonly adopted uniform missingness assumption where all the π_{ij} are the same (see, e.g. Candès and Tao, 2010; Davenport et al., 2014).* R3 *is a standard incoherent condition that is commonly assumed for matrix completion to avoid spiky low-rank matrices (Candès and Recht, 2009; Jain et al., 2013).* R4 *requires that the non-zero singular values of $\mathbf{M}^*$ are in the same asymptotic order. In addition, we restrict the analysis to the case where $\eta \geq -1$, because otherwise $\|\mathbf{M}^*\|_{\max} \ll 1$ and the asymptotic regime is less interesting. We note that* R4 *can be relaxed to a more general asymptotic regime allowing $\sigma_r(\mathbf{M}^*)$ and $\sigma_1(\mathbf{M}^*)$ to have different asymptotic order, and we provide the error analysis under a more general setting in the appendix.* R5 *and* R6 *require the expected number of non-missing observations for each row and column to be large enough.* R7 *requires the initial F-consistent estimator to have a sufficiently small estimation error in scaled Frobenius norm. In Corollary 8 below, we give sufficient conditions for* R5 – R7 *under the three specific GLFMs described in Section 2.*

Remark 7. *Let $\hat{\mathbf{M}}_{CJMLE}$ and $\hat{\mathbf{M}}_{NBE}$ denote the constrained joint maximum likelihood estimator and nuclear-norm-constrained-based estimator described in Section 3.3, respectively. Also let $\tilde{\mathbf{M}}_{CJMLE}$ and $\tilde{\mathbf{M}}_{NBE}$ be the corresponding refined estimators by applying Algorithm 1. Theorem 5 indicates that with high probability $\|\tilde{\mathbf{M}}_{CJMLE} - \mathbf{M}^*\|_{\max} \lesssim (\log(np))^2 \pi^{-1}(n \wedge p)^{-1/2}$ and $\|\tilde{\mathbf{M}}_{NBE} - \mathbf{M}^*\|_{\max} \lesssim (\log(np))^2 \pi^{-3/4}(n \wedge p)^{-1/4}$ when r is bounded, under suitable regularity conditions. Because $\hat{\mathbf{M}}_{CJMLE}$ is asymptotically minimax when $\pi \sim 1$ in Frobenius norm, we also have that $\tilde{\mathbf{M}}_{CJMLE}$ is asymptotically minimax in the matrix max norm.*

In the following corollary, we provide sufficient conditions for *R5 - R7* under specific GLFMs discussed earlier.

Corollary 8. *Assume that $\lim_{n,p\to\infty} \mathbb{P}(\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F \leq e_{\mathbf{M},F}) = 1$ for some non-random $e_{\mathbf{M},F}$. Then,* (5) *holds under one of the following specific models and asymptotic requirements.*

1. *Data follow a binomial factor model and the following asymptotic requirements hold:* R2 – R4 *and* R5B*: $p\pi \gg (n \vee p)^{\epsilon_0} r^{(3+4\eta)\vee 7}$;* R6B*: $n\pi \gg (n \vee p)^{\epsilon_0} r^5$;* R7B: $(np)^{-1/2} e_{\mathbf{M},F} \ll (n \wedge p)^{-\epsilon_0} \pi^{1/2} r^{-7/2}$*;* R8B: $k_1 = \cdots = k_p = k \sim 1$*; and* R9B: $\rho \lesssim \log(n \wedge p)^{1-\epsilon_0}$ *for some* $\epsilon_0 > 0$.

2. *Data follow a normal factor model and the following asymptotic requirements hold:* R1 – R4*;* R5N*: $p\pi \gg (\log(np))^3 r^{(1+2\eta)\vee 5}$;* R6N*: $n\pi \gg (\log(np))^2 r^3$; and* R7N*: $(np)^{-1/2} e_{\mathbf{M},F} \ll (\log(np))^{-2} \pi^{1/2} r^{-5/2}$.*

3. *Data follow a Poisson factor model and the following asymptotic requirements hold:* R2 - R4*,* R5B – R7B *and* R10P*: $r^{1+\eta} \lesssim (\log(n \wedge p))^{1-\epsilon_0}$ for some $\epsilon_0 > 0$.*

In the first part of the above corollary, *R1* automatically holds because the dispersion parameter $\phi_j = 1$ in the binomial model.

Remark 9. *We comment on the asymptotic requirements in the above corollary.* R5B, R6B, R5N *and* R6N *require that rank r is relatively small comparing with $(n \wedge p)\pi$, and it can grow at most of the order $\{(n \wedge p)\pi\}^{\nu_1}$ for some constant $\nu_1 \in (0, 1)$. Conditions* R5B *and* R6B *are slightly stronger than* R5N *and* R6N*, because $\kappa_3^* = 0$ for the normal model while $\kappa_3^* \sim 1$ for the binomial model. Conditions* R7B *and* R7N *require the scaled Frobenius norm of the initial estimator to be small.*

Many F-consistent estimators, including CJMLE and NBE, have the error rate $(np)^{-1/2}e_{\mathbf{M},F} \sim ((n \wedge p)\pi)^{-\nu_2}$ *for some* $\nu_2 \in (0,1)$. *For these estimators,* R7B *and* R7N *require that* $r \lesssim ((n \wedge p)\pi)^{\nu_3}\pi^{1/2}$ *for some* $\nu_3 \in (0,1)$. *Condition* R8B *requires the* k_j*s to be the same for different* $j \in [p]$ *and are bounded. This condition can be easily relaxed to a more general setting with varying but bounded* k_j*s. Condition* R9B *requires that* ρ *grows much slower than* n *and* p. *Similar assumptions are made for 1-bit matrix completion (Davenport et al., 2014; Cai and Zhou, 2013). For Poisson factor models,* R10P *can be achieved either by an arbitrary* r *with* $\eta = -1$ *or by* $r \lesssim (\log(n \wedge p))^{(1-\epsilon_0)/(1+\eta)}$ *with* $\eta > -1$.

4.3 Error Analysis with Data Splitting

Theorem 10. *Assume that* $\lim_{n,p\to\infty} \mathbb{P}(\|\hat{\mathbf{M}}_{\mathcal{N}_k\cdot} - \mathbf{M}^*_{\mathcal{N}_k\cdot}\|_F \leq e_{\mathbf{M},F}) = 1$ *for some non-random* $e_{\mathbf{M},F}$ $(k = 1, 2)$*, and* $\tilde{\mathbf{M}}$ *is obtained by Algorithm 2. Assume asymptotic requirements* R1 - R6 *in Theorem 5 hold as* $n, p \to \infty$. *Also, assume the following asymptotic requirements:*

R7' $(np)^{-1/2}e_{\mathbf{M},F} \ll (\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-2} \min[r^{-5/2}, (\kappa_3^*)^{-1}r^{-7/2}]$.

Then, with probability converging to 1*, estimating equations in steps 3 and 4 of Algorithm 2 have a unique solution and*

$$\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \lesssim (\delta_2^*)^{-2}(\kappa_2^*)^2 \log^2(np) r^{5/2}\Big[\{(p \wedge n)\pi\}^{-1/2} + (np)^{-1/2}e_{\mathbf{M},F}\Big]. \tag{6}$$

In particular, if we further assume that $r \sim 1$*, then, the asymptotic regime requirements* R5*,* R6*, and* R7' *can be simplified as* $p\pi \gg (\log(np))^3$*,* $n\pi \gg (\log(np))^2$ *and* $(np)^{-1/2}e_{\mathbf{M},F} \ll (\log(np))^{-2}$*, and we have that with probability converging to* 1*,* $\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \lesssim (\log(np))^2[\{(n \wedge p)\pi\}^{-1/2} + (np)^{-1/2}e_{\mathbf{M},F}]$.

Remark 11. *There are two main differences between Theorem 5 and Theorem 10. First, the asymptotic requirement* R7 *has an extra factor* $\pi^{1/2}$ *when compared with* R7'*. Second, the error rate* (5) *has an extra* $\pi^{-1/2}$ *factor when compared with* (6)*. Thus, when* $\pi \ll 1$*, Algorithm 1 requires stronger regularity conditions and has a larger error rate. Additional results under a more general asymptotic regime are provided in the appendix.*

The following corollary give sufficient conditions for *R7'* to hold under specific GLFMs.

Corollary 12. *Assume that* $\lim_{n,p\to\infty} \mathbb{P}(\|\hat{\mathbf{M}}^{(k)}_{\mathcal{N}_k\cdot} - \mathbf{M}^*_{\mathcal{N}_k\cdot}\|_F \leq e_{\mathbf{M},F}) = 1$ *for some non-random* $e_{\mathbf{M},F}$ $(k = 1, 2)$*. Then,* (6) *holds under one of the following specific models and asymptotic requirements.*

1. *Data follow a binomial factor model and the following asymptotic requirements hold:* R2 - R4*,* R5B*,* R6B*,* R8B*,* R9B*, and* R7'B*:* $(np)^{-1/2}e_{\mathbf{M},F} \ll (n \wedge p)^{-\epsilon_0}r^{-7/2}$ *for some* $\epsilon_0 > 0$.

2. *Data follow a normal factor model and the following asymptotic requirements hold:* R1 - R4*,* R5N*,* R6N*, and* R7'N*:* $(np)^{-1/2}e_{\mathbf{M},F} \ll (\log(np))^{-2}r^{-5/2}$.

3. *Data follow a Poisson factor model and that asymptotic requirements* R2 - R4*,* R5B*,* R6B*,*R7'B*, and* R10P *hold.*

Remark 9 still applies to Corollary 12, except that now we have a better rate when π is close to zero.

Procedure	Initial	Refinement	Procedure	Initial	Refinement
1	NBE		5	CJMLE	
2	NBE	Algorithm 1	6	CJMLE	Algorithm 1
3	NBE	Algorithm 2	7	CJMLE	Algorithm 2
4	NBE	Algorithm 3	8	CJMLE	Algorithm 3

Table 1: Estimation procedures compared in a simulation study.

Setting	n	p	r	π	Setting	n	p	r	π
1	400	200	3	0.6	4	400	200	3	0.2
2	800	400	3	0.6	5	800	400	3	0.2
3	1600	800	3	0.6	6	1600	800	3	0.2

Table 2: Simulation settings. All the variables are ordinal (with $k_j = 5$), for which the Binomial model is assumed.

5. Simulation Study

We evaluate the proposed methods via a simulation study. Eight estimation procedures are considered as listed in Table 1. For Algorithm 3, five data splittings are performed. These procedures are applied under 24 simulation settings, where n, p, r, $\pi_{\max} = \pi_{\min} = \pi$, and variable types are varied. Settings 1-6 are listed in Table 2, where all the variables follow Binomial distribution with $k_j = 5$. The rest of the settings and additional details on data generation can be found in the appendix. For each simulation setting, 100 simulations are conducted.

The procedures are evaluated under two loss functions, the scaled Frobenius norm $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F/\sqrt{np}$ and the max norm $\|\hat{\mathbf{M}} - \mathbf{M}^*\|_{\max}$. The results for Settings 1-6 are given in Figures 1 and 2, and those for the other settings show similar patterns and are given in the appendix. First, for each procedure and given r and π, both the scaled Frobenius norm and the max norm decay as n and p grow simultaneously. Second, comparing the two figures, we see that the error rates are larger under Settings 4-6 than those under Settings 1-3 given the same n, p, and r, as the proportion of missing entries is higher under Settings 4-6. Third, Procedure 1 (i.e., NBE with no refinement) has larger error rates than its refined versions (Procedures 2-4), suggesting that the refinement procedures reduce the error of the initial NBE. Fourth, we see that Procedures 5 and 6 perform similarly, which is expected as they are asymptotically equivalent, as discussed in Remark 7. Fifth, comparing Procedures 2 and 6, we see that the refined NBE and the refined CJMLE have very similar performance. Similar patterns are observed when comparing Procedures 3 and 7 and when comparing Procedures 4 and 8. At first glance, it may seem a little counter-intuitive. According to Theorems 5 and 10, the error in the max norm of a refined estimator is upper bounded by the error in the scaled Frobenius norm of its initial estimator, and thus, we would expect the CJMLE-based refinements to have smaller errors in the max norm than the NBE-based refinements. The pattern under the current settings may be explained by the SVD steps in Algorithms 1, 2, and 3 that project the initial estimate to the space of rank-r matrices. Under these settings, the initial NBE after projection tends to approximate the CJMLE. We note that this is not always the case under other settings. Under settings 23 and 24 (see their results in the appendix), the CJMLE tends to outperform the projected NBE, and thus, the CJMLE-based refinements tend to outperform the

NBE-based refinements. Finally, comparing within Procedures 2-4 and comparing within Procedures 6-8, we see that Algorithm 1 leads to better empirical performance regardless of the value of π, even though Algorithm 2 has a faster theoretical convergence speed when π approaches 0. We conjecture that for CJMLE and NBE, the resulting $\hat{\mathbf{A}}$ in Step 2 of Algorithm 1 does not have a high dependence with any rows of $\mathbf{\Omega}$ when ω_{ij}s are uniformly sampled, and thus, the upper bound in (5) may be improved in this case. We also observe that Algorithm 3 outperforms Algorithm 2 through aggregating results from multiple runs Algorithm 2. By running Algorithm 2 five times, Algorithm 3 has a similar performance as Algorithm 1.

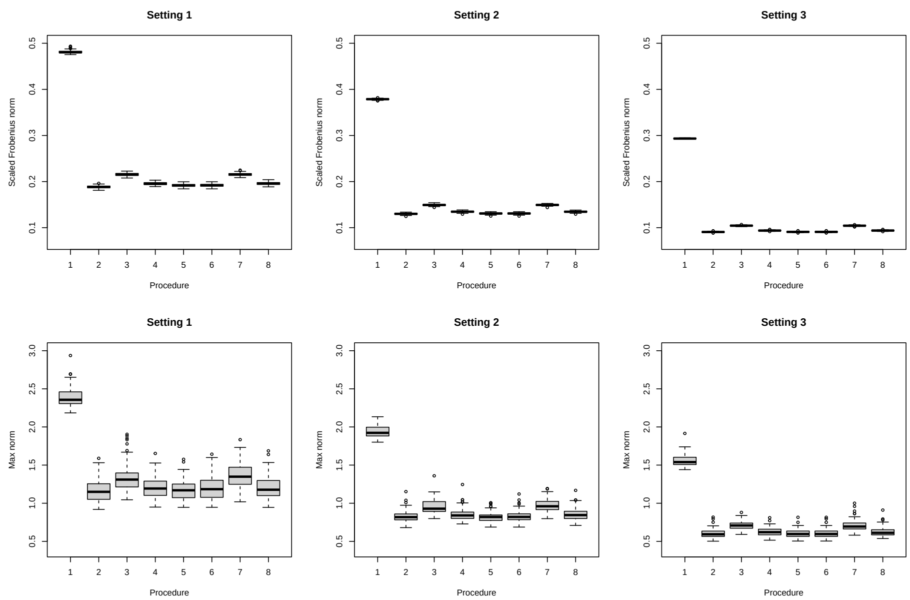

Figure 1: Results from Simulation Settings 1-3. The panels on the first row show the results based on the scaled Frobenius norm, and those on the second row show the results based on the max norm. In each panel, the box plots show the results of the eight procedures in Table 1, each constructed from 100 independent simulations.

6. Real Data Examples

6.1 Collaborative Filtering

We apply the proposed method to a MovieLens dataset for movie recommendation (Harper and Konstan, 2015). The dataset contains 943 users' ratings on 1,682 movies. Only 6.3% of the data entries are observed. For each movie, the raw ratings take integer values from 1 to 5. We transform the values from 0 to 4, and then apply the binomial factor model with $k_j = 4$ for all j. The goal is to predict the unobserved entries for movie recommendations.

The eight procedures in Table 1 are considered, with candidate rank $r = 1, 2, 3$, and 4. To evaluate the procedures, we split the data into training and test datasets, where the training and test sets contain 80% and 20% of the observed entries, respectively. We estimate the $\mathbf{M}$ matrix using the training set and then evaluate the prediction accuracy by the test-set log-likelihood at the estimated

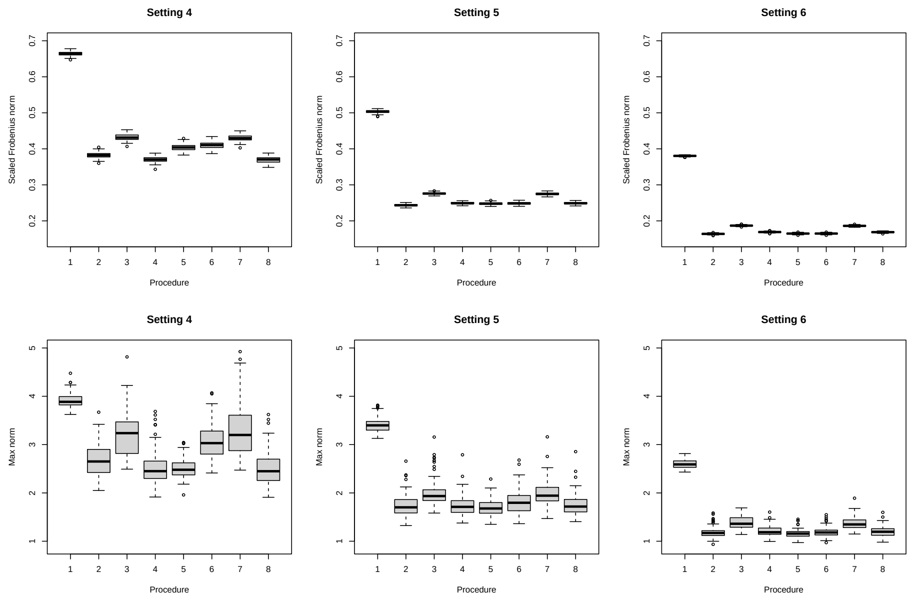

Figure 2: Results from Simulation Settings 4-6. The plots can be interpreted similarly as those in Figure 1.

	Procedure Index							
Rank	1	2	3	4	5	6	7	8
1	-48928	-49247	-49397	-49253	-49256	-49266	-49266	-49163
2	-53201	-49505	-49767	-48875	-48437	-48493	-48654	-48341
3	-56091	-49284	-49754	-48570	-49022	-49217	-48837	**-48207**
4	-56235	-49633	-50037	-48611	-51192	-51986	-49174	-48271

Table 3: Test-set log-likelihoods for the MovieLens data. The eight procedures are listed in Table 1.

M. A larger log-likelihood function value implies a higher prediction accuracy. The results are given in Table 3. The refinement methods improve the test-set log-likelihood of the NBE when $r = 2, 3, 4$ but not when $r = 1$, likely due to the rank-one model being too restrictive for the current data. Turning to the results from the CJMLE and its refinements, we see that Procedures 5 and 6 tend to perform similarly. We also see that Procedure 8, which is a refinement of CJMLE by Algorithm 3, tends to improve the test-set log-likelihood of CJMLE under all values of r. Procedure 7 also performs fine, despite its relatively high variance brought by performing data splitting only once in Algorithm 2. The good performance of Procedures 7 and 8 is likely due to that the distribution of the data missingness indicators ω_{ij} is far from a uniform distribution. Instead, their distribution likely depends on the true signal matrix (i.e., people may be more likely to have watched movies that they like), which may lead to dependence between the initial estimate $\hat{\mathbf{A}}$ and some rows of $\boldsymbol{\Omega}$ when data splitting is not performed. Such dependence leads to a larger estimation error. The largest test-set log-likelihood is given by Procedure 8 (i.e., CJMLE refined by Algorithm 3) when $r = 3$.

	Procedure Index							
Rank	1	2	3	4	5	6	7	8
1	-67205	-67938	-67958	-67921	-67587	-67516	-68204	-68140
2	-71620	-68556	-68733	-67749	**-63250**	-63313	-64914	-64842
3	-75816	-70092	-70067	-69151	-65476	-65370	-68611	-67693
4	-77632	-72365	-72238	-71640	-72320	-72648	-79466	-75989

Table 4: Test-set log-likelihoods for the PISA data. The eight procedures are listed in Table 1.

6.2 Large-scale Assessment in Education

We apply the proposed method to data from the 2018 Program for International Student Assessment (PISA; OECD, 2019a), a large-scale international educational survey operated by the Organization for Economic Co-operation and Development (OECD). We consider a subset of the PISA 2018 dataset, containing 9,970 students' responses to 415 assessment items. The students were from 37 OECD countries. The 415 assessment items measure four knowledge domains, including mathematics, science, reading, and global competence. A matrix sampling design is adopted in PISA 2018, under which each student was only assigned a subset of assessment items. Consequently, only 15.5% of the entries are observed in the dataset. Under this matrix sampling design, it is not sensible to directly compare students' performance based on their total scores, as the students answered different assessment items, and the items measure different knowledge domains and are not equally difficult. Among these items, 396 items are dichotomously scored, and 19 items have score levels $0, 1$ and 2. The goal is to predict students' performance on the items they did not receive in order to compare the performance based on the entire set of items.

We apply the binomial factor model. Similar to the above analysis, we split 80% and 20% of the data into training and test sets and evaluate the prediction accuracy by the test-set log-likelihood. The eight procedures in Table 1 are considered, with candidate rank $r = 1, 2, 3,$ and 4. The results are given in Table 4. First, the refinement methods tend to improve the test-set log-likelihood given by the NBE, except for the case when $r = 1$. The results given by the CJMLE and its refinement by Algorithm 1 are similar under all values of r. They tend to be better than the refinements given by Algorithms 2 and 3, likely due to that the variance brought by data splitting is high in this analysis. Second, the largest test-set log-likelihood is achieved by the CJMLE when the rank $r = 2$. The test-set log-likelihoods of the CJMLE and its refinement by Algorithm 1 are similar when $r = 2$, and they tend to substantially outperform the rest. In the analysis of PISA data, each of the knowledge domains is believed to correspond to at least one latent factor. Thus, four- or higher-dimensional factor models are typically adopted to jointly model the item responses (see Chapter 9, page 22, OECD, 2019b). Our results suggest that a lower-dimensional factor model may have better prediction performance, though not necessarily have better performance in terms of statistical inference and interpretation. This finding is closely related to the discussion in psychometrics regarding the value of subscores (Haberman, 2008).

7. Discussions

This note concerns matrix completion for mixed data under a GLFM framework. It proposes entry-wise consistent methods for estimating GLFMs based on a partially observed data matrix. Probabilistic error bounds are established for the matrix max norm under sensible asymptotic regimes (see Section 4), and they are extended under a more general asymptotic regime in the appendix. These error bounds imply the entrywise consistency and, further, characterize the asymptotic behaviors of the proposed methods. With these error bounds, optimal results are established under suitable asymptotic regimes. The proposed procedures are applied to two real data examples, one on movie recommendation and the other on large-scale educational assessment. For the movie recommendation example, the best predictive model is a rank-three model obtained by refining the CJMLE with Algorithm 3. For the educational assessment example, a rank-two model given by the CJMLE turns out to be the most predictive one.

The current work can be extended in several directions. First, some popular factor models, such as the probit model for binary data considered in Davenport et al. (2014), are not exponential family GLFMs. We believe that our refinement procedures and their theory can be extended to many other models beyond the exponential family GLFM. This is because the theoretical properties of these procedures mainly rely on the convexity of the loss function with respect to $\mathbf{M}$, which still holds under many other non-linear factor models. Second, the optimal rate for estimating GLFMs is worth future investigation. We currently do not know whether our upper bounds are minimax optimal when the dimension r diverges. Sharp lower bounds need to be developed to answer this question.

Acknowledgments

This research is partially supported by National Science Foundation CAREER DMS-2143844.

Appendix

This appendix provides additional theoretical results, proof of the theorems, and additional simulation results.

Appendix A. Proof of Theorem 10 and Additional Theoretical Results for Algorithm 2 with Data splitting

In this section, we obtain the error bound for

$$\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \leq \max(\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_1}(\tilde{\mathbf{A}}^{(1)})^T - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_{\max}, \|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}(\tilde{\mathbf{A}}^{(2)})^T - \mathbf{M}^*_{\mathcal{N}_2\cdot}\|_{\max}).$$

We will provide detailed analysis for $\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_1}(\tilde{\mathbf{A}}^{(1)})^T - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_{\max}$. The analysis of $\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}(\tilde{\mathbf{A}}^{(2)})^T - \mathbf{M}^*_{\mathcal{N}_2\cdot}\|_{\max}$ is similar and is thus omitted. For the ease of presentation, we drop the superscript (1) in $\hat{\mathbf{A}}^{(1)}$ when the context is clear. Recall that $\mathbf{M}^*$ has the SVD $\mathbf{M}^* = \mathbf{U}^*_r\mathbf{D}^*_r(\mathbf{V}^*_r)^T$ where $\mathbf{U}^*_r \in \mathbb{R}^{n\times r}$, $\mathbf{V}^*_r \in \mathbb{R}^{p\times r}$ denote the left and right singular matrices, and $\mathbf{D}^*_r = \mathrm{diag}(\sigma_1(\mathbf{M}^*), \cdots, \sigma_r(\mathbf{M}^*))$.

The rest of the section is organized as follows. In Section A.1, we obtain an error bound for $\|\hat{\mathbf{A}} - \mathbf{A}^*\|_F$ where $\mathbf{A}^* = \mathbf{V}^*_r\hat{\mathbf{P}}$ for a carefully chosen orthogonal matrix $\hat{\mathbf{P}}$. In Section A.2, we provide non-asymptotic and non-probabilistic bounds for solutions to the non-linear estimation equations used in Step 3 and 4 in the proposed Algorithm 2. In Section A.3, we obtain non-asymptotic probabilistic bounds for terms involved in Section A.2. In Section A.4, we put together results in Sections A.1 – A.3 and obtain asymptotic error bounds for $\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*\|_{2\to\infty}$ (Lemma 38), $\|\tilde{\mathbf{A}} - \mathbf{A}^*\|_{2\to\infty}$ (Lemma 39), and $\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_1}(\tilde{\mathbf{A}}^{(1)})^T - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_{\max}$ (Lemma 40) where $\mathbf{\Theta}^* = \mathbf{U}^*_r\mathbf{D}^*_r\hat{\mathbf{P}}$. Finally, we provide additional theoretical results for Algorithm 2 in Section A.5 and the proof of Theorem 10 in Section A.6.

Throughout the analysis, for real number operators, we calculate multiplication and division before the max and min operators ('$\vee$' and '$\wedge$') unless otherwise specified. For example, $u(xy \vee z/w) = u\max(xy, z/w)$ for real numbers x, y, u, w, z. For two events A and B, we say 'event A has probability at least $1-\epsilon$ on event B', if $\mathbb{P}(A^c \cap B) \leq \epsilon$. Note that $\mathbb{P}(A) \geq 1 - \epsilon - \mathbb{P}(B^c)$ in this case.

A.1 Error Analysis for $\hat{\mathbf{A}}$

In this section, we provide an error bound for $\hat{\mathbf{A}}$ given an error bound for $\hat{\mathbf{M}}_{\mathcal{N}_1\cdot}$.

Lemma 13. *Let $\psi_r = \sigma_r(\mathbf{M}^*_{\mathcal{N}_1\cdot}) \wedge \sigma_r(\mathbf{M}^*_{\mathcal{N}_2\cdot})$ and $\psi_1 = \sigma_r(\mathbf{M}^*_{\mathcal{N}_1\cdot}) \vee \sigma_r(\mathbf{M}^*_{\mathcal{N}_2\cdot})$. If $\|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot} - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_2 \leq 2^{-1}\psi_r$, $\|\mathbf{V}^*_r\|_{2\to\infty} \leq C_2$ and rank$(\mathbf{M}^*) = r$, then there exists an orthogonal matrix $\hat{\mathbf{P}} \in \mathbb{R}^{r\times r}$ satisfying*

$$\|\hat{\mathbf{A}} - \mathbf{V}^*_r\hat{\mathbf{P}}\|_F \leq 8\psi_r^{-1}\|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot} - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_F. \tag{7}$$

Proof [Proof of Lemma 13] According to Weyl's inequality and the assumption that $\|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot} - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_2 \leq 2^{-1}\psi_r$, $\sigma_r(\hat{\mathbf{M}}_{\mathcal{N}_1\cdot}) \geq \sigma_r(\mathbf{M}^*_{\mathcal{N}_1\cdot}) - \|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot} - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_2 \geq 2^{-1}\sigma_r(\mathbf{M}^*_{\mathcal{N}_1\cdot}) \geq 2^{-1}\psi_r$. Thus the gaps of singular value satisfies

$$\min\left[\min_{1\leq i\leq r, j>r}\{\sigma_i(\hat{\mathbf{M}}_{\mathcal{N}_1\cdot}) - \sigma_j(\mathbf{M}^*_{\mathcal{N}_1\cdot})\}, \min_{1\leq i\leq r}\sigma_i(\hat{\mathbf{M}}_{\mathcal{N}_1\cdot})\right] = \min\left\{\sigma_r(\hat{\mathbf{M}}_{\mathcal{N}_1\cdot}), \sigma_r(\mathbf{M}^*_{\mathcal{N}_1\cdot})\right\} \geq 2^{-1}\psi_r. \tag{8}$$

Let $\mathbf{V}^*_{r,\mathcal{N}_1\cdot} \in \mathbb{R}^{p\times r}$ be the right singular value matrix corresponding to the top-r singular values of $\mathbf{M}^*_{\mathcal{N}_1\cdot}$ and

$$\mathbf{P}^\dagger = \arg\min_{\mathbf{P}\in\mathcal{O}_r}\|\hat{\mathbf{V}}_r - \mathbf{V}^*_{r,\mathcal{N}_1\cdot}\mathbf{P}\|_F, \tag{9}$$

where $\mathcal{O}_r$ denotes the set of all $r \times r$ orthogonal matrices. According to the above equations and Wedin's sine angle theorem (Wedin, 1972),

$$\|\hat{\mathbf{V}}_r - \mathbf{V}^*_{r,\mathcal{N}_1\cdot}\mathbf{P}^\dagger\|_F = \inf_{\mathbf{P}\in\mathcal{O}_r}\|\hat{\mathbf{V}}_r - \mathbf{V}^*_{r,\mathcal{N}_1\cdot}\mathbf{P}\|_F \leq \frac{2\|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot} - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_F}{\sigma_r(\hat{\mathbf{M}}_{\mathcal{N}_1\cdot})} \leq \frac{4\|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot} - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_F}{\psi_r}. \tag{10}$$

On the other hand, since $\sigma_r(\mathbf{M}^*_{\mathcal{N}_1\cdot}) \geq \psi_r > 0$, the column space of $(\mathbf{M}^*_{\mathcal{N}_1\cdot})^T$ is the same as the columns space of $\mathbf{V}^*_{r,\mathcal{N}_1\cdot}$ and that of $\mathbf{V}^*_r$. This implies that there exists an orthogonal matrix $\bar{\mathbf{P}} \in \mathbb{R}^{r\times r}$ such that $\mathbf{V}^*_{r,\mathcal{N}_1\cdot} = \mathbf{V}^*_r\bar{\mathbf{P}}$, which further implies that for the orthogonal matrix

$$\hat{\mathbf{P}} = \bar{\mathbf{P}}\mathbf{P}^\dagger, \tag{11}$$

we have $\|\hat{\mathbf{V}}_r - \mathbf{V}^*_r\hat{\mathbf{P}}\|_F \leq 4\psi_r^{-1}\|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot} - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_F$. According to Algorithm 2, $\hat{\mathbf{A}}$ is the projection of $\hat{\mathbf{V}}_r$ to the set $\{\mathbf{A} \in \mathbb{R}^{p\times r} : \|\mathbf{A}\|_{2\to\infty} \leq C_2\}$ and $\|\mathbf{V}^*_r\hat{\mathbf{P}}\|_{2\to\infty} = \|\mathbf{V}^*_r\|_{2\to\infty} \leq C_2$. Thus,

$$\|\hat{\mathbf{A}} - \mathbf{V}^*_r\hat{\mathbf{P}}\|_F \leq \|\hat{\mathbf{A}} - \hat{\mathbf{V}}_r\|_F + \|\hat{\mathbf{V}}_r - \mathbf{V}^*_r\hat{\mathbf{P}}\|_F \leq 2\|\hat{\mathbf{V}}_r - \mathbf{V}^*_r\hat{\mathbf{P}}\|_F \leq 8\psi_r^{-1}\|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot} - \mathbf{M}^*_{\mathcal{N}_1\cdot}\|_F. \tag{12}$$

■

The next lemma is obtained by directly applying Lemma 13.

Lemma 14. *If* $\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot}-\mathbf{M}^*_{\mathcal{N}_1\cdot}\|_F\geq e_{\mathbf{M},F})=0$, $e_{\mathbf{M},F}$ *is a non-random number (depending on* n *and* p*),* $\|\mathbf{V}_r^*\|_{2\to\infty}\leq C_2$ *and* $e_{\mathbf{M},F}\leq 2^{-1}\psi_r$*, then*

$$\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{A}}-\mathbf{V}_r^*\hat{\mathbf{P}}\|_F\geq e_{\mathbf{A},F})=0, \tag{13}$$

where $\hat{\mathbf{P}}$ *is defined in* (11) *and* $e_{\mathbf{A},F}=8\psi_r^{-1}e_{\mathbf{M},F}$.

A.2 Non-probabilistic Bounds for Solutions to Estimating Equations

Recall that for each $i\in[n]$, the partial score function corresponding to $\boldsymbol{\theta}_i$ is

$$S_{1,i}(\boldsymbol{\theta}_i;\mathbf{A}):=\frac{\partial}{\partial\boldsymbol{\theta}_i}\ell(\boldsymbol{\Theta},\mathbf{A})=\phi^{-1}\sum_{j=1}^{p}\omega_{ij}\{y_{ij}-b'(\mathbf{a}_j^T\boldsymbol{\theta}_i)\}\mathbf{a}_j \tag{14}$$

The next lemma provides a non-probabilistic bound for the solution to the partial score equation $S_{1,i}(\boldsymbol{\theta}_i,\mathbf{A})=\mathbf{0}_r$.

Lemma 15. *Let* $\boldsymbol{\Theta}^*\in\mathbb{R}^{n\times r}$ *and* $\mathbf{A}^*\in\mathbb{R}^{p\times r}$ *be such that* $\mathbf{M}^*=\boldsymbol{\Theta}^*(\mathbf{A}^*)^T$ *and* $\mathbf{Z}=(z_{ij})$ *with* $z_{ij}=y_{ij}-b'(m^*_{ij})$ *and* $diag(\boldsymbol{\Omega}_{i\cdot}):=diag(\omega_{i1},\cdots,\omega_{ip})$. *If* $\|\boldsymbol{\Theta}^*\|_{2\to\infty}\leq C_1$, $\|\mathbf{A}^*\|_{2\to\infty},\|\mathbf{A}\|_{2\to\infty}\leq C_2$ *and there exists* $\xi>0$ *such that*

$$\begin{aligned}&2\sigma_r^{-1}(\mathcal{I}_{1,i}(\mathbf{A}))\{\|\mathbf{Z}_{i\cdot}diag(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(C_2(C_1+\xi))\}\\ \leq&\xi\leq 2^{-1}\{\gamma_{1,i}(\mathbf{A})\kappa_3(C_2(C_1+\xi)\}^{-1}\sigma_K(\mathcal{I}_{1,i}(\mathbf{A})),\end{aligned} \tag{15}$$

where we define $\mathbf{Z}_{i\cdot}=(z_{ij})_{j\in[p]}\in\mathbb{R}^{1\times p}$,

$$\mathbf{B}_{1,i}(\mathbf{A}):=\sum_{j=1}^{p}\omega_{ij}b''(m^*_{ij})\mathbf{a}_j(\mathbf{a}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*\in\mathbb{R}^r, \tag{16}$$

$$\mathcal{I}_{1,i}(\mathbf{A}):=\sum_{j=1}^{p}\omega_{ij}b''(m^*_{ij})\mathbf{a}_j(\mathbf{a}_j)^T, \tag{17}$$

and

$$\beta_{1,i}(\mathbf{A}):=\sup_{\|\mathbf{u}\|=1}\sum_j\omega_{ij}((\mathbf{a}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*)^2|\mathbf{a}_j^T\mathbf{u}|\text{ and }\gamma_{1,i}(\mathbf{A}):=\sup_{\|\mathbf{u}\|=1}\sum_j\omega_{ij}|\mathbf{a}_j^T\mathbf{u}|^3, \tag{18}$$

then, there is $\tilde{\boldsymbol{\theta}}_i$ *such that* $\|\tilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i^*\|\leq\xi$ *and* $S_{1,i}(\tilde{\boldsymbol{\theta}}_i;\mathbf{A})=\mathbf{0}$.

Proof [Proof of Lemma 15] Let $\boldsymbol{\theta}$ be a vector such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}_i^*\|=\xi$ and let $m_{ij}=\mathbf{a}_j^T\boldsymbol{\theta}_i$. Consider the Taylor expansion of $\phi S_{1,i}(\boldsymbol{\theta};\mathbf{A})$,

$$\begin{aligned}\phi S_{1,i}(\boldsymbol{\theta};\mathbf{A})=&\sum_j\omega_{ij}(y_{ij}-b'(m^*_{ij}))\mathbf{a}_j-\sum_j\omega_{ij}(b'(m_{ij})-b'(m^*_{ij}))\mathbf{a}_j\\ =&\mathbf{A}^T\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{Z}_{i\cdot}^T-\sum_j\omega_{ij}b''(m^*_{ij})(m_{ij}-m^*_{ij})\mathbf{a}_j-2^{-1}\sum_j\omega_{ij}b^{(3)}(\tilde{m}_{ij})(m_{ij}-m^*_{ij})^2\mathbf{a}_j,\end{aligned} \tag{19}$$

for some $\tilde{m}_{ij}$ between m^*_{ij} and m_{ij}. Plugging $m_{ij} - m^*_{ij} = \mathbf{a}_j^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i) + (\mathbf{a}_j - \mathbf{a}^*_j)^T\boldsymbol{\theta}^*_i$ into the above display, we obtain

$$\begin{aligned}\phi S_{1,i}(\boldsymbol{\theta};\mathbf{A}) =&\mathbf{A}^T\mathrm{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{Z}_{i\cdot}^T - \sum_j \omega_{ij} b''(m^*_{ij})\mathbf{a}_j\mathbf{a}_j^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)\\ &- \sum_j \omega_{ij} b''(m^*_{ij})\mathbf{a}_j(\mathbf{a}_j - \mathbf{a}^*_j)^T\boldsymbol{\theta}^*_i - 2^{-1}\sum_j \omega_{ij} b^{(3)}(\tilde{m}_{ij})(m_i - m^*_{ij})^2\mathbf{a}_j\end{aligned}. \tag{20}$$

Multiplying $(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T$ on both sides, we obtain

$$\begin{aligned}&\phi(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T S_{1,i}(\boldsymbol{\theta};\mathbf{A})\\ =&(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\mathbf{A}^T\mathrm{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{Z}_{i\cdot}^T - (\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\sum_j \omega_{ij} b''(m^*_{ij})\mathbf{a}_j\mathbf{a}_j^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)\\ &- (\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\sum_j \omega_{ij} b''(m^*_{ij})\mathbf{a}_j(\mathbf{a}_j - \mathbf{a}^*_j)^T\boldsymbol{\theta}^*_i\\ &- 2^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\sum_j \omega_{ij} b^{(3)}(\tilde{m}_{ij})(m_i - m^*_{ij})^2\mathbf{a}_j.\end{aligned} \tag{21}$$

Recall that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*_i\| = \xi$. Using inequalities about matrix products and singular values, we have the following upper bounds for the first three terms on the right-hand side of the above display.

$$|(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\mathbf{A}^T\mathrm{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{Z}_{i\cdot}^T| \leq \xi\|\mathbf{A}^T\mathrm{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{Z}_{i\cdot}^T\| = \xi\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|, \tag{22}$$

$$- (\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\sum_j \omega_{ij} b''(m^*_{ij})\mathbf{a}_j\mathbf{a}_j^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i) \leq -\xi^2\sigma_r(\mathcal{I}_{1,i}(\mathbf{A})), \tag{23}$$

where $\sigma_r(\mathcal{I}_{1,i}(\mathbf{A}))$ denotes the r-th largest singular value of $\mathcal{I}_{1,i}(\mathbf{A})$, and

$$|(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\sum_j \omega_{ij} b''(m^*_{ij})\mathbf{a}_j(\mathbf{a}_j - \mathbf{a}^*_j)^T\boldsymbol{\theta}^*_i| = \|(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\mathbf{B}_{1,i}\| \leq \xi\|\mathbf{B}_{1,i}\|. \tag{24}$$

Now we analyze the last term $2^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\sum_j \omega_{ij} b^{(3)}(\tilde{m}_{ij})(m_i - m^*_{ij})^2\mathbf{a}_j$. Note that $|\tilde{m}_{ij}| \leq |m_{ij^*}| \vee |m_{ij}| \leq (C_1 + \xi)C_2$ and $m_{ij} - m^*_{ij} = \mathbf{a}_j^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i) + (\mathbf{a}_j - \mathbf{a}^*_j)^T\boldsymbol{\theta}^*_i$, we have

$$\begin{aligned}&2^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\sum_j b^{(3)}(\tilde{m}_{ij})(m_i - m^*_{ij})^2\mathbf{a}_j\\ \leq&2^{-1}\kappa_3((C_1+\xi)C_2)\xi\sup_{\|\mathbf{u}\|=1}\sum_j \omega_{ij}((\mathbf{a}_j - \mathbf{a}^*_j)^T\boldsymbol{\theta}^*_i + \xi\mathbf{a}_j^T\mathbf{u})^2|\mathbf{a}_j^T\mathbf{u}|\\ \leq&\kappa_3((C_1+\xi)C_2)\{\xi\sup_{\|\mathbf{u}\|=1}\sum_j \omega_{ij}((\mathbf{a}_j - \mathbf{a}^*_j)^T\boldsymbol{\theta}^*_i)^2|\mathbf{a}_j^T\mathbf{u}| + \xi^3\sup_{\|\mathbf{u}\|=1}\sum_j \omega_{ij}|\mathbf{a}_j^T\mathbf{u}|^3\}\\ =&\kappa_3((C_1+\xi)C_2)(\xi\beta_{1,i} + \xi^3\gamma_{1,i}).\end{aligned} \tag{25}$$

Combining the analysis with (21), (22), (23), and (24), we obtain

$$\begin{aligned}(\boldsymbol{\theta} - \boldsymbol{\theta}^*_i)^T\phi S_{1,i}(\boldsymbol{\theta};\mathbf{A}) \leq& - \sigma_r(\mathcal{I}_{1,i}(\mathbf{A}))\xi^2 + \gamma_{1,i}\kappa_3((C_1+\xi)C_2)\xi^3\\ &+ \{\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\| + \|\mathbf{B}_{1,i}\| + \beta_{1,i}\kappa_3((C_1+\xi)C_2)\}\xi.\end{aligned} \tag{26}$$

Now, we view the right-hand side of the above inequality as a cubic function in ξ. For any cubic function $f(x) = -ax^2 + bx^3 + cx$ with $a, b, c > 0$, it is easy to verify that if $2c/a \le x \le a/(2b)$, then $f(x) \le 0$. Applying this result, we can see that $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_i^*\|=\xi}(\boldsymbol{\theta} - \boldsymbol{\theta}_i^*)^T S_{1,i}(\boldsymbol{\theta}; \mathbf{A}) \le 0$, if the following inequalities hold:

$$\begin{aligned} &2\sigma_K^{-1}(\mathcal{I}_{1,i}(\mathbf{A}))\{\|\mathbf{Z}_{i\cdot}\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3((C_1+\xi)C_2)\} \\ \le&\xi \le 2^{-1}\{\gamma_{1,i}\kappa_3((C_1+\xi)C_2)\}^{-1}\sigma_r(\mathcal{I}_{1,i}(\mathbf{A})). \end{aligned} \tag{27}$$

According to Result 6.3.4 in Ortega and Rheinboldt (2000), $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_i^*\|=\xi}(\boldsymbol{\theta} - \boldsymbol{\theta}_i^*)^T S_{1,i}(\boldsymbol{\theta}; \mathbf{A}) \le 0$ implies that there is a solution $S_{1,i}(\tilde{\boldsymbol{\theta}}; \mathbf{A}) = \mathbf{0}$ satisfying $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_i^*\| \le \xi$. ■

Next, we simplify the result of Lemma 15 to obtain a more user-friendly version in the next lemma.

Lemma 16. *Let* $\boldsymbol{\Theta}^* \in \mathbb{R}^{n\times r}$ *and* $\mathbf{A}^* \in \mathbb{R}^{p\times r}$ *be such that* $\mathbf{M}^* = \boldsymbol{\Theta}^*(\mathbf{A}^*)^T$ *and* $\mathbf{Z} = (z_{ij})$ *with* $z_{ij} = y_{ij} - b'(m_{ij}^*)$. *If* $\|\mathbf{A}^*\|_{2\to\infty} \le C_2$ *and* $\|\mathbf{A}\|_{2\to\infty} \le C_2$, *and*

$$\begin{aligned} &\|\mathbf{Z}_{i\cdot}\mathit{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(3C_1C_2) \\ \le&\min\left\{2^{-2}(\gamma_{1,i}(\mathbf{A}))^{-1}(\kappa_3(3C_1C_2))^{-1}\sigma_r^2(\mathcal{I}_{1,i}(\mathbf{A})), 2^{-1}\sigma_r(\mathcal{I}_{1,i}(\mathbf{A}))C_1\right\}, \end{aligned} \tag{28}$$

then, there is $\tilde{\boldsymbol{\theta}}_i$ *such that* $S_{1,i}(\tilde{\boldsymbol{\theta}}; \mathbf{A}) = \mathbf{0}$, *and*

$$\|\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*\| \le 2\sigma_r^{-1}(\mathcal{I}_{1,i}(\mathbf{A}))\{\|\mathbf{Z}_{i\cdot}\mathit{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(3C_1C_2)\}. \tag{29}$$

Moreover, the solution $\tilde{\boldsymbol{\theta}}_i$ *also satisfies* $\|\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*\| \le C_1$.

Proof [Proof of Lemma 16]

Let $\xi = 2\sigma_r^{-1}(\mathcal{I}_{1,i}(\mathbf{A}))\{\|\mathbf{Z}_{i\cdot}\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(3C_1C_2)\}$. By the assumption that $\|\mathbf{Z}_{i\cdot}\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(3C_1C_2) \le 2^{-1}\sigma_r(\mathcal{I}_{1,i})C_1$, we have $\xi \le C_1$. Thus,

$$\kappa_3(C_2(C_1+\xi)) \le \kappa_3(3C_1C_2). \tag{30}$$

This implies

$$\begin{aligned} &2\sigma_r^{-1}(\mathcal{I}_{1,i}(\mathbf{A}))\{\|\mathbf{Z}_{i\cdot}\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(C_2(C_1+\xi))\} \\ \le&2\sigma_r^{-1}(\mathcal{I}_{1,i}(\mathbf{A}))\{\|\mathbf{Z}_{i\cdot}\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(3C_1C_2)\}. \end{aligned} \tag{31}$$

Because the right-hand side of the above inequality equals ξ, it is simplified as

$$2\sigma_r^{-1}(\mathcal{I}_{1,i}(\mathbf{A}))\{\|\mathbf{Z}_{i\cdot}\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(C_2(C_1+\xi))\} \le \xi. \tag{32}$$

On the other hand, according to the assumption that $\|\mathbf{Z}_{i\cdot}\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(3C_1C_2) \le 2^{-2}\gamma_{1,i}^{-1}(\kappa_3(3C_1C_2))^{-1}\sigma_r^2(\mathcal{I}_{1,i}(\mathbf{A}))$, we further have

$$\begin{aligned} \xi =&2\sigma_r^{-1}(\mathcal{I}_{1,i}(\mathbf{A}))\{\|\mathbf{Z}_{i\cdot}\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\|+\|\mathbf{B}_{1,i}(\mathbf{A})\|+\beta_{1,i}(\mathbf{A})\kappa_3(3C_1C_2)\} \\ \le&2^{-1}\gamma_{1,i}^{-1}(\kappa_3(3C_1C_2))^{-1}\sigma_r(\mathcal{I}_{1,i}(\mathbf{A})) \\ \le&2^{-1}\{\gamma_{1,i}(\mathbf{A})\kappa_3(C_2(C_1+\xi)\}^{-1}\sigma_r(\mathcal{I}_{1,i}(\mathbf{A})). \end{aligned} \tag{33}$$

Equations (32) and (33) together imply (15). By Lemma 15, there is $\tilde{\boldsymbol{\theta}}_i$ such that $\|\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*\| \leq \xi$ and $S_{1,i}(\tilde{\boldsymbol{\theta}}; \mathbf{A}) = \mathbf{0}$. We complete the proof by noting that $\xi = 2\sigma_r^{-1}(\mathcal{I}_{1,i}(\mathbf{A}))\{\|\mathbf{Z}_{i\cdot}\text{diag}(\boldsymbol{\Omega}_{i\cdot})\mathbf{A}\| + \|\mathbf{B}_{1,i}(\mathbf{A})\| + \beta_{1,i}(\mathbf{A})\kappa_3(3C_1C_2)\} \leq 2\sigma_r^{-1}(\mathcal{I}_{1,i}) \cdot 2^{-1}\sigma_r(\mathcal{I}_{1,i}(\mathbf{A}))C_1 = C_1$. ■

By symmetry, we also have the following non-probabilistic and non-asymptotic analysis for $\tilde{\mathbf{A}}$. For each $j \in [p]$, the estimating equation for $\mathbf{a}_j$ based on $\boldsymbol{\Theta}_{\mathcal{N}_2}$ and $\boldsymbol{\Omega}_{\mathcal{N}_2\cdot}$ is defined as

$$S_{2,j}(\mathbf{a}_j; \boldsymbol{\Theta}_{\mathcal{N}_2}) := \phi^{-1} \sum_{i \in \mathcal{N}_2} \omega_{ij}\{y_{ij} - b'(\mathbf{a}_j^T \boldsymbol{\theta}_i)\}\boldsymbol{\theta}_i. \tag{34}$$

Let

$$\mathbf{B}_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2}) = \sum_{i \in \mathcal{N}_2} \omega_{ij} b''(m_{ij}^*)\boldsymbol{\theta}_i(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)^T \mathbf{a}_j^* \in \mathbb{R}^r, \tag{35}$$

$$\mathcal{I}_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2}) = \sum_{i \in \mathcal{N}_2} \omega_{ij} b''(m_{ij}^*)\boldsymbol{\theta}_i(\boldsymbol{\theta}_i)^T, \tag{36}$$

and

$$\beta_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2}) = \sup_{\|\mathbf{u}\|=1} \sum_{i \in \mathcal{N}_2} \omega_{ij}((\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^*)^T \mathbf{a}_j^*)^2 |\boldsymbol{\theta}_j^T \mathbf{u}| \text{ and } \gamma_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2}) = \sup_{\|\mathbf{u}\|=1} \sum_{i \in \mathcal{N}_2} \omega_{ij} |\boldsymbol{\theta}_i^T \mathbf{u}|^3, \tag{37}$$

Lemma 17. *Let $\boldsymbol{\Theta}_{\mathcal{N}_2}^*$ and $\mathbf{A}^*$ be such that $\mathbf{M}_{\mathcal{N}_2\cdot}^* = \boldsymbol{\Theta}_{\mathcal{N}_2}^*(\mathbf{A}^*)^T$ and $\mathbf{Z} = (z_{ij})$ with $z_{ij} = y_{ij} - b'(m_{ij}^*)$ and $diag(\boldsymbol{\Omega}_{\mathcal{N}_2,j}) := diag((\omega_{ij})_{i \in \mathcal{N}_2})$. If $\|\boldsymbol{\Theta}_{\mathcal{N}_2}\|, \|\boldsymbol{\Theta}_{\mathcal{N}_2}^*\|_{2\to\infty} \leq C_1$, $\|\mathbf{A}^*\|_{2\to\infty} \leq C_2$ and*

$$\begin{aligned} &\|\mathbf{Z}_{\mathcal{N}_2,j}^T diag(\boldsymbol{\Omega}_{\mathcal{N}_2,j})\boldsymbol{\Theta}_{\mathcal{N}_2}\| + \|\mathbf{B}_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2})\| + \beta_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2})\kappa_3(3C_1C_2) \\ &\leq \min\left\{2^{-2}\gamma_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2})^{-1}(\kappa_3(3C_1C_2))^{-1}\sigma_r^2(\mathcal{I}_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2})), 2^{-1}\sigma_r(\mathcal{I}_{2,j}(\mathbf{A}))C_2\right\} \end{aligned} \tag{38}$$

where $\mathbf{Z}_{\mathcal{N}_2,j} = (z_{ij})_{i \in \mathcal{N}_2}$, then, there is $\tilde{\mathbf{a}}$ such that $S_{2,j}(\tilde{\mathbf{a}}; \boldsymbol{\Theta}_{\mathcal{N}_2}) = \mathbf{0}_r$, and

$$\|\tilde{\mathbf{a}}_j - \mathbf{a}_j^*\| \leq 2\sigma_r^{-1}(\mathcal{I}_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2})))\{\|\mathbf{Z}_{\mathcal{N}_2,j}^T diag(\boldsymbol{\Omega}_{\mathcal{N}_2,j})\boldsymbol{\Theta}_{\mathcal{N}_2}\| + \|\mathbf{B}_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2})\| + \beta_{2,j}(\boldsymbol{\Theta}_{\mathcal{N}_2})\kappa_3(3C_1C_2)\}. \tag{39}$$

Moreover, $\tilde{\mathbf{a}}_j$ satisfies that $\|\tilde{\mathbf{a}}_j - \mathbf{a}_j^*\| \leq C_2$.

Proof [Proof of Lemma 17] The lemma follows similar proof as that of Lemma 15 and Lemma 16 with $(\mathbf{A}, \mathbf{A}^*, C_1, C_2)$ replaced by $(\boldsymbol{\Theta}_{\mathcal{N}_2}, \boldsymbol{\Theta}_{\mathcal{N}_2}^*, C_2, C_1)$. We omit the details. ■

A.3 Non-asymptotic Probablistic Analysis

Recall that $\mathbf{M}^*$ has the SVD $\mathbf{M}^* = \mathbf{U}_r^*\mathbf{D}_r^*\mathbf{V}_r^*$. In this section, we first provide non-asymptotic bounds for each term in Lemma 16 with $\mathbf{A}$ replaced by $\hat{\mathbf{A}}$ and $\mathbf{A}^*$ replaced by $\mathbf{V}_r^*\hat{\mathbf{P}}$ where $\hat{\mathbf{P}}$ is defined in (11). Recall that $\hat{\mathbf{A}} = \hat{\mathbf{A}}^{(1)}$ is constructed based on $\hat{\mathbf{M}}_{\mathcal{N}_1\cdot}$ using data $\{Y_{ij}\omega_{ij}, \omega_{ij}\}_{i \in \mathcal{N}_1, j \in [p]}$, and thus, independent with $\{y_{ij}, \omega_{ij}\}_{j \in [p]}$ for all $i \in \mathcal{N}_2$. The results in this section hold in general for any estimator $\hat{\mathbf{A}}$ that is independent with $\{\omega_{ij}, Y_{ij}\omega_{ij}\}_{i \in \mathcal{N}_2, j \in [p]}$, including the proposed one.

After the analysis for terms in Lemma 16, we provide non-asymptotic analysis for terms in Lemma 17 with $\boldsymbol{\Theta}_{\mathcal{N}_2}$ replaced by $\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}$ and $\boldsymbol{\Theta}_{\mathcal{N}_2}^*$ replaced by $\mathbf{U}_r^*\mathbf{D}_r^*\hat{\mathbf{P}}$. Unlike $\hat{\mathbf{A}}$, $\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}$ is dependent with $\{y_{ij}, \omega_{ij}\}_{i \in [p]}$ for $i \in \mathcal{N}_2$. Thus, we will take a different approach for the error analysis of $\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}$.

A.3.1 NON-ASYMPTOTIC BOUND FOR TERMS IN LEMMA 16

Lemma 18 (Upper bound for $\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|$ with data splitting)**.** *Assume $n \geq 2$. $\|\mathbf{M}^*\|_{\max} \leq \rho$ and $\|\hat{\mathbf{A}}\|_{2\to\infty} \leq C_2$. Then, with probability at least $1-(nr)^{-1}$,*

$$\max_{i\in\mathcal{N}_2}\|\mathbf{Z}_{i\cdot}\mathit{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| \leq 8\{\phi^{1/2}(\kappa_2(2\rho+1))^{1/2}C_2\log^{1/2}(nr)r^{1/2}p_{\max}^{1/2} \vee r^{1/2}\phi C_2/(\rho+1)\log(nr)\} \tag{40}$$

where $p_{\max} = \max_{i\in[n]}\sum_j \omega_{ij}$ denotes the maximum number of observations in each row.

Proof [Proof of Lemma 18] We first verify that under the generalized latent factor model, $\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}_{\cdot k}$ is sub-exponential given $\mathbf{\Omega}_{\mathcal{N}_2\cdot} = (\omega_{ij})_{i\in\mathcal{N}_2, j\in[p]}$ and $\hat{\mathbf{A}}$. To see this, consider the moment generating function

$$\begin{aligned}
&\mathbb{E}[\exp(\lambda\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}_{\cdot k})|\mathbf{\Omega}_{\mathcal{N}_2\cdot}, \hat{\mathbf{A}}]\\
=&\prod_{j\in[p]}\mathbb{E}[\lambda Z_{ij}\hat{a}_{jk}\omega_{ij}|\mathbf{\Omega}_{\mathcal{N}_2\cdot}, \hat{\mathbf{A}}]\\
=&\exp\Big[\phi^{-1}\sum_j\omega_{ij}\{b(m^*_{ij}+\lambda\hat{a}_{jk}\phi) - b(m^*_{ij}) - \lambda\hat{a}_{jk}\phi b'(m^*_{ij})\}\Big]\\
=&\exp[2^{-1}\lambda^2\phi\sum_j\omega_{ij}b''(\tilde{m}_{ij})(\hat{a}_{jk})^2]
\end{aligned} \tag{41}$$

for some $\tilde{m}_{ij}$ between m^*_{ij} and $m^*_{ij}+\lambda\hat{a}_{jk}\phi$. Note that here we used the independence between $\hat{\mathbf{A}}$ and $\{z_{ij}\omega_{ij}\}_{i\in\mathcal{N}_2}$ in the first and second equations.

Because $|m^*_{ij}| \leq \rho$ and $|\hat{a}_{jk}| \leq C_2$, for $|\lambda| \leq (\rho+1)/(\phi C_2)$, $\tilde{m}_{ij} \leq \rho + \lambda\phi C_2 \leq 2\rho+1$. Thus, $\mathbb{E}[\exp(\lambda\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}_{\cdot k})|\mathbf{\Omega}_{\mathcal{N}_2\cdot}, \hat{\mathbf{A}}] \leq \exp\{\lambda^2\phi\sum_j\omega_{ij}(\hat{a}_{jk})^2\kappa_2(2\rho+1)/2\}$ for $|\lambda| \leq (\rho+1)/(\phi C_2)$. This implies that $\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}_{\cdot k}$ is sub-exponential (conditional on $(\mathbf{\Omega}_{\mathcal{N}_2\cdot}, \hat{\mathbf{A}})$) with parameters $\nu^2_{ik} = \phi\kappa_2(2\rho+1)\sum_j\omega_{ij}(\hat{a}_{jk})^2 \leq C_2^2\phi\kappa_2(2\rho+1)p_{\max}$ and $\alpha = \phi C_2/(\rho+1)$.

Applying tail probability bound for sub-exponential random variables to $\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}_{\cdot k}$, we have

$$\mathbb{P}(|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}_{\cdot k}| \geq t|\mathbf{\Omega}_{\mathcal{N}_2\cdot}, \hat{\mathbf{A}}) \leq 2(e^{-t^2/(2\nu^2_{ik})} \vee e^{-t/(2\alpha)}) \tag{42}$$

for all positive t. This implies

$$\begin{aligned}
&\mathbb{P}(\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| \geq t|\mathbf{\Omega}_{\mathcal{N}_2\cdot}, \hat{\mathbf{A}})\\
\leq&\sum_{k\in[r]}\mathbb{P}(|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}_{\cdot k}| \geq t/\sqrt{r}|\mathbf{\Omega}_{\mathcal{N}_2\cdot}, \hat{\mathbf{A}})\\
\leq& r\cdot 2(e^{-t^2/(2r\max_k\nu^2_{ik})} \vee e^{-t/(2r^{1/2}\alpha)}).
\end{aligned} \tag{43}$$

Combining results for different i with a union bound, we have

$$\mathbb{P}\Big(\max_{i\in\mathcal{N}_2}\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| \geq t|\mathbf{\Omega}_{\mathcal{N}_1\cdot}, \hat{\mathbf{A}}\Big) \leq 2rn\cdot(e^{-t^2/(2r\max_k\nu^2_{ik})} \vee e^{-t/(2r^{1/2}\alpha)}). \tag{44}$$

For $t = \{8(\log(nr)r\max_{k\in[r]}\nu^2_{ik})^{1/2}\} \vee 8r^{1/2}\alpha\log(nr)$ and $n \geq 2$, the right-hand side of the above inequality is no larger than $(nr)^{-1}$. Because $\nu^2_{ik} \leq \phi\kappa_2(2\rho+1)C_2^2p_{\max}$, we obtain

$$\max_{i\in\mathcal{N}_2}\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| \leq 8\{\phi^{1/2}(\kappa_2(2\rho+1))^{1/2}C_2\log^{1/2}(nr)r^{1/2}p_{\max}^{1/2} \vee r^{1/2}\phi C_2/(\rho+1)\log(nr)\} \tag{45}$$

with probability at least $1-(nr)^{-1}$.

■

Lemma 19 (Upper bound for $\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|$ with data splitting)**.** *Let* $\mathbf{A}^* = \mathbf{V}_r^*\hat{\mathbf{P}}$ *and* $\mathbf{\Theta}^* = \mathbf{U}_r^*\mathbf{D}_r^*\hat{\mathbf{P}}$. *If* $\hat{\mathbf{A}}$ *is independent with* $\{\omega_{ij}\}_{j\in[p]}$ *for* $i \in \mathcal{N}_2$, $\|\hat{\mathbf{A}}\|_{2\to\infty}, \|\mathbf{V}_r^*\|_{2\to\infty} \le C_2$ *and* $\|\mathbf{U}_r\mathbf{D}_r^*\|_{2\to\infty} \le C_1$, *then, for* $n \ge 4$ *with probability at least* $1 - 1/(nr)$,

$$\max_{i\in\mathcal{N}_2}\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\| \le \kappa_2^*\pi_{\max}C_1\|\hat{\mathbf{A}}\|_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F + 64\log(n)\cdot(\pi_{\max}^{1/2}\kappa_2^*C_1C_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F + \kappa_2^*C_1C_2^2) \tag{46}$$

Proof [Proof of Lemma 19] First, by the assumptions and $\hat{\mathbf{P}}$ is orthogonal, $\|\mathbf{\Theta}^*\|_{2\to\infty} = \|\mathbf{U}_r^*\mathbf{D}_r^*\|_{2\to\infty} \le C_1$ and $\|\mathbf{A}^*\|_{2\to\infty} = \|\mathbf{V}_r^*\|_{2\to\infty} \le C_2$. Let

$$\mathbf{S}_j = (\omega_{ij} - \pi_{ij})b''(m_{ij}^*)\hat{\mathbf{a}}_j(\hat{\mathbf{a}}_j - \mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*. \tag{47}$$

Then,

$$\mathbf{B}_{1,i}(\hat{\mathbf{A}}) = \sum_{j=1}^{p}\omega_{ij}b''(m_{ij}^*)\hat{\mathbf{a}}_j(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^* = \sum_{j\in[p]}\mathbf{S}_j + \sum_{j\in[p]}\pi_{ij}b''(m_{ij}^*)\hat{\mathbf{a}}_j(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*. \tag{48}$$

Note that $\mathbf{S}_j$ are independent mean zero random vectors for $j \in [p]$ (conditional on $\hat{\mathbf{A}}$) and

$$\|\mathbf{S}_j\| \le 4\kappa_2^*C_1C_2^2. \tag{49}$$

This allow us to apply the matrix Bernstein inequality (Equation (6.1.5) in Tropp (2015)) to $\sum_{j\in[p]}\mathbf{S}_j \in \mathbb{R}^r$, and obtain

$$\mathbb{P}\Big(\|\sum_{j\in[p]}\mathbf{S}_j\| \ge t|\hat{\mathbf{A}}\Big) \le (r+1)\cdot e^{-\frac{3t^2}{8\nu}} \vee e^{-\frac{3t}{8L}} \le 2r\cdot e^{-\frac{3t^2}{8\nu}} \vee e^{-\frac{3t}{8L}} \tag{50}$$

for $t > 0$ where $\nu = \max\Big\{\Big\|\sum_{j\in[p]}E\{\mathbf{S}_j\mathbf{S}_j^T|\hat{\mathbf{A}}\}\Big\|_2, \Big\|\sum_{j\in[p]}E\{\mathbf{S}_j^T\mathbf{S}_j|\hat{\mathbf{A}}\}\Big\|_2\Big\}$ and $L = 4\kappa_2^*C_1C_2^2 \ge \|\mathbf{S}_j\|$ for all j. Thus, for any $0 < \epsilon < r$

$$\mathbb{P}\Big(\|\sum_{j\in[p]}\mathbf{S}_j\| \ge \{8/3\cdot\log(2r/\epsilon)\}^{1/2}\nu^{1/2} \vee \{(8/3\cdot\log(2r/\epsilon))L\}|\hat{\mathbf{A}}\Big) \le \epsilon. \tag{51}$$

Now we find an upper bound for ν. Since

$$\mathbb{E}\{\mathbf{S}_j\mathbf{S}_j^T|\hat{\mathbf{A}}\} = \pi_{ij}(1-\pi_{ij})\cdot\{b''(m_{ij}^*)\}^2\hat{\mathbf{a}}_j(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*(\boldsymbol{\theta}_i^*)^T(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)\hat{\mathbf{a}}_j^T, \tag{52}$$

and

$$\mathbb{E}\{\mathbf{S}_j^T\mathbf{S}_j|\hat{\mathbf{A}}\} = \pi_{ij}(1-\pi_{ij})\cdot\{b''(m_{ij}^*)\}^2(\boldsymbol{\theta}_i^*)^T(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)\hat{\mathbf{a}}_j^T\hat{\mathbf{a}}_j(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*, \tag{53}$$

we have

$$\max\Big\{\|\mathbb{E}\{\mathbf{S}_j^T\mathbf{S}_j|\hat{\mathbf{A}}\}\|_2, \|\mathbb{E}\{\mathbf{S}_j\mathbf{S}_j^T|\hat{\mathbf{A}}\}\|_2\Big\} \le \pi_{\max}(\kappa_2(\rho))^2C_1^2C_2^2\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|^2 \tag{54}$$

which implies

$$\nu = \max\Big\{\Big\|\sum_{j\in[p]}\mathbb{E}\{\mathbf{S}_j\mathbf{S}_j^T|\hat{\mathbf{A}}\}\Big\|_2, \Big\|\sum_{j\in[p]}\mathbb{E}\{\mathbf{S}_j^T\mathbf{S}_j|\hat{\mathbf{A}}\}\Big\|_2\Big\} \le \pi_{\max}(\kappa_2(\rho))^2 C_1^2 C_2^2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2. \tag{55}$$

Combine the above inequality with (51), we have that with probability at least $1-\epsilon$,

$$\|\sum_{j\in[p]}\mathbf{S}_j\|\le \{8/3\cdot\log(2r/\epsilon)\}^{1/2}\pi_{\max}^{1/2}\kappa_2(\rho)C_1C_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+\{(8/3\cdot\log(2r/\epsilon))\}\cdot 4\kappa_2(\rho)C_1C_2^2 \tag{56}$$

for any $0<\epsilon<r$. Simplifying this inequality, we get that with probability at least $1-\epsilon$,

$$\|\sum_{j\in[p]}\mathbf{S}_j\|\le \{16\cdot\log(r/\epsilon)\}\cdot(\pi_{\max}^{1/2}\kappa_2(\rho)C_1C_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+\kappa_2(\rho)C_1C_2^2) \tag{57}$$

for $\epsilon\in(0,r/10)$.

Next, we obtain an upper bound for $\|\sum_{j\in[p]}\pi_{ij}b''(m_{ij}^*)\hat{\mathbf{a}}_j(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*\|$ as

$$\begin{aligned}
&\|\sum_{j\in[p]}\pi_{ij}b''(m_{ij}^*)\hat{\mathbf{a}}_j(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*\|\\
\le& C_1\|\sum_{j\in[p]}\pi_{ij}b''(m_{ij}^*)\hat{\mathbf{a}}_j(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\|_2\\
=& C_1\|\hat{\mathbf{A}}^T\mathrm{diag}(\pi_{i1}b''(m_{i1}^*),\cdots,\pi_{ip}b''(m_{ip}^*))(\hat{\mathbf{A}}-\mathbf{A}^*)\|_2\\
\le& C_1\|\hat{\mathbf{A}}\|_2\pi_{\max}\kappa_2^*\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F
\end{aligned} \tag{58}$$

Combine the above inequality with (48) and (57), we have

$$\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|\le \kappa_2^*\pi_{\max}C_1\|\hat{\mathbf{A}}\|_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+\{16\cdot\log(r/\epsilon)\}\cdot(\pi_{\max}^{1/2}\kappa_2^*C_1C_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+\kappa_2^*C_1C_2^2) \tag{59}$$

with probability at least $1-\epsilon$ for $\epsilon\in(0,r/10)$. We complete the proof using a union bound for $i\in\mathcal{N}_2$ and $\epsilon=1/(rn^2)$. ■

Remark 20. *The first term $\kappa_2^*\pi_{\max}C_1\|\hat{\mathbf{A}}\|_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F$ in the upper bound is the leading term in the error analysis. To obtain this error bound, we need $\{\omega_{ij}\}_{j\in[p]}$ to be independent with $\hat{\mathbf{A}}$. In contrast, if $\{\omega_{ij}\}_{j\in[p]}$ are dependent with $\hat{\mathbf{A}}$, then the the leading term in the error analysis may be larger (at the order $1/\sqrt{\pi_{\max}}$ in the worst case).*

Lemma 21 (Upper bound for $\beta_{1,i}(\hat{\mathbf{A}})$ with data splitting)**.** *If $\|\mathbf{U}_r^*\mathbf{D}_r^*\|_{2\to\infty}\le C_1$, $\|\hat{\mathbf{A}}\|_{2\to\infty},\|\mathbf{V}_r^*\|_{2\to\infty}\le C_2$, and $\hat{\mathbf{A}}$ is independent with $\{\omega_{ij}\}_{i\in\mathcal{N}_2,j\in[p]}$, then, with probability at least $1-1/n$,*

$$\max_{i\in\mathcal{N}_2}\beta_{1,i}(\hat{\mathbf{A}})\le C_1^2C_2\{\pi_{\max}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2+4\pi_{\max}^{1/2}C_2(\log(n))^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F 4C_2^2\log(n)\}. \tag{60}$$

Proof [Proof of Lemma 21] Recall

$$\beta_{1,i}(\hat{\mathbf{A}})=\sup_{\|\mathbf{u}\|=1}\sum_j\omega_{ij}((\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*)^2|\hat{\mathbf{a}}_j^T\mathbf{u}|\le C_1^2C_2\sum_{j\in[p]}\omega_{ij}\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|^2. \tag{61}$$

Conditional on $\hat{\mathbf{A}}$, $(\omega_{ij}-\pi_{ij})\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|^2$ are independent, mean-zero, bounded by $4C_2^2$, and has the variance $\pi_{ij}(1-\pi_{ij})\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|^4\leq 4\pi_{ij}C_2^2\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|^2$. By Bernstein's inequality for bounded random variables (Theorem 2.10 in Boucheron et al. (2013) with $c=4C_2^2/3$ and $v=4\pi_{ij}C_2^2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2$), for $t>0$

$$\mathbb{P}\Big(\sum_{j\in[p]}(\omega_{ij}-\pi_{ij})\|\hat{\mathbf{a}}_j-\mathbf{a}^*\|^2\geq (8\pi_{ij}C_2^2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2t)^{1/2}+4/3\cdot C_2^2t|\hat{\mathbf{A}}\Big)\leq e^{-t}. \tag{62}$$

Let $t=2\log(n)$ in the above inequality and note that $\pi_{ij}\leq\pi_{\max}$ and $4/3<2$, we have that with probability at least $1-1/n^2$,

$$\sum_{j\in[p]}(\omega_{ij}-\pi_{ij})\|\hat{\mathbf{a}}_j-\mathbf{a}^*\|^2\leq 4\pi_{\max}^{1/2}C_2(\log(n))^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+4C_2^2\log(n). \tag{63}$$

This implies that with probability at least $1-1/n^2$,

$$\begin{aligned}&\sum_{j\in[p]}\omega_{ij}\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|^2\\ \leq&\sum_{j\in[p]}\pi_{ij}\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|^2+4\pi_{\max}^{1/2}C_2(\log(n))^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+4C_2^2\log(n)\\ \leq&\pi_{\max}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2+4\pi_{\max}^{1/2}C_2(\log(n))^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+4C_2^2\log(n).\end{aligned} \tag{64}$$

We complete the proof by combining the above inequality with (61) and applying a union bound for $i\in\mathcal{N}_2$. ∎

Remark 22. *Similar to Remark 20, the above analysis also requires the independence of* $\{\omega_{ij}\}_{j\in[p]}$ *and* $\hat{\mathbf{A}}$ *in order to obtain the leading term* $C_1^2C_2\pi_{\max}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2$.

Lemma 23 (Upper bound for $p_{\max}$). *Recall* $p_{\max}=\max_{i\in[n]}p_i$. *If* $p\pi_{\max}\geq 6\log n$, *then*

$$\mathbb{P}(p_{\max}\geq 2p\pi_{\max})\leq 1/n. \tag{65}$$

Proof [Proof of Lemma 23] First note that $|\omega_{ij}-\pi_{ij}|\leq 1$ and $p_i-\mathbb{E}(p_i)=\sum_j(\omega_{ij}-\pi_{ij})$. We apply the Bernstein inequality (Corollary 2.11 in Boucheron et al. (2013)) and obtain

$$\mathbb{P}(p_i-\mathbb{E}(p_i)\geq p\pi_{\max})\leq\exp\Big\{-\frac{(p\pi_{\max})^2/2}{\sum_j\mathbb{E}(\omega_{ij}-p_{ij})^2+(p\pi_{\max})/3}\Big\}. \tag{66}$$

Because $\sum_j\mathbb{E}(\omega_{ij}-p_{ij})^2=\sum_j Var(\omega_{ij})\leq\sum_j\pi_{ij}\leq p\pi_{\max}$, the above inequality implies,

$$\mathbb{P}(p_i-\mathbb{E}(p_i)\geq p\pi_{\max})\leq\exp\Big\{-\frac{(p\pi_{\max})^2/2}{(p\pi_{\max})+(p\pi_{\max})/3}\Big\}=\exp\big(-\frac{3}{8}p\pi_{\max}\big), \tag{67}$$

which further implies

$$\mathbb{P}(p_i\geq 2p\pi_{\max})\leq\exp(-3p\pi_{\max}/8). \tag{68}$$

Applying a union bound to the above inequality for $i \in [n]$, we obtain

$$\mathbb{P}(\max_{i\in[n]} p_i \geq 2p\pi_{\max}) \leq n\exp(-3p\pi_{\max}/8) \leq 1/n, \tag{69}$$

where the last inequality is due to the assumption that $p\pi_{\max} \geq 6\log n > 16/3 \log n$. ■

Lemma 24 (Upper bound of $\gamma_{1,i}(\hat{\mathbf{A}})$). *If* $\|\hat{\mathbf{A}}\|_{2\to\infty} \leq C_2$ *and* $p\pi_{\max} > 6\log n$*, then with probability at least* $1 - 1/n$*,*

$$\gamma_{1,i}(\hat{\mathbf{A}}) \leq 2p\pi_{\max} C_2^3. \tag{70}$$

Proof [Proof of Lemma 24] The lemma follows by Lemma 23 and the following inequality

$$\gamma_{1,i}(\hat{\mathbf{A}}) = \sup_{\|\mathbf{u}\|=1} \sum_{i=1}^{p} \omega_{ij} |\hat{\mathbf{a}}_j^T \mathbf{u}|^3 \leq p_{\max} C_2^3. \tag{71}$$

■

The next three lemmas together give a lower bound for $\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))$

Lemma 25. *If* $\|\text{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}} - \mathbf{A}^*)\|_2 \leq 2^{-1}\sigma_r(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*)$ *and* $\|\mathbf{M}^*\|_{\max} \leq \rho$*, then*

$$\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}})) \geq 2^{-2}\delta_2(\rho)\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*). \tag{72}$$

Proof [Proof of Lemma 25] For any $|\mathbf{u}| = 1$ and $\mathbf{u} \in \mathbb{R}^r$,

$$\mathbf{u}^T\mathcal{I}_{1,i}(\hat{\mathbf{A}})\mathbf{u} = \sum_{j=1}^{p} \omega_{ij} b''(m_{ij}^*)(\mathbf{u}^T\hat{\mathbf{a}}_j)^2 \geq \delta_2(\rho)\sum_{j=1}^{p} \omega_{ij}(\mathbf{u}^T\hat{\mathbf{a}}_j)^2 \geq \delta_2(\rho)\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}). \tag{73}$$

This implies $\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}})) \geq \delta_2(\rho)\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}})$. By Weyl's inequality, $\sigma_r(\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}) \geq \sigma_r(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) - \|\text{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}} - \mathbf{A}^*)\|_2$. Thus, if $\|\text{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}} - \mathbf{A}^*)\|_2 \leq 2^{-1}\sigma_r(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*)$, then $\sigma_r(\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}) \geq 2^{-1}\sigma_r(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*)$, and thus,

$$\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}})) \geq \delta_2(\rho)\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}) \geq 2^{-2}\delta_2(\rho)\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*). \tag{74}$$

■

The next two lemmas give a lower bound for $\sigma_r(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*)$ and an upper bound for $\|\text{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}} - \mathbf{A}^*)\|_2$.

Lemma 26. *Let* $\mathbf{A}^* = \mathbf{V}_r^*\hat{\mathbf{P}}$ *and let* $\mathbf{\Pi}_{1,i} = \text{diag}(\pi_{i1}, \cdots, \pi_{ip}) = \mathbb{E}(\text{diag}(\mathbf{\Omega}_{i\cdot}))$ *and* $\lambda_{i,\min}^* = \lambda_r((\mathbf{V}_r^*)^T\mathbf{\Pi}_{1,i}\mathbf{V}_r^*) = \lambda_r((\mathbf{A}^*)^T\mathbf{\Pi}_{1,i}\mathbf{A}^*)$*, where* $\lambda_r(\cdot)$ *denotes the r-th largest eigenvalue of a symmetric matrix. If* $\lambda_{\min}^* := \min_{i\in[n]}\lambda_{i,\min}^* \geq 16\|\mathbf{V}_r^*\|_{2\to\infty}^2\log(nr)$*, then*

$$\mathbb{P}\Big(\min_{i\in[n]}\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) \leq 2^{-1}\lambda_{\min}^*\Big) \leq 1/(nr) \tag{75}$$

Moreover, if $\pi_{\min}\sigma_r^2(\mathbf{A}^*) \geq 32\|\mathbf{A}^*\|_{2\to\infty}^2\log(n)$ *and* $n \geq r$*, then*

$$\mathbb{P}\Big(\min_{i\in[n]}\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) \leq 2^{-1}\pi_{\min}\sigma_r^2(\mathbf{A}^*)\Big) \leq 1/(nr). \tag{76}$$

Remark 27. *In the 'moreover part' of the above lemma, $\sigma_r^2(\mathbf{A}^*) = \sigma_r^2(\mathbf{V}_r^*\hat{\mathbf{P}}) = 1$, so it is possible to further simplify the statement of lemma. We keep the current form without simplification so that similar results can be obtained by symmetry for $\mathbf{\Theta}^* = \mathbf{U}_r^*\mathbf{D}_r^*\hat{\mathbf{P}}$, which will be useful for the analysis later.*

Proof [Proof of Lemma 26] First note that $\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) = \sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{V}_r^*\hat{\mathbf{P}}) = \sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{V}_r^*) = \lambda_r((\mathbf{V}_r^*)^T\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{V}_r^*)$. Also note that for all $t \in (0,1)$

$$\begin{aligned}&\mathbb{P}\Big(\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{V}_r^*) \leq (1-t)\lambda_{i,\min}^*\Big)\\ =&\mathbb{P}\Big(\lambda_r(\sum_j \omega_{ij}\mathbf{v}_j^*(\mathbf{v}_j^*)^T) \leq (1-t)\cdot\lambda_r(\sum_j \pi_{ij}\mathbf{v}_j^*(\mathbf{v}_j^*)^T)\Big),\end{aligned} \tag{77}$$

where $\mathbf{v}_j^* \in \mathbb{R}^r$ denotes the j-th row of $\mathbf{V}_r^*$. Note that $\lambda_r\{\mathbb{E}(\sum_{j\in[p]}\omega_{ij}\mathbf{v}_j^*(\mathbf{v}_j^*)^T)\} = \lambda_{i,\min}^*$, $\lambda_1(\omega_{ij}\mathbf{v}_j^*(\mathbf{v}_j^*)^T) \leq \|\mathbf{V}_r^*\|_{2\to\infty}^2$, and $\omega_{ij}\mathbf{v}_j^*(\mathbf{v}_j^*)^T$ are independent for different j. Applying Remark 5.3 in Tropp (2012) to the above probability, we obtain that for all $t \in (0,1)$,

$$\mathbb{P}\Big(\lambda_r(\sum_j \omega_{ij}\mathbf{v}_j^*(\mathbf{v}_j^*)^T) \leq (1-t)\cdot\lambda_{i,\min}^*\Big) \leq r\exp\Big\{-2^{-1}\|\mathbf{V}_r^*\|_{2\to\infty}^{-2}(1-t)^2\lambda_{i,\min}^*\Big\}. \tag{78}$$

Thus,

$$\mathbb{P}\Big(\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) \leq (1-t)\lambda_{i,\min}^*\Big) \leq r\exp\{-2^{-1}\|\mathbf{V}_r^*\|_{2\to\infty}^{-2}(1-t)^2\lambda_{i,\min}^*\}. \tag{79}$$

Let $t = 1/2$ in the above inequality, we obtain

$$\mathbb{P}\Big(\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) \leq 2^{-1}\lambda_{i,\min}^*\Big) \leq r\exp\{-8^{-1}\|\mathbf{V}_r^*\|_{2\to\infty}^{-2}\lambda_{i,\min}^*\}, \tag{80}$$

which further implies

$$\mathbb{P}\Big(\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) \leq 2^{-1}\lambda_{\min}^*\Big) \leq r\exp\{-8^{-1}\|\mathbf{V}_r^*\|_{2\to\infty}^{-2}\lambda_{\min}^*\}. \tag{81}$$

Apply a union bound to the above inequality for different $i \in [n]$, we obtain

$$\mathbb{P}(\min_{i\in[n]}\sigma_r^2(\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) \leq 2^{-1}\lambda_{\min}^*) \leq nr\exp\{-8^{-1}\|\mathbf{V}_r^*\|_{2\to\infty}^{-2}\lambda_{\min}^*\}. \tag{82}$$

The right-hand side of the above inequality is no greater than $(nr)^{-1}$ when $\lambda_{\min}^* \geq 16\|\mathbf{V}_r^*\|_{2\to\infty}^2\log(nr) = 16\|\mathbf{A}^*\|_{2\to\infty}^2\log(nr)$.

The 'moreover' part of the lemma is proved by noting that $\lambda_{i,\min}^* = \lambda_r(\sum_{j\in[p]}\pi_{ij}\mathbf{a}_j^*(\mathbf{a}_j^*)^T) \geq \pi_{\min}\lambda_r(\sum_j \mathbf{a}_j^*(\mathbf{a}_j^*)^T) = \pi_{\min}\sigma_r^2(\mathbf{A}^*)$. ■

Lemma 28. *If $\|\hat{\mathbf{A}}\|_{2\to\infty}, \|\mathbf{V}_r^*\|_{2\to\infty} \leq C_2$ and $\hat{\mathbf{A}}$ is independent with $\{\omega_{ij}\}_{i\in\mathcal{N}_2, j\in[p]}$, then with probability at least $1-1/(nr)$,*

$$\max_{i\in\mathcal{N}_2}\|diag(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}}-\mathbf{A}^*)\|_2^2 \leq \pi_{\max}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2 + 64\log(n)\cdot\{(\pi_{\max}^{1/2}C_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F)\vee C_2^2\} \tag{83}$$

for $n \geq 4$.

Proof [Proof of Lemma 28] Let $\Delta_{a_j} = \hat{\mathbf{a}}_j - \mathbf{a}_j^*$ and $\Delta_{\mathbf{A}} = \hat{\mathbf{A}} - \mathbf{A}^* = (\Delta_{\mathbf{a}_1}^T, \cdots, \Delta_{\mathbf{a}_p}^T)^T$. Conditional on $\hat{\mathbf{A}}$, $(\omega_{ij} - \pi_{ij})\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T$ are independent symmetric matrices satisfying $\|(\omega_{ij} - \pi_{ij})\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T\|_2 \leq \|\Delta_{\mathbf{a}_j}\|_{2\to\infty}^2 \leq 4C_2^2$, and $\|\mathbb{E}\{(\omega_{ij}\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T)^T\omega_{ij}\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T\}\|_2 \leq \pi_{ij}\|\Delta_{\mathbf{a}_j}\|_{2\to\infty}^2\|\Delta_{\mathbf{a}_j}\|^2 \leq 4\pi_{ij}C_2^2\|\Delta_{\mathbf{a}_j}\|^2$. Applying the inequality (6.1.5) in Tropp (2015) to $\sum_{j\in[p]}(\omega_{ij} - \pi_{ij})\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T$, we obtain that for all $t > 0$

$$P\Big(\|\sum_{j\in[p]}(\omega_{ij} - \pi_{ij})\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T\|_2 \geq t|\hat{\mathbf{A}}\Big) \leq 2r \cdot \exp\Big\{-\frac{3t^2}{8\nu} \wedge \frac{3t}{8L}\Big\} \tag{84}$$

where $\nu = 4\pi_{\max}C_2^2\|\Delta_{\mathbf{A}}\|_F^2 \geq \sum_{j\in[p]}\|\mathbb{E}[\{\omega_{ij}\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T\}^T\omega_{ij}\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T]\|$ and $L = 4C_2^2 \geq \|(\omega_{ij} - \pi_{ij})\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T\|_2$.

For $\epsilon \in (0, 1)$, let $t = [\{8/3 \cdot \log(2r/\epsilon)\}^{1/2}\nu^{1/2}] \vee [\{8/3 \cdot \log(2r/\epsilon)\}L]$ in the above inequality, we obtain

$$P\Big(\|\sum_{j\in[p]}(\omega_{ij} - \pi_{ij})\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T\|_2 \geq t|\hat{\mathbf{A}}\Big) \leq \epsilon. \tag{85}$$

Now we give an upper bound for $t = [\{8/3 \cdot \log(2r/\epsilon)\}^{1/2}\nu^{1/2}] \vee [\{8/3 \cdot \log(2r/\epsilon)\}L]$ for $\epsilon \in (0, r/10)$

$$\begin{aligned}
&[\{8/3 \cdot \log(2r/\epsilon)\}^{1/2}\nu^{1/2}] \vee [\{8/3 \cdot \log(2r/\epsilon)\}L] \\
\leq& 8\log(r/\epsilon) \cdot (\nu^{1/2} \vee L) \\
\leq& 32\log(r/\epsilon) \cdot \{(\pi_{\max}^{1/2}C_2\|\Delta_{\mathbf{A}}\|_F) \vee C_2^2\}.
\end{aligned} \tag{86}$$

Thus, with probability at least $1 - \epsilon$,

$$\|\sum_{j\in[p]}(\omega_{ij} - \pi_{ij})\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T\|_2 \leq 32\log(r/\epsilon) \cdot \{(\pi_{\max}^{1/2}C_2\|\Delta_{\mathbf{A}}\|_F) \vee C_2^2\} \tag{87}$$

for $\epsilon \in (0, r/10)$. Applying a union bound to the above result with $\epsilon = 1/(rn^2)$, we have

$$\|\sum_{j\in[p]}(\omega_{ij} - \pi_{ij})\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T\|_2 \leq 64\log(n) \cdot \{(\pi_{\max}^{1/2}C_2\|\Delta_{\mathbf{A}}\|_F) \vee C_2^2\} \tag{88}$$

with probability at least $1 - 1/(nr)$ for all $i \in \mathcal{N}_2$ and $n \geq 4$.

Next, we give an upper bound for $\lambda_1(\sum_{j=1}^p \pi_{ij}\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T)$.

$$\lambda_1(\sum_{j=1}^p \pi_{ij}\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T) \leq \pi_{\max}\lambda_1(\sum_{j=1}^p \Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T) = \pi_{\max}\|\Delta_{\mathbf{A}}\|_2^2 \leq \pi_{\max}\|\Delta_{\mathbf{A}}\|_F^2. \tag{89}$$

Combining the above two inequalities and note that $\|\mathrm{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}} - \mathbf{A}^*)\|_2^2 = \lambda_1(\sum_{j\in[p]}\omega_{ij}\Delta_{\mathbf{a}_j}\Delta_{\mathbf{a}_j}^T)$, we obtain that with probability at least $1 - 1/(nr)$,

$$\|\mathrm{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}} - \mathbf{A}^*)\|_2^2 \leq \pi_{\max}\|\Delta_{\mathbf{a}_j}\|_F^2 + 64\log(n) \cdot \{(\pi_{\max}^{1/2}C_2\|\Delta_{\mathbf{A}}\|_F) \vee C_2^2\} \tag{90}$$

for $n \geq 4$. ■

A.3.2 NON-ASYMPTOTIC BOUND FOR TERMS IN LEMMA 17

Let $n_{\max} = \max_{j\in[p]} \sum_{i\in[n]} \omega_{ij}$ be the maximal number of observations in each column.

Lemma 29. *If $n\pi_{\max} \geq 6\log(p)$, then $\mathbb{P}(n_{\max} \geq 2n\pi_{\max}) \leq 1/p$.*

Proof [Proof of Lemma 29] The proof is similar to that of Lemma 24. We ommit the details. ■

Lemma 30. *With probability at least $1-1/(np)$, $\|\mathbf{Z}\|_{\max}\leq 8\log(np)\{(\phi\kappa_2^*)^{1/2}\vee 1\}$*

Proof [Proof of Lemma 30] Note that the moment generating function for z_{ij} is $\mathbb{E}(\exp(\lambda z_{ij})) = \exp\{\phi^{-1}(b(m_{ij}^*+\lambda)-b(m_{ij}^*)-\lambda b'(m_{ij^*}))\} = \exp\{2^{-1}\lambda^2\phi b''(\tilde{m}_{ij})\}$ for some $\tilde{m}_{ij}$ between m_{ij}^* and $m_{ij}^*+\lambda$. Thus, z_{ij} is sub-exponential with $\nu^2=\phi\kappa_2^*$ and $\alpha=1$, which implies $\mathbb{P}(|Z_{ij}|\geq t)\leq 2e^{-t^2/(2\phi\kappa_2^*)}\vee e^{-t/2}$. Thus,

$$\mathbb{P}(\|\mathbf{Z}\|_{\max}\geq t)\leq 2(np)(e^{-t^2/(2\phi\kappa_2^*)}\vee e^{-t/2}) \tag{91}$$

Let $t=8\log(np)\{(\phi\kappa_2^*)^{1/2}\vee 1\}$ in the above probability bound. We see that the right-hand side is no larger than $(np)^{-1}$.

■

Lemma 31 (Upper bound for $\|\mathbf{Z}_{\mathcal{N}_2,j}^T\text{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|$)**.** *Assume that $n\pi_{\max}\geq 6\log(p)$. With probability at least $1-3/p-\mathbb{P}(\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|_{2\to\infty}>2C_1)$,*

$$\begin{aligned}&\max_{j\in[p]}\|\mathbf{Z}_{\mathcal{N}_2,j}^T diag(\mathbf{\Omega}_{\mathcal{N}_2,j})\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|\\ \leq&16\{\phi^{1/2}(\kappa_2^*)^{1/2}C_1\log^{1/2}(pr)r^{1/2}(n\pi_{\max})^{1/2}\vee r^{1/2}\phi C_1/(\rho+1)\log(pr)\}\\ &+16\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}_{\mathcal{N}_2}^*\|_{2\to\infty}\cdot n\pi_{\max}\log(np)\{(\kappa_2^*\phi)^{1/2}\vee 1\}\end{aligned} \tag{92}$$

on the event $\{\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|_{2\to\infty}\leq 2C_1\}$.

Proof [Proof of Lemma 31] With similar derivations as that for the inequality (40), we have that with probability at least $1-1/(pr)$,

$$\max_{j\in[p]}\|\mathbf{Z}_{\mathcal{N}_2,j}^T\text{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\mathbf{\Theta}_{\mathcal{N}_2}^*\|\leq 16\{\phi^{1/2}(\kappa_2^*)^{1/2}C_1\log^{1/2}(pr)r^{1/2}n_{\max}^{1/2}\vee r^{1/2}\phi C_1/(\rho+1)\log(pr)\}. \tag{93}$$

Note that

$$\|\mathbf{Z}_{\mathcal{N}_2,j}^T\text{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}_{\mathcal{N}_2}^*)\|=\|\sum_{i\in\mathcal{N}_2}\omega_{ij}z_{ij}(\tilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i^*)\|\leq\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}_{\mathcal{N}_2}^*\|_{2\to\infty}\|\mathbf{Z}\|_{\max}n_{\max}. \tag{94}$$

Thus, with probability at least $1-1/(pr)$,

$$\begin{aligned}&\max_{j\in[p]}\|\mathbf{Z}_{\mathcal{N}_2,j}^T\text{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|\\ \leq&\max_{j\in[p]}\|\mathbf{Z}_{\mathcal{N}_2,j}^T\text{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\mathbf{\Theta}_{\mathcal{N}_2}^*\|+\max_{j\in[p]}\|\mathbf{Z}_{\mathcal{N}_2,j}^T\text{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}_{\mathcal{N}_2}^*)\|\\ \leq&16\{\phi^{1/2}(\kappa_2^*)^{1/2}C_1\log^{1/2}(pr)r^{1/2}(n_{\max})^{1/2}\vee r^{1/2}\phi C_1/(\rho+1)\log(pr)\}\\ &+\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}_{\mathcal{N}_2}^*\|_{2\to\infty}\|\mathbf{Z}\|_{\max}n_{\max}\end{aligned} \tag{95}$$

Combine the above display with Lemma 29 and Lemma 30, we have that with probability at least $1-3/p$,

$$\begin{aligned}&\max_{j\in[p]}\|\mathbf{Z}_{\mathcal{N}_2,j}^T\mathrm{diag}(\boldsymbol{\Omega}_{\mathcal{N}_2,j})\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}\|\\ \leq&16\{\phi^{1/2}(\kappa_2^*)^{1/2}C_1\log^{1/2}(pr)r^{1/2}(n\pi_{\max})^{1/2}\vee r^{1/2}\phi C_1/(\rho+1)\log(pr)\}\\ &+16\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}-\boldsymbol{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}\cdot n\pi_{\max}\log(np)\{(\kappa_2^*\phi)^{1/2}\vee 1\}\end{aligned} \tag{96}$$

■

Lemma 32 (Upper bound for $\|\mathbf{B}_{2,j}(\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2})\|$). *Assume that $n\pi_{\max}\geq 6\log(p)$. With probability at least $1-1/p$,*

$$\max_{j\in[p]}\|\mathbf{B}_{2,j}(\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2})\|\leq 4C_1C_2\kappa_2^*n\pi_{\max}\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}-\boldsymbol{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}, \tag{97}$$

on the event $\{\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}\|_{2\to\infty}\leq 2C_1\}$.

Proof [Proof of Lemma 32]

$$\|\mathbf{B}_{2,j}(\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2})\|=\|\sum_{i\in\mathcal{N}_2}\omega_{ij}b''(m_{ij}^*)\tilde{\boldsymbol{\theta}}_i(\tilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i^*)^T\mathbf{a}_j^*\|\leq 2C_1C_2\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}-\boldsymbol{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}\max_{ij}b''(m_{ij}^*)n_{\max} \tag{98}$$

According to Lemma 29 and noting that $\max_{ij}b''(m_{ij}^*)\leq\kappa_2^*$, we further have that with probability at least $1-1/p$,

$$\max_{j\in[p]}\|\mathbf{B}_{2,j}(\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2})\|\leq 4C_1C_2\kappa_2^*n\pi_{\max}\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}-\boldsymbol{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty} \tag{99}$$

■

Lemma 33. *Assume that $n\pi_{\max}\geq 6\log(p)$. With probability at least $1-1/p$,*

$$\max_{j\in[p]}\beta_{2,j}(\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2})\leq 4C_1C_2^2\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\|_{2\to\infty}^2n\pi_{\max} \tag{100}$$

on the event $\{\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}\|_{2\to\infty}\leq 2C_1\}$.

Proof [Proof of Lemma 33]

$$\beta_{2,j}(\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2})=\sup_{\|\mathbf{u}\|=1}\sum_{i\in\mathcal{N}_2}\omega_{ij}((\tilde{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i^*)^T\mathbf{a}_j^*)^2|\tilde{\boldsymbol{\theta}}_j^T\mathbf{u}|\leq 2C_1C_2^2\|\tilde{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\|_{2\to\infty}^2n_{\max} \tag{101}$$

The proof is completed by combining the above inequality with Lemma 29 ■

Lemma 34. *Assume that $n\pi_{\max}\geq 6\log(p)$. With probability at least $1-1/p$,*

$$\max_{j\in[p]}\gamma_{2,j}(\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2})\leq 16C_1^3n\pi_{\max} \tag{102}$$

on the event $\{\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}\|_{2\to\infty}\leq 2C_1\}$.

Proof [Proof of Lemma 34]

$$\gamma_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}) = \sup_{\|\mathbf{u}\|=1} \sum_j \omega_{ij} |\tilde{\boldsymbol{\theta}}_i^T \mathbf{u}|^3 \leq 8C_1^3 n_{\max} \tag{103}$$

Combine this with Lemma 29, we complete the proof. ■

Lemma 35. *Assume that* $\mathbb{P}(\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty} \leq e_{\mathbf{\Theta},2\to\infty}) \geq 1-\epsilon$ *for some non-random* $e_{\mathbf{\Theta},2\to\infty}$, $n\pi_{\max} \geq 6\log(p)$, $\pi_{\min}\sigma_r^2(\mathbf{\Theta}^*_{\mathcal{N}_2}) \geq 32\|\mathbf{\Theta}^*_{\mathcal{N}_2}\|^2_{2\to\infty}\log(p)$, $p \geq r$, *and* $2e^2_{\mathbf{\Theta},2\to\infty} n\pi_{\max} \leq 2^{-3}\pi_{\min}\sigma_r^2(\mathbf{\Theta}^*_{\mathcal{N}_2})$. *Then, with probability at least* $1 - 2/p - \epsilon$

$$\mathcal{I}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}) \geq 2^{-2}\delta_2(\rho)\pi_{\min}\sigma_r^2(\mathbf{\Theta}^*) \geq 2^{-2}\delta_2(\rho)\pi_{\min}\psi_r^2 \tag{104}$$

Proof [Proof of Lemma 35] First note that

$$\|\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2})\|_2^2 = \|\sum_{i\in\mathcal{N}_2} \omega_{ij}(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*)(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^*)^T\|_2 \leq \|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2}\|^2_{2\to\infty} \cdot n_{\max} \tag{105}$$

Combine the above inequality with Lemma 29, we have that with probability at least $1 - 1/p$,

$$\|\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2})\|_2^2 \leq 2\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2}\|^2_{2\to\infty} \cdot n\pi_{\max}. \tag{106}$$

On the other hand, with similar argument as those in the proof of Lemma 26, we have that if $\pi_{\min}\sigma_r^2(\mathbf{\Theta}^*_{\mathcal{N}_2}) \geq 32\|\mathbf{\Theta}^*_{\mathcal{N}_2}\|^2_{2\to\infty}\log(p)$ and $p \geq r$, then

$$\mathbb{P}\Big(\min_{i\in[n]} \sigma_r^2(\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\mathbf{\Theta}^*_{\mathcal{N}_2}) \leq 2^{-1}\pi_{\min}\sigma_r^2(\mathbf{\Theta}^*_{\mathcal{N}_2})\Big) \leq 1/(pr) \tag{107}$$

Thus, if $2e^2_{\mathbf{\Theta},2\to\infty} n\pi_{\max} \leq 2^{-3}\pi_{\min}\sigma_r^2(\mathbf{\Theta}^*_{\mathcal{N}_2})$, then with probability at least $1 - \epsilon - 2/p$,

$$\|\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2})\|_2 \leq 2^{-1}\sigma_r(\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\mathbf{\Theta}^*_{\mathcal{N}_2}).$$

With similar arguments as those for Lemma 25, we have that with probability $1 - 2/p - \epsilon$,

$$\min_{j\in[p]} \mathcal{I}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}) \geq 2^{-2}\delta_2(\rho)\pi_{\min}\sigma_r^2(\mathbf{\Theta}^*_{\mathcal{N}_2}) \geq 2^{-2}\delta_2(\rho)\pi_{\min}\psi_r^2 \tag{108}$$

where the last inequality in the above display holds because $\mathbf{\Theta}^*_{\mathcal{N}_2} = (\mathbf{U}^*_r)_{\mathcal{N}_2\cdot}\mathbf{D}^*_r$ and as a result $\sigma_r(\mathbf{\Theta}^*) = \sigma_r(\mathbf{M}^*_{\mathcal{N}_2\cdot}) \geq \psi_r$.

■

A.3.3 BOUNDS FOR ψ_1 AND ψ_r

Lemma 36. *Let* $\mathbf{R} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ *be the singular value decomposition of a non-random matrix* $\mathbf{R}$ *with* $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{V} \in \mathbb{R}^{p\times r}$ *and* $\mathbf{D} = diag(\sigma_1(\mathbf{R}, \cdots, \sigma_r(\mathbf{R}))$, *and let* $g_i \sim$ *Bernoulli*$(1/2)$ *be i.i.d. random variables.*

Then,

$$\mathbb{P}(\sigma_r^2(\mathbf{R}_{\mathcal{G}}) \leq 2^{-2}\sigma_r^2(\mathbf{R})) \leq r\exp\left[-2^{-3}\sigma_r^2(\mathbf{R})/\{\|\mathbf{U}\|^2_{2\to\infty}\sigma_1^2(\mathbf{R})\}\right], \tag{109}$$

where $\mathcal{G} = \{i : g_i = 1\}$ *and* $\mathbf{R}_{\mathcal{G}} = (r_{ij})_{i\in\mathcal{G}}$. *In particular, if* $\sigma_r^2(\mathbf{R})/\{\|\mathbf{U}\|^2_{2\to\infty}\sigma_1^2(\mathbf{R})\} \gg \log(r)$, *then with probability converging to* 1, $\sigma_r(\mathbf{R}) \lesssim \sigma_r(\mathbf{R}_{\mathcal{G}}) \leq \sigma_1(\mathbf{R}_{\mathcal{G}}) \leq \sigma_1(\mathbf{R})$.

Proof First, as $\mathbf{R}_{\mathcal{G}}$ is a submatrix of $\mathbf{R}$, we have $\sigma_1(\mathbf{R}_{\mathcal{G}}) \leq \sigma_1(\mathbf{R})$. In the rest of the proof, we show that (109) holds. Let $\mathbf{T} = \mathbf{U}\mathbf{D} \in \mathbb{R}^{n\times r}$. Then, $\mathbf{R}_{\mathcal{G}} = \mathbf{T}_{\mathcal{G}}\mathbf{V}^T$ and $\sigma_r^2(\mathbf{R}_{\mathcal{G}}) = \lambda_r(\mathbf{R}_{\mathcal{G}}\mathbf{R}_{\mathcal{G}}^T) = \lambda_r(\mathbf{T}_{\mathcal{G}}\mathbf{T}_{\mathcal{G}}^T) = \lambda_r(\mathbf{T}_{\mathcal{G}}^T\mathbf{T}_{\mathcal{G}}) = \lambda_r(\sum_{i\in[n]} g_i\mathbf{t}_i\mathbf{t}_i^T)$ where $\mathbf{t}_i = \mathbf{T}_{i\cdot}^T$ indicates the i-th row of the matrix $\mathbf{T}$.

Note that for each i, $g_i\mathbf{t}_i\mathbf{t}_i^T$ is positive semi-definite, and $\lambda_1(g_i\mathbf{t}_i\mathbf{t}_i^T) \leq \|\mathbf{t}_i\|^2 \leq \|\mathbf{T}\|_{2\to\infty}^2$. Also, $\lambda_r(\mathbb{E}(\sum_{i\in[n]} g_i\mathbf{t}_i\mathbf{t}_i^T)) = 2^{-1}\lambda_r(\mathbf{T}^T\mathbf{T}) = 2^{-1}\sigma_r^2(\mathbf{R})$. Applying the weak Chernoff bounds for matrices (inequalities on page 61 of Tropp (2015) under equations (5.1.7) with $t = 1/2$), we obtain

$$\mathbb{P}(\lambda_r(\sum_{i\in[n]} g_i\mathbf{t}_i\mathbf{t}_i^T) \leq 2^{-2}\sigma_r^2(\mathbf{R})) \leq re^{-2^{-3}\sigma_r^2(\mathbf{R})/\|\mathbf{T}\|_{2\to\infty}^2}. \tag{110}$$

We complete the proof by noting that $\|\mathbf{T}\|_{2\to\infty} \leq \|\mathbf{U}\|_{2\to\infty}\sigma_1(\mathbf{R})$. ■

A.4 Asymptotic Analysis

In this section, we provide asymptotic analysis of the estimators based on the non-asymptotic bounds established in previous sections.

Lemma 37 (Asymptotic bounds for ψ_1 and ψ_r). *Recall that* $\psi_1 = \sigma_1(\mathbf{M}^*_{\mathcal{N}_1\cdot}) \vee \sigma_1(\mathbf{M}^*_{\mathcal{N}_2\cdot})$ *and* $\psi_r = \sigma_r(\mathbf{M}^*_{\mathcal{N}_1\cdot}) \wedge \sigma_r(\mathbf{M}^*_{\mathcal{N}_2\cdot})$. *If* $\sigma_r^2(\mathbf{M}^*)/\sigma_1^2(\mathbf{M}^*) \gg \|\mathbf{U}_r^*\|_{2\to\infty}^2 \log(r)$, *then with probability converging to* 1, $\sigma_r(\mathbf{M}^*) \lesssim \psi_r \leq \psi_1 \leq \sigma_1(\mathbf{M}^*)$.

Proof [Proof of Lemma 37] This lemma is a direct application of Lemma 36 with $\mathbf{R}$, $\mathbf{U}$, and $\mathcal{G}$ replaced by $\mathbf{M}^*$, $\mathbf{U}_r^*$ and $\mathcal{N}_1$ (or $\mathcal{N}_2$). We omit the details. ■

Lemma 38 (Asymptotic analysis for $\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}$). *Let* $\mathbf{A}^* = \mathbf{V}_r^*\hat{\mathbf{P}}$, $\mathbf{\Theta}^* = \mathbf{U}_r^*\mathbf{D}_r^*\hat{\mathbf{P}}$, *where* $\hat{\mathbf{P}}$ *is defined in* (11). *Assume that* $\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{A}} - \mathbf{A}^*\|_F \leq e_{\mathbf{A},F}) = 1$. *Assume the following asymptotic regime holds:*

1. $\phi \lesssim 1$;
2. $\|\mathbf{U}_r^*\|_{2\to\infty} \lesssim (r/n)^{1/2}$, $\|\mathbf{V}_r^*\|_{2\to\infty} \lesssim (r/p)^{1/2}$, $C_2 \sim (r/p)^{1/2}$;
3. $(np)^{1/2}r^{\eta_2} \lesssim \sigma_r(\mathbf{M}^*) \leq \sigma_1(\mathbf{M}^*) \lesssim (np)^{1/2}r^{\eta_1}$, *for constants* η_1 *and* η_2;
4. $$\begin{aligned}&p\pi_{\min}\\ \gg&(\delta_2^*)^{-4}(\kappa_2^*)^2(\log(n))^2\\ &\cdot\max\{r^{1\vee(1+2\eta_1)\vee(1-2\eta_2)}(\pi_{\max}/\pi_{\min}), (\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^3 r^{5\vee(3+2\eta_1)\vee(3+4\eta_1)}\};\end{aligned}$$
5. $e_{\mathbf{A},F} \ll (\kappa_2^*)^{-1}(\delta_2^*)^2\min\{r^{-(\eta_1-\eta_2)}(\pi_{\min}/\pi_{\max}), (\kappa_3^*)^{-1}r^{-2-\eta_1}(\pi_{\min}/\pi_{\max})^2\}$;
6. *and* $n \gg r^{1+2(\eta_1-\eta_2)}\log(r)$.

Then, with probability converging to 1*, there is* $\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} = (\tilde{\boldsymbol{\theta}}_i^T)_{i\in\mathcal{N}_2} \in \mathbb{R}^{|\mathcal{N}_2|\times r}$ *such that* $S_{1,i}(\tilde{\boldsymbol{\theta}}_i; \hat{\mathbf{A}}) = \mathbf{0}$ *for all* $i \in \mathcal{N}_2$*, and*

$$\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty} \lesssim \kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min})p^{1/2}\{r(\log(n))^{1/2}(p\pi_{\max})^{-1/2} + r^{1/2+\eta_1}e_{\mathbf{A},F}\}. \tag{111}$$

Moreover, $\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}$ *defined above satisfies* $\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty} \leq C_1$*, and* $\tilde{\boldsymbol{\theta}}_i$ *is the unique solution to the optimization problem* $\max_{\boldsymbol{\theta}_i\in\mathbb{R}^r}\sum_{j\in[p]}\omega_{ij}\{y_{ij}\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j - b(\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j)\}$ *for all* $i \in \mathcal{N}_2$*.*

Proof [Proof of Lemma 38]

First, we provide analysis on the asymptotic regime. Note that $\kappa_2^* \geq \kappa_2(0) \gtrsim 1$ and $\delta_2^* \leq \delta_2(0) \lesssim 1$. Then, the 4-th requirement on the asymptotic regime, i.e.,

$$\begin{aligned} &p\pi_{\min} \\ \gg&(\delta_2^*)^{-4}(\kappa_2^*)^2(\log(n))^2\max\{r^{1\vee(1+2\eta_1)\vee(1-2\eta_2)}(\pi_{\max}/\pi_{\min}), (\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^3 r^{5\vee(3+2\eta_1)\vee(3+4\eta_1)}\}; \end{aligned} \tag{112}$$

implies the following asymptotic regimes,

$$p\pi_{\min} \gg \begin{cases} \max[\log(n), r(\log n)^2, r^{1+2\eta_1}\log(n)], \\ (\kappa_3^*)^2(\kappa_2^*)^{-2}r^{3+2\eta_1}\log(n), \\ (\kappa_3^*)^2(\kappa_2^*)^{-2}r^{3+4\eta_1}\log(n), \\ (\kappa_2^*)^2(\kappa_3^*)^2(\delta_2^*)^{-4}(\pi_{\max}/\pi_{\min})^3 r^5(\log(n)), \\ (\pi_{\max}/\pi_{\min})(\kappa_2^*)^2(\delta_2^*)^{-2}r^{1-2\eta_2}\log(n). \end{cases} \tag{113}$$

Similarly, the 5-th requirement on the asymptotic regime, i.e.,

$$e_{\mathbf{A},F} \ll (\kappa_2^*)^{-1}(\delta_2^*)^2\min\{r^{-(\eta_1-\eta_2)}(\pi_{\min}/\pi_{\max}), (\kappa_3^*)^{-1}r^{-2-\eta_1}(\pi_{\min}/\pi_{\max})^2\} \tag{114}$$

implies

$$e_{\mathbf{A},F} \ll \begin{cases} r^{-1-\eta_1}(\kappa_3^*)^{-1}\kappa_2^*, \\ (\pi_{\min}/\pi_{\max})^{1/2}, \\ (\kappa_2^*)^{-1}\delta_2^* r^{-(\eta_1-\eta_2)}(\pi_{\min}/\pi_{\max}), \\ (\kappa_3^*)^{-1}(\kappa_2^*)^{-1}(\delta_2^*)^2 r^{-2-\eta_1}(\pi_{\min}/\pi_{\max})^2, \end{cases} \tag{115}$$

because $\eta_1 - \eta_2 \geq 0$ and $-1 - 2\eta_1 > -2 - \eta_1$. According to the 6-th asymptotic requirement, $n \gg r^{1+2(\eta_1-\eta_2)}\log(r)$, which implies $\sigma_r^2(\mathbf{M}^*)/\sigma_1^2(\mathbf{M}^*) \gg \|\mathbf{U}_r^*\|_{2\to\infty}^2\log(r)$ and the assumption for Lemma 37 holds. Thus, with probability converging to 1,

$$(np)^{1/2}r^{\eta_2} \lesssim \psi_r \leq \psi_1 \leq (np)^{1/2}r^{\eta_1}. \tag{116}$$

Also, we have

$$r^{1/2+\eta_2}p^{1/2} \lesssim C_1 \lesssim r^{1/2+\eta_1}p^{1/2}, C_2 \lesssim r^{1/2}p^{-1/2}, \text{ and } C_1C_2 \lesssim r^{1+\eta_1}. \tag{117}$$

Throughout the proof, we restrict the analysis on the event $\{\|\hat{\mathbf{A}} - \mathbf{A}^*\|_F \leq e_{\mathbf{A},F}\} \cap \{p_{\max} \leq 2p\pi_{\max}\} \cap \{(np)^{1/2}r^{\eta_2} \lesssim \psi_r \leq \psi_1 \leq (np)^{1/2}r^{\eta_1}\}$, which has probability converging to 1 by the lemma's assumption, (113), (116), and Lemma 24. On this event, we have that with probability at least $1 - 1/n$,

$$\max_{i\in\mathcal{N}_2}\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|\leq 32\{\phi^{1/2}(\kappa_2^*)^{1/2}C_2\log^{1/2}(n)r^{1/2}(p\pi_{\max})^{1/2}\vee r^{1/2}\phi C_2/(\rho+1)\log(n)\}, \tag{118}$$

according to Lemma 18. Under the asymptotic regime that $\phi\lesssim 1$, $C_2\lesssim (r/p)^{1/2}$, the above inequality implies

$$\max_{i\in\mathcal{N}_2}\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|\lesssim(\kappa_2^*)^{1/2}r\log^{1/2}(n)\pi_{\max}^{1/2}\vee rp^{-1/2}\log(n). \tag{119}$$

Note that $\kappa_2^*\gtrsim 1$. According to (113), $p\pi_{\min}\gg r(\log n)^2$, which implies $rp^{-1/2}\log(n)\ll(\kappa_2^*)^{1/2}r\log^{1/2}(n)\pi_{\max}^{1/2}$. Thus, the above display implies

$$\max_{i\in\mathcal{N}_2}\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|\lesssim(\kappa_2^*)^{1/2}r\log^{1/2}(n)\pi_{\max}^{1/2}\lesssim\kappa_2^*r\log^{1/2}(n)\pi_{\max}^{1/2} \tag{120}$$

with probability converging to 1. Next, according to Lemma 19, with probability converging to 1, we have

$$\begin{aligned}&\max_{i\in\mathcal{N}_2}\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|\\ \leq&\kappa_2^*\pi_{\max}C_1\|\hat{\mathbf{A}}\|_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+64\log(n)\cdot(\pi_{\max}^{1/2}\kappa_2^*C_1C_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+\kappa_2^*C_1C_2^2\log(n)).\end{aligned} \tag{121}$$

According to (117), $C_1C_2^2\lesssim r^{3/2+\eta_1}p^{-1/2}$. Also, note that $\|\hat{\mathbf{A}}\|_2\leq 1$. Thus, the above display implies that with probability converging to 1,

$$\max_{i\in\mathcal{N}_2}\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|\lesssim\kappa_2^*\{\pi_{\max}r^{1/2+\eta_1}p^{1/2}e_{\mathbf{A},F}+r^{1+\eta_1}(\pi_{\max})^{1/2}\log(n)e_{\mathbf{A},F}+r^{3/2+\eta_1}p^{-1/2}\log(n)\}. \tag{122}$$

According to (113), $p\pi_{\min}\gg r(\log n)^2$, which implies $\pi_{\max}^{1/2}r^{1+\eta_1}\log(n)\ll\pi_{\max}r^{1/2+\eta_1}p^{1/2}$. Thus, (122) implies that with probability converging to 1,

$$\max_{i\in\mathcal{N}_2}\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|\lesssim\kappa_2^*(\pi_{\max}r^{1/2+\eta_1}p^{1/2}e_{\mathbf{A},F}+r^{3/2+\eta_1}p^{-1/2}\log(n)). \tag{123}$$

According to (113), $p\pi_{\min}\gg r^{1+2\eta_1}\log(n)$, which implies $r^{3/2+\eta_1}p^{-1/2}\log(n)\lesssim r\log^{1/2}(n)\pi_{\max}^{1/2}$. This, together with equations (120) and (123), we have

$$\max_{i\in\mathcal{N}_2}\{\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|+\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|\}\lesssim\kappa_2^*\{r\log^{1/2}(n)\pi_{\max}^{1/2}+\pi_{\max}r^{1/2+\eta_1}p^{1/2}e_{\mathbf{A},F}\} \tag{124}$$

with probability converging to 1.

We proceed to the analysis of $\max_{i\in\mathcal{N}_2}\beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*$. According to Lemma 21, with probability $1-1/n$

$$\max_{i\in\mathcal{N}_2}\beta_{1,i}(\hat{\mathbf{A}})\leq C_1^2C_2\{\pi_{\max}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2+4\pi_{\max}^{1/2}C_2(\log(n))^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F+4C_2^2\log(n)\}. \tag{125}$$

Note that $C_1^2 C_2 \lesssim r^{3/2+2\eta_1} p^{1/2}$. Thus, the above display implies

$$\max_{i \in \mathcal{N}_2} \beta_{1,i}(\hat{\mathbf{A}})\kappa_3^* \leq \kappa_3^* r^{3/2+2\eta_1} p^{1/2} \{\pi_{\max} e_{\mathbf{A},F}^2 + \pi_{\max}^{1/2} r^{1/2} p^{-1/2} (\log(n))^{1/2} e_{\mathbf{A},F} + r p^{-1} \log(n)\}. \tag{126}$$

First, according to (115), $e_{\mathbf{A},F} \lesssim r^{-1-\eta_1}(\kappa_3^*)^{-1}\kappa_2^*$, which implies $\kappa_3^* r^{3/2+2\eta_1} p^{1/2} \pi_{\max} e_{\mathbf{A},F}^2 \lesssim \kappa_2^* \pi_{\max} r^{1/2+\eta_1} p^{1/2} e_{\mathbf{A},F}$. Second, according to (113), $p\pi_{\min} \gg (\kappa_3^*)^2(\kappa_2^*)^{-2} r^{3+2\eta_1}\log(n)$, which implies $\kappa_3^* r^{3/2+2\eta_1} p^{1/2} \cdot \pi_{\max}^{1/2} r^{1/2} p^{-1/2} (\log(n))^{1/2} e_{\mathbf{A},F} \lesssim \kappa_2^* \pi_{\max} r^{1/2+\eta_1} p^{1/2} e_{\mathbf{A},F}$. Third, according to (113), $p\pi_{\min} \gg (\kappa_3^*)^2(\kappa_2^*)^{-2} r^{3+4\eta_1}\log(n)$, which implies $\kappa_3^* r^{3/2+2\eta_1} p^{1/2} \cdot r p^{-1}\log(n) \ll \kappa_2^* r \log^{1/2}(n) \pi_{\max}^{1/2}$. Thus, (126) implies that with probability converging to one,

$$\max_{i \in \mathcal{N}_2} \beta_{1,i}(\hat{\mathbf{A}})\kappa_3^* \lesssim \kappa_2^* \{r \log^{1/2}(n) \pi_{\max}^{1/2} + \pi_{\max} r^{1/2+\eta_1} p^{1/2} e_{\mathbf{A},F}\}. \tag{127}$$

Equations (124) and (127) together imply that with probability converging to 1

$$\max_{i \in \mathcal{N}_2} \{\|\mathbf{Z}_{i\cdot} \mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| + \|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\| + \beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*\} \lesssim \kappa_2^* \{r \log^{1/2}(n) \pi_{\max}^{1/2} + \pi_{\max} r^{1/2+\eta_1} p^{1/2} e_{\mathbf{A},F}\}. \tag{128}$$

Next, we find a lower bound for $\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))$. Note that $\sigma_r(\mathbf{A}^*) = 1$ and $\|\mathbf{A}^*\|_{2\to\infty}^2 \lesssim r/p$ by assumption. Under the asymptotic regime that $p\pi_{\min} \gg r(\log(n))^2$, $\pi_{\min}\sigma_r^2(\mathbf{A}^*) \geq 32\|\mathbf{A}^*\|_{2\to\infty}^2 \log(n)$ for n large enough. According to Lemma 26, with probability at least $1 - 1/(nr)$,

$$\min_{i \in \mathcal{N}_2} \sigma_r^2(\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) \geq 2^{-1}\pi_{\min} \tag{129}$$

for n and p large enough. According to Lemma 28, with probability converging to 1,

$$\max_{i \in \mathcal{N}_2} \|\mathrm{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}} - \mathbf{A}^*)\|_2^2 \lesssim \pi_{\max} e_{\mathbf{A},F}^2 + \pi_{\max}^{1/2} (r/p)^{1/2} \log(n) e_{\mathbf{A},F} + (r/p)\log(n). \tag{130}$$

First, according to (115), $e_{\mathbf{A},F} \ll (\pi_{\min}/\pi_{\max})^{1/2}$, which implies $\pi_{\max} e_{\mathbf{A},F}^2 \ll \pi_{\min}$. Second, according to (113) and (115), $e_{\mathbf{A},F} \ll (\pi_{\min}/\pi_{\max})^{1/2}$ and $\pi_{\min} p \gg r(\log(n))^2$, which implies $e_{\mathbf{A},F} \ll (\pi_{\min}/\pi_{\max})^{1/2}(\pi_{\min} p)^{1/2} r^{-1/2} (\log(n))^{-1}$. This further implies $\pi_{\max}^{1/2}(r/p)^{1/2}\log(n) e_{\mathbf{A},F} \ll \pi_{\min}$. Third, according to (113), $p\pi_{\min} \gg r(\log(n))^2$, which implies $(r/p)\log(n) \ll \pi_{\min}$. Combining the analysis, we have that with probability converging to one,

$$\max_{i \in \mathcal{N}_2} \|\mathrm{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}} - \mathbf{A}^*)\|_2^2 \ll \pi_{\min}. \tag{131}$$

Combining the above display with (129) and using Lemma 25, we have that with probability converging to 1,

$$\min_{i \in \mathcal{N}_2} \sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}})) \geq 2^{-3}\delta_2^* \pi_{\min}. \tag{132}$$

So far, we have obtained upper bounds for $\max_{i \in \mathcal{N}_2} \{\|\mathbf{Z}_{i\cdot} \mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| + \|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\| + \beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*\}$ and a lower bound for $\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))$. In the rest of the proof, we restrict our analysis on the event that (128) and (132) hold. To proceed, we verify conditions of of Lemma 16. According to Lemma 24, on the event $p_{\max} \leq 2p\pi_{\max}$, $\max_{i\in\mathcal{N}_2}\gamma_{1,i}(\hat{\mathbf{A}}) \lesssim p\pi_{\max}(r/p)^{3/2}$. This and (132) implies with probability tending to 1

$$\begin{aligned}&\min_{i\in\mathcal{N}_2}\left\{(\gamma_{1,i}(\hat{\mathbf{A}}))^{-1}(\kappa_3(3C_1C_2))^{-1}\sigma_r^2(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))\right\}\\ \gtrsim&(p\pi_{\max})^{-1}(r/p)^{-3/2}(\kappa_3^*)^{-1}\pi_{\min}^2(\delta_2^*)^2\\ =&(\kappa_3^*)^{-1}(\delta_2^*)^2p^{1/2}r^{-3/2}\pi_{\min}^2/\pi_{\max}.\end{aligned}\tag{133}$$

According to (113), $p\pi_{\min} \gg (\kappa_2^*)^2(\kappa_3^*)^2(\delta_2^*)^{-4}(\pi_{\max}/\pi_{\min})^3r^5(\log(n))$, which implies $\kappa_2^*r\log^{1/2}(n)\pi_{\max}^{1/2} \ll (\kappa_3^*)^{-1}(\delta_2^*)^2p^{1/2}r^{-3/2}\pi_{\min}^2/\pi_{\max}$. According to (115) $e_{\mathbf{A},F} \ll (\kappa_3^*)^{-1}(\kappa_2^*)^{-1}(\delta_2^*)^2r^{-2-\eta_1}(\pi_{\min}/\pi_{\max})^2$, which implies $\kappa_2^*\pi_{\max}r^{1/2+\eta_1}p^{1/2}e_{\mathbf{A},F} \ll (\kappa_3^*)^{-1}(\delta_2^*)^2p^{1/2}r^{-3/2}\pi_{\min}^2/\pi_{\max}$. Combining the analysis, we have $\kappa_2^*r\log^{1/2}(n)\pi_{\max}^{1/2} + \kappa_2^*\pi_{\max}r^{1/2+\eta_1}p^{1/2}e_{\mathbf{A},F} \ll (\kappa_3^*)^{-1}(\delta_2^*)^2p^{1/2}r^{-3/2}\pi_{\min}^2/\pi_{\max}$. This, together with (133) implies

$$\max_{i\in\mathcal{N}_2}\{\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|+\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|+\beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*\} \ll \min_{i\in\mathcal{N}_2}\{(\gamma_{1,i}(\hat{\mathbf{A}}))^{-1}(\kappa_3(3C_1C_2))^{-1}\sigma_r^2(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))\}.\tag{134}$$

Next, according to (132) and $C_1 = \{\|\mathbf{U}_r^*\|_{2\to\infty}\vee(r/n)^{1/2}\}\cdot\sigma_1(\mathbf{M}^*)$

$$\min_{i\in\mathcal{N}_2}\{\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))C_1\} \gtrsim \delta_2^*\pi_{\min}(r/n)^{1/2}(np)^{1/2}r^{\eta_2} \gtrsim \delta_2^*\pi_{\min}r^{1/2+\eta_2}p^{1/2}.\tag{135}$$

According to (113), $p\pi_{\min} \gg (\pi_{\max}/\pi_{\min})(\kappa_2^*)^2(\delta_2^*)^{-2}r^{1-2\eta_2}\log(n)$, which implies $\kappa_2^*r\log^{1/2}(n)\pi_{\max}^{1/2} \ll \delta_2^*\pi_{\min}r^{1/2+\eta_2}p^{1/2}$. According to (115), $e_{\mathbf{A},F} \ll (\kappa_2^*)^{-1}\delta_2^*(\pi_{\min}/\pi_{\max})r^{-(\eta_1-\eta_2)}$, which implies $\kappa_2^*\pi_{\max}r^{1/2+\eta_1}p^{1/2}e_{\mathbf{A},F} \ll \delta_2^*\pi_{\min}r^{1/2+\eta_2}p^{1/2}$. Combining the analysis and (133), we get

$$\max_{i\in\mathcal{N}_2}\{\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|+\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|+\beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*\} \ll \min_{i\in\mathcal{N}_2}\{\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))C_1\}.\tag{136}$$

According to (134) and (136), conditions of Lemma 16 are satisfied. According to Lemma 16 and (128) and (132), with probability converging to 1, there exists $\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} = (\tilde{\boldsymbol{\theta}}_i^T)_{i\in\mathcal{N}_2} \in \mathbb{R}^{|\mathcal{N}_2|\times r}$ such that $S_{1,i}(\tilde{\boldsymbol{\theta}}_i;\hat{\mathbf{A}}) = \mathbf{0}$ for all $i\in\mathcal{N}_2$, and

$$\begin{aligned}&\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}\\ \leq&\max_{i\in\mathcal{N}_2}\left[(\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}})))^{-1}\{\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|+\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|+\beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*\}\right]\\ \lesssim&(\delta_2^*\pi_{\min})^{-1}\kappa_2^*\{r\log^{1/2}(n)\pi_{\max}^{1/2} + \pi_{\max}r^{1/2+\eta_1}p^{1/2}e_{\mathbf{A},F}\}\\ =&\kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min})p^{1/2}\{r(\log(n))^{1/2}(p\pi_{\max})^{-1/2} + r^{1/2+\eta_1}e_{\mathbf{A},F}\},\end{aligned}\tag{137}$$

and $\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}\leq C_1$. Moreover, $\tilde{\boldsymbol{\theta}}_i$ described above is the unique solution to to the optimization problem $\max_{\boldsymbol{\theta}_i\in\mathbb{R}^r}\sum_{j\in[p]}\omega_{ij}\{y_{ij}\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j - b(\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j)\}$ for all $i\in\mathcal{N}_2$ because this optimization is strictly convex by (132). ■

Lemma 39 (Asymptotic analysis for $\tilde{\mathbf{A}}$). *Assume that* $\lim_{n,p\to\infty}\mathbb{P}(\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}\leq e_{\mathbf{\Theta},2\to\infty}) = 1$. *Assume the the following asymptotic regime holds,* *1.* $\phi \lesssim 1$;

2. $\|\mathbf{U}_r^*\|_{2\to\infty} \lesssim (r/n)^{1/2}$, $\|\mathbf{V}_r^*\|_{2\to\infty} \lesssim (r/p)^{1/2}$, $C_2 \sim (r/p)^{1/2}$;

3. $(np)^{1/2} r^{\eta_2} \lesssim \sigma_r(\mathbf{M}^*) \le \sigma_1(\mathbf{M}^*) \lesssim (np)^{1/2} r^{\eta_1}$;

4.

$$\begin{aligned} & n\pi_{\min} \\ \gg & (\kappa_2^*)^2 (\delta_2^*)^{-4} (\log(np))^2 \\ & \cdot \max\left\{(\pi_{\max}/\pi_{\min}) r^{(1+2\eta_1-2\eta_2)\vee(1+2\eta_1-4\eta_2)}, (\kappa_3^*)^2 (\pi_{\max}/\pi_{\min})^3 r^{5+8\eta_1-8\eta_2}\right\}; \end{aligned} \tag{138}$$

5. $e_{\mathbf{\Theta},2\to\infty} \le C_1$ *and*

$$\begin{aligned} & e_{\mathbf{\Theta},2\to\infty} \\ \ll & (\delta_2^*)^2 (\kappa_2^*)^{-1} p^{1/2} (\log(np))^{-1} \\ & \cdot \min\left\{(\pi_{\min}/\pi_{\max}) r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)}, (\kappa_3^*)^{-1} (\pi_{\min}/\pi_{\max})^2 r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}\right\}. \end{aligned} \tag{139}$$

Then, with probability converging to 1, *there is* $\tilde{\mathbf{A}} = (\tilde{\mathbf{a}}_j^T)_{j\in[p]} \in \mathbb{R}^{p\times r}$ *such that* $S_{2,j}(\tilde{\mathbf{a}}_j; \tilde{\mathbf{\Theta}}_{\mathcal{N}_2}) = \mathbf{0}$ *for all* $j \in [p]$, $\|\tilde{\mathbf{A}} - \mathbf{A}^*\| \le C_2$, *and*

$$\begin{aligned} & \|\tilde{\mathbf{A}} - \mathbf{A}^*\|_{2\to\infty} \\ \lesssim & \kappa_2^* (\delta_2^*)^{-1} (\pi_{\max}/\pi_{\min}) r^{-2\eta_2} \log(np) p^{-1/2} \left\{ r^{1+\eta_1} (n\pi_{\max})^{-1/2} + r^{(1+\eta_1)\vee 0} p^{-1/2} e_{\mathbf{\Theta},2\to\infty} \right\}. \end{aligned} \tag{140}$$

Moreover, $\tilde{\mathbf{a}}_j$ *defined above is the unique solution to the optimization problem* $\max_{\mathbf{a}_j\in\mathbb{R}^r} \sum_{i\in\mathcal{N}_2} \omega_{ij}\{y_{ij}\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j - b(\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j)\}$ *for all* $j \in [p]$.

Proof [Proof of Lemma 39] First, the 4-th condition on the asymptotic regime, i.e.,

$$\begin{aligned} & n\pi_{\min} \\ \gg & (\kappa_2^*)^2 (\delta_2^*)^{-4} (\log(np))^2 \max\left\{(\pi_{\max}/\pi_{\min}) r^{(1+2\eta_1-2\eta_2)\vee(1+2\eta_1-4\eta_2)}, (\kappa_3^*)^2 (\pi_{\max}/\pi_{\min})^3 r^{5+8\eta_1-8\eta_2}\right\} \end{aligned} \tag{141}$$

implies the following asymptotic regime holds

$$n\pi_{\min} \gg \begin{cases} \log(p), \\ r^{1+2\eta_1-2\eta_2}\log(p), \\ (\kappa_2^*)^2(\kappa_3^*)^2(\delta_2^*)^{-4}(\pi_{\max}/\pi_{\min})^3 r^{5+8\eta_1-8\eta_2}(\log(np))^2, \\ (\kappa_2^*)^2(\delta_2^*)^{-2}(\pi_{\max}/\pi_{\min}) r^{1+2\eta_1-4\eta_2}\log^2(np), \end{cases} \tag{142}$$

and $n \gg r^{1+2(\eta_1-\eta_2)}\log(r)$, which ensures that the conditions of Lemma 37 holds, and thus, $(np)^{1/2} r^{\eta_2} \lesssim \psi_r \le \psi_2 \lesssim (np)^{1/2} r^{\eta_2}$ with probability converging to 1.

The 5-th condition on the asymptotic regime, i.e.,

$$
\begin{aligned}
&e_{\mathbf{\Theta},2\to\infty}\\
\ll&(\delta_2^*)^2(\kappa_2^*)^{-1}p^{1/2}(\log(np))^{-1}\\
&\cdot\min\{(\pi_{\min}/\pi_{\max})r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)},(\kappa_3^*)^{-1}(\pi_{\min}/\pi_{\max})^2r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}\}
\end{aligned}
\tag{143}
$$

implies

$$
e_{\mathbf{\Theta},2\to\infty}\ll\begin{cases}p^{1/2}r^{1/2+\eta_2}\lesssim C_1,\\ \kappa_2^*(\kappa_3^*)^{-1}r^{-1/2}p^{1/2}\log(np),\\ (\pi_{\min}/\pi_{\max})^{1/2}p^{1/2}r^{\eta_2},\\ (\kappa_2^*)^{-1}(\kappa_3^*)^{-1}(\delta_2^*)^2(\pi_{\min}/\pi_{\max})^2p^{1/2}r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}(\log(np))^{-1},\\ (\kappa_2^*)^{-1}\delta_2^*(\pi_{\min}/\pi_{\max})r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)}(\log(np))^{-1}p^{1/2},\end{cases}
\tag{144}
$$

where we used $\eta_2>-1/2-\eta_1+2\eta_2$ because $\eta_1-\eta_2\geq 0$.

Throughout the proof, we restrict the analysis on the event $\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}\leq e_{\mathbf{\Theta},2\to\infty}\leq C_1$, which has probability converging to 1 as $n,p\to\infty$, according to the assumption of the lemma and (144). This also implies that $\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|\leq 2C_1$ with probability converging to 1. According to Lemma 31 and under the asymptotic regime $n\pi_{\max}\gg\log(p)$, with probability converging to 1,

$$
\begin{aligned}
&\max_{j\in[p]}\|\mathbf{Z}^T_{\mathcal{N}_2,j}\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|\\
\leq&16\{\phi^{1/2}(\kappa_2^*)^{1/2}C_1\log^{1/2}(pr)r^{1/2}(n\pi_{\max})^{1/2}\vee r^{1/2}\phi C_1/(\rho+1)\log(pr)\}\\
&+16\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}\cdot n\pi_{\max}\log(np)\{(\kappa_2^*\phi)^{1/2}\vee 1\}\\
\lesssim&(\kappa_2^*)^{1/2}p^{1/2}r^{1/2+\eta_1}\log^{1/2}(p)r^{1/2}(n\pi_{\max})^{1/2}+r^{1/2}p^{1/2}r^{1/2+\eta_1}\log(p)\}\\
&+e_{\mathbf{\Theta},2\to\infty}n\pi_{\max}\log(np)(\kappa_2^*)^{1/2}\\
\lesssim&(\kappa_2^*)^{1/2}r^{1+\eta_1}p^{1/2}n^{1/2}\pi_{\max}^{1/2}\log^{1/2}(p)+e_{\mathbf{\Theta},2\to\infty}n\pi_{\max}\log(n\vee p)(\kappa_2^*)^{1/2},
\end{aligned}
\tag{145}
$$

where we used $r^{1/2}p^{1/2}r^{1/2+\eta_1}\log(p)\lesssim p^{1/2}r^{1+\eta_1}\log^{1/2}(p)(n\pi_{\max})^{1/2}$ under the asymptotic regime $n\pi_{\max}\gg\log(p)$ for the last inequality.

According to Lemma 32, with probability converging to 1,

$$
\max_{j\in[p]}\|\mathbf{B}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\|\leq 4C_1C_2\kappa_2^*n\pi_{\max}\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}\lesssim\kappa_2^*r^{1+\eta_1}n\pi_{\max}e_{\mathbf{\Theta},2\to\infty}
\tag{146}
$$

According to Lemma 33, with probability converging to 1,

$$
\max_{j\in[p]}\beta_{2,j}(\tilde{\mathbf{\Theta}}^*_{\mathcal{N}_2})\leq 4C_1C_2^2\|\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}^*_{\mathcal{N}_2}\|^2_{2\to\infty}n\pi_{\max}\lesssim r^{3/2+\eta_1}p^{-1/2}e^2_{\mathbf{\Theta},2\to\infty}n\pi_{\max}.
\tag{147}
$$

Combining the above analysis, we obtain that with probability converging to 1,

$$\begin{aligned}
&\max_{j\in[p]}\{\|\mathbf{Z}_{\mathcal{N}_2,j}^T\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|+\|\mathbf{B}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\|+\beta_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\kappa_3^*\}\\
\lesssim&(\kappa_2^*)^{1/2}r^{1+\eta_1}p^{1/2}n^{1/2}\pi_{\max}^{1/2}\log^{1/2}(p)+e_{\mathbf{\Theta},2\to\infty}n\pi_{\max}\log(n\vee p)(\kappa_2^*)^{1/2}\\
&+\kappa_2^*r^{1+\eta_1}n\pi_{\max}e_{\mathbf{\Theta},2\to\infty}+r^{3/2+\eta_1}p^{-1/2}e_{\mathbf{\Theta},2\to\infty}^2n\pi_{\max}\kappa_3^*\\
\lesssim&(\kappa_2^*)^{1/2}r^{1+\eta_1}p^{1/2}n^{1/2}\pi_{\max}^{1/2}\log^{1/2}(p)\\
&+\kappa_2^*r^{(1+\eta_1)\vee 0}\log(np)n\pi_{\max}e_{\mathbf{\Theta},2\to\infty}+r^{3/2+\eta_1}p^{-1/2}e_{\mathbf{\Theta},2\to\infty}^2n\pi_{\max}\kappa_3^*.
\end{aligned}\tag{148}$$

Under the asymptotic regime that $e_{\mathbf{\Theta},2\to\infty}\lesssim\kappa_2^*(\kappa_3^*)^{-1}r^{-1/2}p^{1/2}\log(np)$, $r^{3/2+\eta_1}p^{-1/2}e_{\mathbf{\Theta},2\to\infty}^2n\pi_{\max}\kappa_3^*\lesssim\kappa_2^*r^{1+\eta_1}\log(np)n\pi_{\max}e_{\mathbf{\Theta},2\to\infty}$. Thus, the above inequality implies

$$\begin{aligned}
&\max_{j\in[p]}\{\|\mathbf{Z}_{\mathcal{N}_2,j}^T\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|+\|\mathbf{B}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\|+\beta_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\kappa_3^*\}\\
\lesssim&\kappa_2^*r^{1+\eta_1}p^{1/2}n^{1/2}\pi_{\max}^{1/2}\log(np)+\kappa_2^*r^{(1+\eta_1)\vee 0}\log(np)n\pi_{\max}e_{\mathbf{\Theta},2\to\infty}.
\end{aligned}\tag{149}$$

Next, we derive a lower bound for $\sigma_r(\mathcal{I}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}))$. Under the asymptotic regime $n\pi_{\min}\gg r^{1+2\eta_1-2\eta_2}\log(p)$, and $e_{\mathbf{\Theta},2\to\infty}\ll(\pi_{\min}/\pi_{\max})^{1/2}p^{1/2}r^{\eta_2}$, we have $n\pi_{\max}\gg\log(p)$, $\pi_{\min}(np)r^{2\eta_2}\gg r^{1+2\eta_1}p\log(p)$, and $e_{\mathbf{\Theta},2\to\infty}^2n\pi_{\max}\ll\pi_{\min}(np)r^{2\eta_2}$. Note that $\sigma_r^2(\mathbf{\Theta}_{\mathcal{N}_2}^*)\geq\sigma_r^2(\mathbf{M}_{\mathcal{N}_2,\cdot}^*)\geq\psi_r^2\gtrsim(np)r^{2\eta_2}$ and $\|\mathbf{\Theta}_{\mathcal{N}_2}^*\|_{2\to\infty}\lesssim(r/n)^{1/2}\psi_1\lesssim r^{1/2+\eta_1}p^{1/2}$. Thus, under the same asymptotic regime, conditions of Lemma 35 hold. Therefore, with probability converging to 1,

$$\sigma_r(\mathcal{I}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}))\geq 2^{-2}\delta_2^*\pi_{\min}\psi_r^2\gtrsim\delta_2^*\pi_{\min}(np)r^{2\eta_2}.\tag{150}$$

Note that

$$\begin{aligned}
&\min_j\{2^{-2}(\gamma_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}))^{-1}(\kappa_3^*)^{-1}\sigma_r^2(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\}\\
\gtrsim&(C_1^3n\pi_{\max})^{-1}(\kappa_3^*)^{-1}(\delta_2^*\pi_{\min}\psi_r^2)^2\\
\gtrsim&((p^{1/2}r^{1/2+\eta_1})^3n\pi_{\max})^{-1}(\kappa_3^*)^{-1}(\delta_2^*)^2\pi_{\min}^2(np)^2r^{4\eta_2}\\
=&(\kappa_3^*)^{-1}(\delta_2^*)^2(\pi_{\min}^2/\pi_{\max})p^{1/2}nr^{-3/2-3\eta_1+4\eta_2}.
\end{aligned}\tag{151}$$

Under the asymptotic regime $n\pi_{\min}\gg(\kappa_2^*)^2(\kappa_3^*)^2(\delta_2^*)^{-4}(\pi_{\max}/\pi_{\min})^3r^{5+8\eta_1-8\eta_2}(\log(np))^2$, we have $\kappa_2^*r^{1+\eta_1}\log(np)p^{1/2}n^{1/2}\pi_{\max}^{1/2}\ll(\kappa_3^*)^{-1}(\delta_2^*)^2(\pi_{\min}^2/\pi_{\max})p^{1/2}nr^{-3/2-3\eta_1+4\eta_2}$. Under the asymptotic regime $e_{\mathbf{\Theta},2\to\infty}\ll(\kappa_2^*)^{-1}(\kappa_3^*)^{-1}(\delta_2^*)^2(\pi_{\min}/\pi_{\max})^2p^{1/2}r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}\cdot(\log(np))^{-1}$, we have $\kappa_2^*r^{(1+\eta_1)\vee 0}\log(np)\cdot n\pi_{\max}e_{\mathbf{\Theta},2\to\infty}\ll(\kappa_3^*)^{-1}(\delta_2^*)^2(\pi_{\min}^2/\pi_{\max})p^{1/2}n\cdot r^{-3/2-3\eta_1+4\eta_2}$. Combining the analysis, we have $\kappa_2^*r^{1+\eta_1}p^{1/2}n^{1/2}\pi_{\max}^{1/2}\log^{1/2}(np)+\kappa_2^*r^{(1+\eta_1)\vee 0}\cdot\log(np)n\pi_{\max}e_{\mathbf{\Theta},2\to\infty}\ll(\kappa_3^*)^{-1}(\delta_2^*)^2(\pi_{\min}^2/\pi_{\max})p^{1/2}nr^{-3/2-3\eta_1+4\eta_2}$. This further implies

$$\|\mathbf{Z}_{\mathcal{N}_2,j}^T\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|+\|\mathbf{B}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\|+\beta_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\kappa_3^*\ll 2^{-2}(\gamma_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}))^{-1}(\kappa_3^*)^{-1}\sigma_r^2(\mathcal{I}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}))\tag{152}$$

for all j. According to (150), $\sigma_r(\mathcal{I}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}))C_2\gtrsim\delta_2^*\pi_{\min}(np)r^{2\eta_2}(r/p)^{1/2}\gtrsim\delta_2^*\pi_{\min}np^{1/2}r^{1/2+2\eta_2}$. According to (142), $n\pi_{\min}\gg(\kappa_2^*)^2(\delta_2^*)^{-2}(\pi_{\max}/\pi_{\min})r^{1+2\eta_1-4\eta_2}\log^2(np)$, which implies

$\kappa_2^* r^{1+\eta_1}\log(np)p^{1/2}n^{1/2}\pi_{\max}^{1/2} \ll \delta_2^*\pi_{\min}np^{1/2}r^{1/2+2\eta_2}$. According to (144), $e_{\mathbf{\Theta},2\to\infty} \ll (\kappa_2^*)^{-1}\delta_2^*(\pi_{\min}/\pi_{\max})r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)}(\log(np))^{-1}p^{1/2}$, which implies $\kappa_2^* r^{(1+\eta_1)\vee 0}\log(np)\cdot n\pi_{\max}e_{\mathbf{\Theta},2\to\infty} \ll \delta_2^*\pi_{\min}np^{1/2}r^{1/2+2\eta_2}$. Combine the analysis, we obtain

$$\|\mathbf{Z}_{\mathcal{N}_2,j}^T\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|+\|\mathbf{B}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\|+\beta_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\kappa_3^* \ll \sigma_r(\mathcal{I}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}))C_2 \tag{153}$$

for all j.

The inequalities (152) and (153) verify conditions of Lemma 17 (with C_1 replaced by $2C_1$). According to Lemma 17 and combining (149) and (150), with probability converging to 1,

$$\begin{aligned}
&\|\tilde{\mathbf{A}}-\mathbf{A}^*\|_{2\to\infty}\\
\le&\max_{j\in[p]}\sigma_r^{-1}(\mathcal{I}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}))\{\|\mathbf{Z}_{\mathcal{N}_2,j}^T\mathrm{diag}(\mathbf{\Omega}_{\mathcal{N}_2,j})\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\|+\|\mathbf{B}_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\|+\beta_{2,j}(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2})\kappa_3^*\}\\
\lesssim&\kappa_2^*(\delta_2^*)^{-1}\pi_{\min}^{-1}(np)^{-1}r^{-2\eta_2}\Big\{r^{1+\eta_1}p^{1/2}n^{1/2}\pi_{\max}^{1/2}\log(np)+r^{(1+\eta_1)\vee 0}\log(np)n\pi_{\max}e_{\mathbf{\Theta},2\to\infty}\Big\}\\
\lesssim&\kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min})r^{-2\eta_2}\log(np)p^{-1/2}\Big\{r^{1+\eta_1}(n\pi_{\max})^{-1/2}+r^{(1+\eta_1)\vee 0}p^{-1/2}e_{\mathbf{\Theta},2\to\infty}\Big\}.
\end{aligned} \tag{154}$$

According to (153), $\|\tilde{\mathbf{A}}-\mathbf{A}^*\|_{2\to\infty}\le C_2$. In addition, $\tilde{\mathbf{a}}_j$ is the unique solution to to the optimization problem $\max_{\mathbf{a}_j\in\mathbb{R}^r}\sum_{i\in\mathcal{N}_2}\omega_{ij}\{y_{ij}\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j-b(\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j)\}$ for all j because this optimization is strictly convex by (150).

■

Lemma 40 (Asymptotic analysis for $\tilde{\mathbf{M}}_{\mathcal{N}_2\cdot}=\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\tilde{\mathbf{A}}^T$)**.** *Assume that* $\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{M}}_{\mathcal{N}_1\cdot}-\mathbf{M}^*_{\mathcal{N}_1\cdot}\|_F\le e_{\mathbf{M},F})=1$*, and the following asymptotic regime holds:*

1. $\phi\lesssim 1$;
2. $\|\mathbf{U}_r^*\|_{2\to\infty}\lesssim(r/n)^{1/2}$, $\|\mathbf{V}_r^*\|_{2\to\infty}\lesssim(r/p)^{1/2}$, $C_2\sim(r/p)^{1/2}$;
3. $(np)^{1/2}r^{\eta_2}\lesssim\sigma_r(\mathbf{M}^*)\le\sigma_1(\mathbf{M}^*)\lesssim(np)^{1/2}r^{\eta_1}$ *for some constants* η_1 *and* η_2;
4. $$\begin{aligned}&p\pi_{\min}\\ \gg&(\kappa_2^*)^4(\delta_2^*)^{-6}(\log(np))^3\\ &\cdot\max\Big[(\pi_{\max}/\pi_{\min})^3r^{(1+2\eta_1)\vee(3+2\eta_1-4\eta_2)\vee(1-4\eta_2)},\\ &\qquad(\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^5r^{(3+2\eta_1)\vee(3+4\eta_1)\vee\{7+8(\eta_1-\eta_2)\}\vee(5+6\eta_1-8\eta_2)}\Big];\end{aligned} \tag{155}$$
5. $$\begin{aligned}&n\pi_{\min}\\ \gg&(\kappa_2^*)^2(\delta_2^*)^{-4}(\log(np))^2\max\{(\pi_{\max}/\pi_{\min})r^{(1+2\eta_1-2\eta_2)\vee(1+2\eta_1-4\eta_2)},\\ &\qquad(\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^3r^{5+8\eta_1-8\eta_2}\};\end{aligned} \tag{156}$$

6.

$$\begin{aligned}&(np)^{-1/2}e_{\mathbf{M},F}\\ \ll&(\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-1}(\pi_{\min}/\pi_{\max})^3\min\big[r^{(-\eta_1+\eta_2)\wedge(-1-2\eta_1+3\eta_2)\wedge(-\eta_1+3\eta_2)},\\ &\qquad(\kappa_3^*)^{-1}r^{(-2-\eta_1)\vee\{-3-5(\eta_1-\eta_2)\}\wedge(-2-4\eta_1+5\eta_2)}\big].\end{aligned} \tag{157}$$

Then, with probability converging to 1*,*

$$\begin{aligned}&\|\tilde{\mathbf{M}}_{\mathcal{N}_2\cdot}-\mathbf{M}^*_{\mathcal{N}_2\cdot}\|_{\max}\\ \lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2(\pi_{\max}/\pi_{\min})^2\log^{3/2}(np)\Big[r^{(5/2+2\eta_1-2\eta_2)\vee(3/2+\eta_1-2\eta_2)}\{(p\wedge n)\pi_{\max}\}^{-1/2}\\ &+r^{(2+3\eta_1-3\eta_2)\vee(1+2\eta_1-3\eta_2)}(np)^{-1/2}e_{\mathbf{M},F}\Big].\end{aligned} \tag{158}$$

Proof [Proof of Lemma 40] First, we analyze the asymptotic regime assumption. The 4-th condition of the asymptotic regime, i.e.,

$$\begin{aligned}&p\pi_{\min}\\ \gg&(\kappa_2^*)^4(\delta_2^*)^{-6}(\log(np))^3\\ &\cdot\max\Big[(\pi_{\max}/\pi_{\min})^3r^{(1+2\eta_1)\vee(3+2\eta_1-4\eta_2)\vee(1-4\eta_2)},\\ &\qquad(\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^5r^{(3+2\eta_1)\vee(3+4\eta_1)\vee\{7+8(\eta_1-\eta_2)\}\vee(5+6\eta_1-8\eta_2)}\Big]\end{aligned} \tag{159}$$

implies

$$\begin{aligned}&p\pi_{\min}\\ \gg&\begin{cases}(\delta_2^*)^{-4}(\kappa_2^*)^2(\log(n))^2\max\{r^{1\vee(1+2\eta_1)\vee(1-2\eta_2)}(\pi_{\max}/\pi_{\min}),(\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^3r^{5\vee(3+2\eta_1)\vee(3+4\eta_1)}\},\\ (\kappa_2^*)^4(\delta_2^*)^{-6}(\pi_{\max}/\pi_{\min})^3(\log(np))^3r^{(3+2\eta_1-4\eta_2)\vee(1-4\eta_2)},\\ (\kappa_3^*)^2(\kappa_2^*)^4(\delta_2^*)^{-6}(\pi_{\max}/\pi_{\min})^5r^{\{7+8(\eta_1-\eta_2)\}\vee(5+6\eta_1-8\eta_2)}(\log(np))^3,\end{cases}\end{aligned} \tag{160}$$

where we used the fact $1\le(1+2\eta_1)\vee(1-2\eta_2)$, $3+2\eta_1-4\eta_2>2-2\eta_2$, and $7+8(\eta_1-\eta_2)>5$.

The 6-th condition of the asymptotic regime, i.e.,

$$\begin{aligned}&(np)^{-1/2}e_{\mathbf{M},F}\\ \ll&(\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-1}(\pi_{\min}/\pi_{\max})^3\min\big[r^{(-\eta_1+\eta_2)\wedge(-1-2\eta_1+3\eta_2)\wedge(-\eta_1+3\eta_2)},\\ &\qquad(\kappa_3^*)^{-1}r^{(-2-\eta_1)\vee\{-3-5(\eta_1-\eta_2)\}\wedge(-2-4\eta_1+5\eta_2)}\big]\end{aligned} \tag{161}$$

implies

$$(np)^{-1/2}e_{\mathbf{M},F}\ll\begin{cases}r^{\eta_2},\\ r^{\eta_2}(\kappa_2^*)^{-1}(\delta_2^*)^2\min\{r^{-(\eta_1-\eta_2)}(\pi_{\min}/\pi_{\max}),(\kappa_3^*)^{-1}r^{-2-\eta_1}(\pi_{\min}/\pi_{\max})^2\},\\ (\kappa_2^*)^{-2}(\delta_2^*)^3(\pi_{\min}/\pi_{\max})^2(\log(np))^{-1}r^{(-1-2\eta_1+3\eta_2)\wedge(-\eta_1+3\eta_2)},\\ (\kappa_2^*)^{-2}(\delta_2^*)^3(\pi_{\min}/\pi_{\max})^3(\log(np))^{-1}(\kappa_3^*)^{-1}r^{\{-3-5(\eta_1-\eta_2)\}\wedge(-2-4\eta_1+5\eta_2)},\end{cases} \tag{162}$$

where we used the fact that $\eta_2 \geq -1 - 2\eta_1 + 3\eta_2$ and $\eta_2 - (\eta_1 - \eta_2) \geq -1 - 2\eta_1 + 3\eta_2$.

According to (162), $e_{\mathbf{M},F} \ll (np)^{1/2} r^{\eta_2} \lesssim \psi_r$, which implies that the conditions for Lemma 14 holds. Thus, with probability converging to 1, $\|\hat{\mathbf{A}} - \mathbf{A}^*\|_F \leq e_{\mathbf{A},F}$, where $e_{\mathbf{A},F} = 8\psi_r^{-1} e_{\mathbf{M},F}$. Note that $e_{\mathbf{A},F} \lesssim r^{-\eta_2}(np)^{-1/2} e_{\mathbf{M},F}$. According to (162), $e_{\mathbf{M},F} \ll (np)^{1/2} r^{\eta_2} (\kappa_2^*)^{-1} (\delta_2^*)^2 \min\{r^{-(\eta_1-\eta_2)}(\pi_{\min}/\pi_{\max}), (\kappa_3^*)^{-1} r^{-2-\eta_1}(\pi_{\min}/\pi_{\max})^2\}$, which implies $e_{\mathbf{A},F} \ll (\kappa_2^*)^{-1}(\delta_2^*)^2 \min\{r^{-(\eta_1-\eta_2)}(\pi_{\min}/\pi_{\max}), (\kappa_3^*)^{-1} r^{-2-\eta_1}(\pi_{\min}/\pi_{\max})^2\}$. According to (160), $(\delta_2^*)^{-4}(\kappa_2^*)^2(\log(n))^2 \max\{r^{1\vee(1+2\eta_1)\vee(1-2\eta_2)}(\pi_{\max}/\pi_{\min}), (\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^3 r^{5\vee(3+2\eta_1)\vee(3+4\eta_1)}\} \ll p\pi_{\min}$. Thus, the asymptotic regime of Lemma 38 is satisfied.

According to Lemma 38, $\|\hat{\mathbf{\Theta}}_{\mathcal{N}_2} - \mathbf{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty} \leq e_{\mathbf{\Theta}_{\mathcal{N}_2},2\to\infty}$, with probability converging to 1, for $e_{\mathbf{\Theta}_{\mathcal{N}_2},2\to\infty}$ satisfying

$$\begin{aligned} & e_{\mathbf{\Theta}_{\mathcal{N}_2},2\to\infty} \\ \sim & \kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min}) p^{1/2}\{r(\log(n))^{1/2}(p\pi_{\max})^{-1/2} + r^{1/2+\eta_1} e_{\mathbf{A},F}\} \\ \lesssim & \kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min}) p^{1/2}\{r(\log(n))^{1/2}(p\pi_{\max})^{-1/2} + r^{1/2+\eta_1} \cdot r^{-\eta_2}(np)^{-1/2} e_{\mathbf{M},F}\}. \end{aligned} \tag{163}$$

Next, we verify that the asymptotic regime of Lemma 39 is satisfied. We first verify conditions about $e_{\mathbf{\Theta},2\to\infty}$. According to (160),

$$p\pi_{\min} \gg (\kappa_2^*)^4 (\delta_2^*)^{-6} (\pi_{\max}/\pi_{\min})^3 (\log(np))^3 r^{(3+2\eta_1-4\eta_2)\vee(1-4\eta_2)},$$

which implies

$$\begin{aligned} & \kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min}) p^{1/2} \cdot r(\log(n))^{1/2}(p\pi_{\max})^{-1/2} \\ \ll & (\delta_2^*)^2(\kappa_2^*)^{-1} p^{1/2}(\log(np))^{-1}(\pi_{\min}/\pi_{\max}) r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)}. \end{aligned} \tag{164}$$

According to (160), $p\pi_{\min} \gg (\kappa_3^*)^2(\kappa_2^*)^4(\delta_2^*)^{-6}(\pi_{\max}/\pi_{\min})^5 r^{\{7+8(\eta_1-\eta_2)\}\vee(5+6\eta_1-8\eta_2)}(\log(np))^3$, which implies

$$\begin{aligned} & \kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min}) p^{1/2} r(\log(n))^{1/2}(p\pi_{\max})^{-1/2} \\ \ll & (\delta_2^*)^2(\kappa_2^*)^{-1} p^{1/2}(\log(np))^{-1}(\kappa_3^*)^{-1}(\pi_{\min}/\pi_{\max})^2 r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}. \end{aligned} \tag{165}$$

According to (162), $(np)^{-1/2} e_{\mathbf{M},F} \ll (\kappa_2^*)^{-2}(\delta_2^*)^3(\pi_{\min}/\pi_{\max})^2(\log(np))^{-1} r^{(-1-2\eta_1+3\eta_2)\wedge(-\eta_1+3\eta_2)}$, which implies

$$\begin{aligned} & \kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min}) p^{1/2} \cdot r^{1/2+\eta_1} \cdot r^{-\eta_2}(np)^{-1/2} e_{\mathbf{M},F} \\ \ll & (\delta_2^*)^2(\kappa_2^*)^{-1} p^{1/2}(\log(np))^{-1}(\pi_{\min}/\pi_{\max}) r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)}. \end{aligned} \tag{166}$$

According to (162), $(np)^{-1/2} e_{\mathbf{M},F} \ll (\kappa_2^*)^{-2}(\delta_2^*)^3(\pi_{\min}/\pi_{\max})^3(\log(np))^{-1}(\kappa_3^*)^{-1} r^{\{-3-5(\eta_1-\eta_2)\}\wedge(-2-4\eta_1+5\eta_2)}$, which implies

$$\begin{aligned} & \kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min}) p^{1/2} \cdot r^{1/2+\eta_1} \cdot r^{-\eta_2}(np)^{-1/2} e_{\mathbf{M},F} \\ \ll & (\delta_2^*)^2(\kappa_2^*)^{-1} p^{1/2}(\log(np))^{-1}(\kappa_3^*)^{-1}(\pi_{\min}/\pi_{\max})^2 r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}. \end{aligned} \tag{167}$$

Combining the equations (164)–(167), we have

$$\begin{aligned}
&\kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min})p^{1/2}\{r(\log(n))^{1/2}(p\pi_{\max})^{-1/2}+r^{1/2+\eta_1-\eta_2}(np)^{-1/2}e_{\mathbf{M}_{\mathcal{N}_1,\cdot},F}\}\\
\ll&(\delta_2^*)^2(\kappa_2^*)^{-1}p^{1/2}(\log(np))^{-1}\\
&\cdot\min\{(\pi_{\min}/\pi_{\max})r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)},(\kappa_3^*)^{-1}(\pi_{\min}/\pi_{\max})^2r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}\}
\end{aligned}\tag{168}$$

which implies $e_{\mathbf{\Theta}_{\mathcal{N}_2},2\to\infty}$ satisfies the 5-th condition of the asymptotic regime of Lemma 39.

On the other hand, according to the lemma's assumption,

$$\begin{aligned}
&n\pi_{\min}\\
\gg&(\kappa_2^*)^2(\delta_2^*)^{-4}(\log(np))^2\max\{(\pi_{\max}/\pi_{\min})r^{(1+2\eta_1-2\eta_2)\vee(1+2\eta_1-4\eta_2)},(\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^3r^{5+8\eta_1-8\eta_2}\}.
\end{aligned}\tag{169}$$

Thus, the other requirements for the asymptotic regime in Lemma 39 are also satisfied.

According to Lemma 39, we have $\|\tilde{\mathbf{A}}-\mathbf{A}^*\|_{2\to\infty}\leq e_{\mathbf{A},2\to\infty}$ with probability converging to 1, where

$$e_{\mathbf{A},2\to\infty}\sim\kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min})r^{-2\eta_2}\log(np)\Big\{r^{1+\eta_1}p^{-1/2}(n\pi_{\max})^{-1/2}+r^{(1+\eta_1)\vee 0}p^{-1/2}e_{\mathbf{\Theta},2\to\infty}\Big\}.\tag{170}$$

Combining the above display with (163), we further have

$$\begin{aligned}
&e_{\mathbf{A},2\to\infty}\\
\lesssim&\kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min})r^{-2\eta_2}\log(np)p^{-1/2}\Big[r^{1+\eta_1}(n\pi_{\max})^{-1/2}\\
&+r^{(1+\eta_1)\vee 0}p^{-1/2}\\
&\quad\cdot\kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min})p^{1/2}\{r(\log(n))^{1/2}(p\pi_{\max})^{-1/2}+r^{1/2+\eta_1}\cdot r^{-\eta_2}(np)^{-1/2}e_{\mathbf{M},F}\}\Big]\\
\lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2(\log(np))^{3/2}(\pi_{\max}/\pi_{\min})^2p^{-1/2}\Big[r^{(2+\eta_1-2\eta_2)\vee(1-2\eta_2)}\{(p\wedge n)\pi_{\max}\}^{-1/2}\\
&\qquad+r^{(3/2+2\eta_1-3\eta_2)\vee(1/2+\eta_1-3\eta_2)}(np)^{-1/2}e_{\mathbf{M},F}\Big].
\end{aligned}\tag{171}$$

Now, we combine the above analysis to find an upper bound for $\|\tilde{\mathbf{M}}_{\mathcal{N}_2\cdot}-\mathbf{M}^*_{\mathcal{N}_2}\|_{\max}$. Recall that $\tilde{\mathbf{M}}_{\mathcal{N}_2\cdot}=\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\tilde{\mathbf{A}}^T$. Thus, for $\hat{\mathbf{P}}\in\mathcal{O}_{r\times r}$ defined in (11), and $\mathbf{\Theta}^*_{\mathcal{N}_2}=(\mathbf{U}_r^*)_{\mathcal{N}_2\cdot}\mathbf{D}_r^*\hat{\mathbf{P}}$, $\mathbf{A}^*=\mathbf{V}_r^*\hat{\mathbf{P}}$, we have

$$\begin{aligned}
&\tilde{\mathbf{M}}_{\mathcal{N}_2\cdot}-\mathbf{M}^*_{\mathcal{N}_2\cdot}\\
=&\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\tilde{\mathbf{A}}^T-(\mathbf{U}_r^*)_{\mathcal{N}_2\cdot}\mathbf{D}_r^*(\mathbf{V}_r^*)^T\\
=&\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\tilde{\mathbf{A}}^T-(\mathbf{U}_r^*)_{\mathcal{N}_2\cdot}\mathbf{D}_r^*\hat{\mathbf{P}}(\mathbf{V}_r^*\hat{\mathbf{P}})^T\\
=&\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}\tilde{\mathbf{A}}^T-\mathbf{\Theta}^*_{\mathcal{N}_2}(\mathbf{A}^*)^T\\
=&(\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}-\mathbf{\Theta}^*_{\mathcal{N}_2})(\mathbf{A}^*)^T+\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}(\tilde{\mathbf{A}}-\mathbf{A}^*)^T.
\end{aligned}\tag{172}$$

Therefore, according to Lemma 38, with probability converging to 1,

$$\begin{aligned}&\|\tilde{\mathbf{M}}_{\mathcal{N}_2\cdot}-\mathbf{M}^*_{\mathcal{N}_2\cdot}\|_{\max}\\ \le&\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}-\boldsymbol{\Theta}^*_{\mathcal{N}_2}\|_{2\to\infty}\|\mathbf{A}^*\|_{2\to\infty}+\|\tilde{\mathbf{A}}-\mathbf{A}^*\|_{2\to\infty}\|\tilde{\boldsymbol{\Theta}}_{\mathcal{N}_2}\|_{2\to\infty}\\ \lesssim&(r/p)^{1/2}e_{\boldsymbol{\Theta},2\to\infty}+p^{1/2}r^{1/2+\eta_1}e_{\mathbf{A},2\to\infty}.\end{aligned}\tag{173}$$

Combine the above inequality with (163) and (171), we obtain

$$\begin{aligned}&\|\tilde{\mathbf{M}}_{\mathcal{N}_2\cdot}-\mathbf{M}^*_{\mathcal{N}_2\cdot}\|_{\max}\\ \lesssim&(r/p)^{1/2}\cdot\kappa_2^*(\delta_2^*)^{-1}(\pi_{\max}/\pi_{\min})p^{1/2}\{r(\log(n))^{1/2}(p\pi_{\max})^{-1/2}+r^{1/2+\eta_1-\eta_2}(np)^{-1/2}e_{\mathbf{M},F}\}\\ &+p^{1/2}r^{1/2+\eta_1}(\delta_2^*)^{-2}(\kappa_2^*)^2(\log(np))^{3/2}(\pi_{\max}/\pi_{\min})^2p^{-1/2}\\ &\cdot\Big[r^{(2+\eta_1-2\eta_2)\vee(1-2\eta_2)}\{(p\wedge n)\pi_{\max}\}^{-1/2}+r^{(3/2+2\eta_1-3\eta_2)\vee(1/2+\eta_1-3\eta_2)}(np)^{-1/2}e_{\mathbf{M},F}\Big]\\ \lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2(\pi_{\max}/\pi_{\min})^2\log^{3/2}(np)\Big[r^{(5/2+2\eta_1-2\eta_2)\vee(3/2+\eta_1-2\eta_2)}\{(p\wedge n)\pi_{\max}\}^{-1/2}\\ &\qquad+r^{(2+3\eta_1-3\eta_2)\vee(1+2\eta_1-3\eta_2)}(np)^{-1/2}e_{\mathbf{M},F}\Big].\end{aligned}\tag{174}$$

■

A.5 Additional theoretical results for Algorithm 2 with data splitting

We provide the following theoretical result for $\tilde{\mathbf{M}}$ obtained from Algorithm 2 that extends Theorem 10 to allow $\sigma_r(\mathbf{M}^*)$ and $\sigma_1(\mathbf{M}^*)$ growing at different asymptotic orders and $\pi_{\min}$ and $\pi_{\max}$ decaying at different orders.

Lemma 41 (Asymptotic analysis for $\tilde{\mathbf{M}}$ with data splitting)**.** *Assume that* $\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{M}}_{\mathcal{N}_k\cdot}-\mathbf{M}^*_{\mathcal{N}_k\cdot}\|_F\le e_{\mathbf{M},F})=1$ *($k=1,2$), and the following asymptotic regime holds:*

1. $\phi\lesssim 1$*;*
2. $\|\mathbf{U}_r^*\|_{2\to\infty}\lesssim(r/n)^{1/2}$, $\|\mathbf{V}_r^*\|_{2\to\infty}\lesssim(r/p)^{1/2}$, $C_2\sim(r/p)^{1/2}$*;*
3. $(np)^{1/2}r^{\eta_2}\lesssim\sigma_r(\mathbf{M}^*)\le\sigma_1(\mathbf{M}^*)\lesssim(np)^{1/2}r^{\eta_1}$ *for some constants* η_1 *and* η_2*;*
4. $$\begin{aligned}&p\pi_{\min}\\ \gg&(\kappa_2^*)^4(\delta_2^*)^{-6}(\log(np))^3\\ &\cdot\max\Big[(\pi_{\max}/\pi_{\min})^3r^{(1+2\eta_1)\vee(3+2\eta_1-4\eta_2)\vee(1-4\eta_2)},\\ &\quad(\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^5r^{(3+2\eta_1)\vee(3+4\eta_1)\}\vee\{7+8(\eta_1-\eta_2)\vee(5+6\eta_1-8\eta_2)}\Big];\end{aligned}\tag{175}$$
5. $$\begin{aligned}&n\pi_{\min}\\ \gg&(\kappa_2^*)^2(\delta_2^*)^{-4}(\log(np))^2\max\{(\pi_{\max}/\pi_{\min})r^{(1+2\eta_1-2\eta_2)\vee(1+2\eta_1-4\eta_2)},(\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^3r^{5+8\eta_1-8\eta_2}\};\end{aligned}\tag{176}$$

6.

$$
\begin{aligned}
&(np)^{-1/2}e_{\mathbf{M},F}\\
\ll&(\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-1}(\pi_{\min}/\pi_{\max})^3\min\big[r^{(-\eta_1+\eta_2)\wedge(-1-2\eta_1+3\eta_2)\wedge(-\eta_1+3\eta_2)},\\
&\qquad(\kappa_3^*)^{-1}r^{(-2-\eta_1)\vee\{-3-5(\eta_1-\eta_2)\}\wedge(-2-4\eta_1+5\eta_2)}\big].
\end{aligned}
\tag{177}
$$

Then, with probability converging to 1*, estimating equations in steps 3 and 4 of Algorithm 2 have a unique solution and*

$$
\begin{aligned}
&\|\tilde{\mathbf{M}}-\mathbf{M}^*\|_{\max}\\
\lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2(\pi_{\max}/\pi_{\min})^2\log^{3/2}(np)\Big[r^{(5/2+2\eta_1-2\eta_2)\vee(3/2+\eta_1-2\eta_2)}\{(p\wedge n)\pi_{\max}\}^{-1/2}\\
&\qquad+r^{(2+3\eta_1-3\eta_2)\vee(1+2\eta_1-3\eta_2)}(np)^{-1/2}e_{\mathbf{M},F}\Big].
\end{aligned}
\tag{178}
$$

Proof [Proof of Lemma 41] Recall that $\tilde{\mathbf{M}}=(\tilde{m}_{ij})_{i\in[n],j\in[p]}$, where $(\tilde{m}_{ij})_{i\in\mathcal{N}_1,j\in[p]}=\tilde{\mathbf{\Theta}}^{(2)}_{\mathcal{N}_1}(\tilde{\mathbf{A}}^{(2)})^T$ and $(\tilde{m}_{ij})_{i\in\mathcal{N}_2,j\in[p]}=\tilde{\mathbf{\Theta}}^{(1)}_{\mathcal{N}_2}(\tilde{\mathbf{A}}^{(1)})^T$. The error rate for $(\tilde{m}_{ij})_{i\in\mathcal{N}_2,j\in[p]}=\tilde{\mathbf{\Theta}}^{(1)}_{\mathcal{N}_2}(\tilde{\mathbf{A}}^{(1)})^T$ is obtained by Lemma 40, and the error rate of $(\tilde{m}_{ij})_{i\in\mathcal{N}_1,j\in[p]}$ is obtained by swapping $(\hat{\mathbf{A}}^{(1)},\tilde{\mathbf{\Theta}}^{(1)}_{\mathcal{N}_2},\tilde{\mathbf{A}}^{(1)},\mathcal{N}_1)$ with $(\hat{\mathbf{A}}^{(2)},\tilde{\mathbf{\Theta}}^{(2)}_{\mathcal{N}_1},\tilde{\mathbf{A}}^{(2)},\mathcal{N}_2)$ in the proof of Lemma 40.

The uniqueness of the solution to estimating equations in steps 3 and 4 of Algorithm 2 is proved by the uniqueness property in Lemma 38 and 39.

■

A.6 Proof of Theorem 10

Proof [Proof of Theorem 10] Note that when $\pi_{\min}\sim\pi_{\max}\sim\pi$ and $\eta_1=\eta_2=\eta$, the 4-th asymptotic requirement in Lemma 41 becomes

$$
p\pi\gg(\kappa_2^*)^4(\delta_2^*)^{-6}(\log(np))^3\cdot\max\Big[r^{(1+2\eta)\vee(3-2\eta)\vee(1-4\eta)},(\kappa_3^*)^2r^{(3+2\eta)\vee(3+4\eta)\vee\{7\vee(5-2\eta)\}}\Big].
\tag{179}
$$

When $\eta\geq-1$, the above requirement is implied by

$$
p\pi\gg(\kappa_2^*)^4(\delta_2^*)^{-6}(\log(np))^3\cdot\max\Big[r^{(1+2\eta)\vee5},(\kappa_3^*)^2r^{(3+4\eta)\vee7}\Big],
\tag{180}
$$

which is the asymptotic requirement *R5*.

Similarly, the 5-th asymptotic requirement in Lemma 41 becomes

$$
n\pi\gg(\kappa_2^*)^2(\delta_2^*)^{-4}(\log(np))^2\max\{r^{1\vee(1-2\eta)},(\kappa_3^*)^2r^5\},
$$

which is implied by the asymptotic requirement *R6*: $n\pi\gg(\kappa_2^*)^2(\delta_2^*)^{-4}(\log(np))^2\max\{r^3,(\kappa_3^*)^2r^5\}$.

The 6-th asymptotic requirement becomes

$$(np)^{-1/2}e_{\mathbf{M},F} \ll (\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-1}\min\left[r^{0\wedge(-1+\eta)\wedge(2\eta)}, (\kappa_3^*)^{-1}r^{(-2-\eta)\wedge(-3)\wedge(-2+\eta)}\right], \tag{181}$$

and is implied by $(np)^{-1/2}e_{\mathbf{M},F} \ll (\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-1}\min\left[r^{-2}, (\kappa_3^*)^{-1}r^{-3}\right]$, and further implied by the asymptotic requirement *R7'*.

Thus, under *R1-R6* and *R7'*, the conditions of Lemma 41 is satisfied and with probability converging to 1,

$$\begin{aligned}
&\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \\
\lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2\log^{3/2}(np)\Big[r^{(5/2+2\eta_1-2\eta_2)\vee(3/2+\eta_1-2\eta_2)}\{(p\wedge n)\pi\}^{-1/2} \\
&\qquad + r^{(2+3\eta_1-3\eta_2)\vee(1+2\eta_1-3\eta_2)}(np)^{-1/2}e_{\mathbf{M},F}\Big] \\
\lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2\log^{3/2}(np)\Big[r^{5/2\vee(3/2-\eta)}\{(p\wedge n)\pi\}^{-1/2} + r^{2\vee(1-\eta)}(np)^{-1/2}e_{\mathbf{M},F}\Big] \\
\lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2\log^{3/2}(np)\Big[r^{5/2}\{(p\wedge n)\pi\}^{-1/2} + r^{2}(np)^{-1/2}e_{\mathbf{M},F}\Big] \\
\lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2\log^{2}(np)r^{5/2}\Big[\{(p\wedge n)\pi\}^{-1/2} + (np)^{-1/2}e_{\mathbf{M},F}\Big].
\end{aligned} \tag{182}$$

The above analysis gives the error bound of $\tilde{\mathbf{M}}$.

To proceed to prove the 'in particular' part of the theorem. We note that if $r \lesssim 1$, then $\sigma_1(\mathbf{M}^*) \sim \sigma_r(\mathbf{M}^*) \sim (np)^{1/2}$ and $C_1 \sim n^{-1/2}\sigma_1(\mathbf{M}^*) \lesssim p^{1/2}$ and $C_2 \sim p^{-1/2}$. As a result, $\|\mathbf{M}^*\|_{\max} \leq C_1C_2 \lesssim 1$ and thus $2\rho + 1 \lesssim 1$. This implies that $\delta_2^* \gtrsim 1$, $\kappa_2^*, \kappa_3^* \lesssim 1$. The proof is completed by combining the above analysis with (182).

■

Appendix B. Proof of Theorem 5 and Additional Theoretical Results for Algorithm 1 without Data Splitting

In this section, we provide analysis for $\tilde{\mathbf{\Theta}}$, $\tilde{\mathbf{A}}$, and $\tilde{\mathbf{M}}$ obtained from Algorithm 1 without data splitting. Let

$$\hat{\mathbf{P}} = \arg\min_{\mathbf{P}\in\mathcal{O}_r}\|\hat{\mathbf{V}}_r - \mathbf{V}_r^*\mathbf{P}\|_F \tag{183}$$

and $\mathbf{A}^* = \mathbf{V}_r^*\hat{\mathbf{P}}$ and $\mathbf{\Theta}^* = \mathbf{U}_r^*\mathbf{D}_r^*\hat{\mathbf{P}}$. With similar derivations as those for Lemma 14, we have the following lemma.

Lemma 42. *If* $\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{M}} - \mathbf{M}^*\|_F \geq e_{\mathbf{M},F}) = 0$, $e_{\mathbf{M},F}$ *is a non-random number (depending on* n *and* p*),* $\|\mathbf{V}_r^*\|_{2\to\infty} \leq C_2$, $e_{\mathbf{M},F} \leq 2^{-1}\sigma_r(\mathbf{M}^*)$*, and* $\hat{\mathbf{P}}$ *is defined in* (183) *then*

$$\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{A}} - \mathbf{V}_r^*\hat{\mathbf{P}}\|_F \geq e_{\mathbf{A},F}) = 0, \tag{184}$$

where $e_{\mathbf{A},F} = 8\sigma_r^{-1}(\mathbf{M}^*)e_{\mathbf{M},F}$.

The rest of the section is organized as follows. In Section B.1, we obtain non-asymptotic probabilistic bounds for terms involved in the estimating equations in Step 3 and 4 of Algorithm 1. In Section B.2, we obtain asymptotic error bounds for $\|\tilde{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{2\to\infty}$ (Lemma 48). In Section B.3, we provide error bound $\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max}$ (Lemma 49) under a general setting. Finally, the proof of Theorem 5 is given in Section B.4.

B.1 Non-asymptotic Analysis

We first analyze each term in Lemma 16 with $\mathbf{A} = \hat{\mathbf{A}}$ obtained from Algorithm 1 without data splitting.

Lemma 43 (Upper bound for $\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|$ without data splitting)**.** *Assume $n \geq 2$. $\|\mathbf{M}^*\|_{\max}\leq \rho$. Assume that $\|\mathbf{A}^*\|_{2\to\infty}\leq C_2$ and $\hat{\mathbf{A}}$ may be dependent with $\mathbf{\Omega}$, $n \geq r$. Then, with probability at least $1-2(nr)^{-1}$,*

$$\begin{aligned}&\max_{i\in[n]}\|\mathbf{Z}_{i\cdot}\textit{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|\\ \leq&8\{\phi^{1/2}(\kappa_2(2\rho+1))^{1/2}C_2\log^{1/2}(nr)r^{1/2}p_{\max}^{1/2}\vee r^{1/2}\phi C_2/(\rho+1)\log(nr)\}\\ &+8\log(np)\{(\phi\kappa_2^*)^{1/2}\vee 1\}\cdot p_{\max}^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F.\end{aligned} \tag{185}$$

Proof [Proof of Lemma 43] Note that

$$\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|\leq \|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*\|+\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}}-\mathbf{A}^*)\| \tag{186}$$

and

$$\begin{aligned}&\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}}-\mathbf{A}^*)\|\\ =&\|\sum_{j=1}^p z_{ij}\omega_{ij}(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)\|\\ \leq&\|\mathbf{Z}\|_{\max}\sum_{j=1}^p\omega_{ij}\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|\\ \leq&\|\mathbf{Z}\|_{\max}p_{\max}^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F.\end{aligned} \tag{187}$$

Combining the above two inequalities and taking maximum over $i \in [n]$, we have

$$\max_{i\in[n]}\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|\leq \max_{i\in[n]}\{\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*\|\}+\|\mathbf{Z}\|_{\max}p_{\max}^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F. \tag{188}$$

For the first term on the right-hand side of the above inequality, we follow a similar proof as that in the proof of Lemma 18 (with $\hat{\mathbf{A}}$ replaced by $\mathbf{A}^*$) and obtain that with probability at least $1-(nr)^{-1}$

$$\max_{i\in[n]}\{\|\mathbf{Z}_{i\cdot}\text{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*\|\}\leq 8\{\phi^{1/2}(\kappa_2(2\rho{+}1))^{1/2}C_2\log^{1/2}(nr)r^{1/2}p_{\max}^{1/2}\vee r^{1/2}\phi C_2/(\rho{+}1)\log(nr)\}. \tag{189}$$

For the second term on the right-hand side of equation (188), we apply Lemma 30 and obtain that with probability at least $1-(np)^{-1}$,

$$\|\mathbf{Z}\|_{\max}p_{\max}^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F\leq 8\log(np)\{(\phi\kappa_2^*)^{1/2}\vee 1\}\cdot p_{\max}^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F. \tag{190}$$

The proof is completed by combining the above two inequalities. ■

Lemma 44 (Upper bound for $\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|$ without data splitting)**.** *Let* $\mathbf{A}^* = \mathbf{V}_r^*\hat{\mathbf{P}}$ *and* $\mathbf{\Theta}^* = \mathbf{U}_r^*\mathbf{D}_r^*\hat{\mathbf{P}}$. *Assume* $\|\hat{\mathbf{A}}\|_{2\to\infty}, \|\mathbf{V}_r^*\|_{2\to\infty}\leq C_2$ *and* $\|\mathbf{U}_r^*\mathbf{D}_r^*\|_{2\to\infty}\leq C_1$, *and* $\hat{\mathbf{A}}$ *may be dependent with* $\mathbf{\Omega}_i$. *Then,*

$$\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|\leq C_1C_2\kappa_2^*p_{\max}^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F. \tag{191}$$

Proof [Proof of Lemma 44] First, by the assumptions and $\hat{\mathbf{P}}$ is orthogonal, $\|\mathbf{\Theta}^*\|_{2\to\infty}= \|\mathbf{U}_r^*\mathbf{D}_r^*\|_{2\to\infty}\leq C_1$ and $\|\mathbf{A}^*\|_{2\to\infty}= \|\mathbf{V}_r^*\|_{2\to\infty}\leq C_2$. Recall that

$$\begin{aligned}\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|=&\|\sum_{j=1}^{p}\omega_{ij}b''(m_{ij}^*)\hat{\mathbf{a}}_j(\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*\|\\ \leq&C_1C_2\sum_{j=1}^{p}\omega_{ij}b''(m_{ij}^*)\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\| \\ \leq&C_1C_2\kappa_2^*\sum_{j=1}^{p}\omega_{ij}\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|.\end{aligned} \tag{192}$$

Applying Cauchy-Schwarz inequality, we further obtain

$$\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|\leq C_1C_2\kappa_2^*p_{\max}^{1/2}\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F. \tag{193}$$

The proof is completed by taking maximum for $i\in[n]$. ■

Lemma 45 (Bound for $\beta_{1,i}(\hat{\mathbf{A}})$, without data splitting)**.** *If* $\|\mathbf{U}_r^*\mathbf{D}_r^*\|_{2\to\infty}\leq C_1$, $\|\hat{\mathbf{A}}\|_{2\to\infty}, \|\mathbf{V}_r^*\|_{2\to\infty}\leq C_2$, *then,*

$$\max_{i\in[n]}\beta_{1,i}(\hat{\mathbf{A}})\leq C_1^2C_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2. \tag{194}$$

Proof [Proof of Lemma 45] Recall

$$\beta_{1,i}(\hat{\mathbf{A}})=\sup_{\|\mathbf{u}\|=1}\sum_j\omega_{ij}((\hat{\mathbf{a}}_j-\mathbf{a}_j^*)^T\boldsymbol{\theta}_i^*)^2|\hat{\mathbf{a}}_j^T\mathbf{u}|\leq C_1^2C_2\sum_{j\in[p]}\omega_{ij}\|\hat{\mathbf{a}}_j-\mathbf{a}_j^*\|^2\leq C_1^2C_2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2. \tag{195}$$

■

Lemma 46 (Bound for $\gamma_{1,i}(\hat{\mathbf{A}})$, without data splitting)**.** *If* $\|\hat{\mathbf{A}}\|_{2\to\infty}\leq C_2$, *then with probability at least* $1-1/n$,

$$\max_{i\in[n]}\gamma_{1,i}(\hat{\mathbf{A}})\leq 2p\pi_{\max}C_2^3. \tag{196}$$

Proof [Proof of Lemma 46] The proof of this Lemma is the same as that of Lemma 24 which does not require the independence between $\hat{\mathbf{A}}$ and $\mathbf{\Omega}_{i\cdot}$. ■

Lemma 47.

$$\max_{i\in[n]}\|\mathit{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}}-\mathbf{A}^*)\|_2^2\leq\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2. \tag{197}$$

Proof [Proof of Lemma 47]

$$\max_{i\in[n]}\|\mathrm{diag}(\boldsymbol{\Omega}_{i\cdot})(\hat{\mathbf{A}}-\mathbf{A}^*)\|_2^2\leq \max_{i\in[n]}\|\mathrm{diag}(\boldsymbol{\Omega}_{i\cdot})\|_2^2\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2= \|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2. \tag{198}$$

■

B.2 Asymptotic Analysis for Algorithm 1 without Data Splitting

Lemma 48 (Asymptotic analysis of $\tilde{\mathbf{A}}$ without data splitting). *Let* $\mathbf{A}^* = \mathbf{V}_r^*\hat{\mathbf{P}}$, $\boldsymbol{\Theta}^* = \mathbf{U}_r^*\mathbf{D}_r^*\hat{\mathbf{P}}$, *and* $\hat{\mathbf{P}}$ *is defined in* (183). *Assume that* $\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{A}}-\mathbf{A}^*\|_F\geq e_{\mathbf{A},F})=1$.

Assume the following asymptotic regime holds:

1. $\phi\sim 1$, $\pi_{\min}\sim\pi_{\max}\sim\pi$;
2. $\|\mathbf{U}_r^*\|_{2\to\infty}\lesssim (r/n)^{1/2}$, $\|\mathbf{V}_r^*\|_{2\to\infty}\lesssim (r/p)^{1/2}$, $C_2\sim (r/p)^{1/2}$;
3. $(np)^{1/2}r^{\eta_2}\lesssim \sigma_r(\mathbf{M}^*)\leq\sigma_1(\mathbf{M}^*)\lesssim (np)^{1/2}r^{\eta_1}$, *and* η_1 *and* η_2 *are constants;*
4. $p\pi\gg(\delta_2^*)^{-4}(\kappa_2^*)^2\log^2(n)\max\{r^{1\vee(1-2\eta_2)},(\kappa_3^*)^2r^5\}$;
5. $e_{\mathbf{A},F}\ll(\kappa_2^*)^{-1}(\delta_2^*)^2(\log(np))^{-1}\min\{r^{0\wedge(-1/2-\eta_1+\eta_2)\wedge(1/2+\eta_2)},(\kappa_3^*)^{-1}r^{(-5/2-\eta_1)\wedge(-3/2)}\}\pi^{1/2}$.

Then, with probability converging to 1, *there is* $\tilde{\boldsymbol{\Theta}}=(\tilde{\boldsymbol{\theta}}_i^T)_{i\in[n]}$ *such that* $S_{1,i}(\tilde{\boldsymbol{\theta}}_i,\hat{\mathbf{A}})=\mathbf{0}$, *for all* $i\in[n]$, $\|\tilde{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\|_{2\to\infty}\leq C_1$, *and*

$$\|\tilde{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\|_{2\to\infty}\lesssim\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}\{r(\log(n))^{1/2}+\log(np)r^{(1+\eta_1)\vee 0}p^{1/2}e_{\mathbf{A},F}\}. \tag{199}$$

Moreover, $\tilde{\boldsymbol{\theta}}_i$ *is the unique solution to the optimization problem* $\max_{\boldsymbol{\theta}_i\in\mathbb{R}^r}\sum_{j\in[p]}\omega_{ij}\{y_{ij}\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j-b(\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j)\}$ *for all* $i\in[n]$.

Proof [Proof of Lemma 48] First, we provide analysis on the asymptotic regime. Note that $\kappa_2^*\geq\kappa_2(0)\gtrsim 1$ and $\delta_2^*\leq\delta_2(0)\lesssim 1$. Then, the 4-th condition on the asymptotic regime, i.e.,

$$p\pi\gg(\delta_2^*)^{-4}(\kappa_2^*)^2\log^2(n)\max\{r^{1\vee(1-2\eta_2)},(\kappa_3^*)^2r^5\} \tag{200}$$

implies the following asymptotic regimes,

$$p\pi\gg\begin{cases}r(\log n)^2,\\(\kappa_2^*)^2(\kappa_3^*)^2(\delta_2^*)^{-4}r^5\log(n),\\(\kappa_2^*)^2(\delta_2^*)^{-2}\log(n)r^{1-2\eta_2}.\end{cases} \tag{201}$$

Similarly, the 5-th condition on the asymptotic regime, i.e.,

$$e_{\mathbf{A},F}\ll(\kappa_2^*)^{-1}(\delta_2^*)^2(\log(np))^{-1}\min\{r^{0\wedge(-1/2-\eta_1+\eta_2)\wedge(1/2+\eta_2)},(\kappa_3^*)^{-1}r^{(-5/2-\eta_1)\wedge(-3/2)}\}\pi^{1/2} \tag{202}$$

implies

$$e_{\mathbf{A},F} \ll \begin{cases} (\kappa_3^*)^{-1} r^{-1/2-\eta_1} \pi^{1/2}, \\ \pi^{1/2}, \\ (\kappa_2^*)^{-1}(\kappa_3^*)^{-1}(\delta_2^*)^2(\log(np))^{-1} r^{(-5/2-\eta_1)\wedge(-3/2)} \pi^{1/2}, \\ \delta_2^*(\kappa_2^*)^{-1}(\log(np))^{-1} r^{(-1/2-\eta_1+\eta_2)\wedge(1/2+\eta_2)} \pi^{1/2}, \end{cases} \tag{203}$$

where we used the fact that $-1/2 - \eta_1 > -5/2 - \eta_1$.

Throughout the proof, we restrict the analysis on the event $\{\|\hat{\mathbf{A}} - \mathbf{A}^*\|_F \leq e_{\mathbf{A},F}\} \cap \{p_{\max} \leq 2p\pi_{\max}\}$, which has probability converging to 1 by the lemma's assumption, and Lemma 23. On this event, we have that with probability at least $1 - 1/n$,

$$\begin{aligned} &\max_{i\in[n]} \|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| \\ \leq &16\{\phi^{1/2}(\kappa_2(2\rho+1))^{1/2} C_2 \log^{1/2}(nr) r^{1/2}(p\pi_{\max})^{1/2} \vee r^{1/2}\phi C_2/(\rho+1)\log(nr)\} \\ &+ 8\{(\phi\kappa_2^*)^{1/2} \vee 1\} \log(np)\cdot (p\pi_{\max})^{1/2} e_{A,F}. \end{aligned} \tag{204}$$

according to Lemma 43. Under the asymptotic regime that $\phi \lesssim 1$, $\pi_{\min} \sim \pi_{\max} \sim \pi$, $C_2 \lesssim (r/p)^{1/2}$, the above inequality implies

$$\max_{i\in\mathcal{N}_2} \|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| \lesssim (\kappa_2^*)^{1/2} r \log^{1/2}(n)\pi^{1/2} + rp^{-1/2}\log(n) + (\kappa_2^*)^{1/2}\log(np) p^{1/2}\pi^{1/2} e_{A,F}. \tag{205}$$

According to (201), $p\pi \gg r(\log n)^2$, which implies $rp^{-1/2}\log(n) \ll (\kappa_2^*)^{1/2} r \log^{1/2}(n)\pi^{1/2}$. Thus, the above display implies

$$\max_{i\in\mathcal{N}_2} \|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| \lesssim (\kappa_2^*)^{1/2} r \log^{1/2}(n)\pi^{1/2} + (\kappa_2^*)^{1/2}\log(np) p^{1/2}\pi^{1/2} e_{A,F} \tag{206}$$

with probability converging to 1.

Next, according to Lemma 44,

$$\max_{i\in[n]} \|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\| \leq C_1 C_2 \kappa_2^* p_{\max}^{1/2} \|\hat{\mathbf{A}} - \mathbf{A}^*\|_F. \tag{207}$$

Note that $C_1 C_2 \lesssim r^{1+\eta_1}$. Thus, the above display implies that with probability converging to one,

$$\max_{i\in[n]} \|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\| \lesssim \kappa_2^* r^{1+\eta_1} p^{1/2}\pi^{1/2} e_{A,F}. \tag{208}$$

Combining equations (206) and (208), we obtain

$$\max_{i\in[n]} \{\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\| + \|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|\} \lesssim \kappa_2^*\{r\log^{1/2}(n)\pi^{1/2} + \log(np) r^{(1+\eta_1)\vee 0} p^{1/2}\pi^{1/2} e_{A,F}\}. \tag{209}$$

Next, we consider $\max_{i\in[n]}\{\beta_{1,i}(\hat{\mathbf{A}})\}\kappa_3^*$. According to Lemma 45, we have

$$\max_{i\in[n]} \beta_{1,i}(\hat{\mathbf{A}}) \leq C_1^2 C_2 \|\hat{\mathbf{A}} - \mathbf{A}^*\|_F^2. \tag{210}$$

Note that $C_1^2 C_2 \lesssim r^{3/2+2\eta_1} p^{1/2}$. Thus, the above display implies

$$\max_{i\in[n]}\{\beta_{1,i}(\hat{\mathbf{A}})\}\kappa_3^* \lesssim \kappa_3^* r^{3/2+2\eta_1} p^{1/2} e_{A,F}^2. \tag{211}$$

According to (203), $e_{A,F} \ll (\kappa_3^*)^{-1} r^{-1/2-\eta_1}\pi^{1/2}$. This implies

$$\kappa_3^* r^{3/2+2\eta_1} p^{1/2} e_{A,F}^2 \lesssim \kappa_2^* \log(np) r^{(1+\eta_1)\vee 0} p^{1/2}\pi^{1/2} e_{A,F}.$$

Thus, combining (209) and (211), we obtain

$$\begin{aligned}&\max_{i\in[n]}\{\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|+\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|+\beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*\}\\ \lesssim&\kappa_2^*\{\log^{1/2}(n) r\pi^{1/2} + \log(np) r^{(1+\eta_1)\vee 0} p^{1/2}\pi^{1/2} e_{A,F}\}.\end{aligned} \tag{212}$$

Next, we find a lower bound for $\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))$. With similar derivations as those for (129), we have

$$\min_{i\in[n]} \sigma_r^2(\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\mathbf{A}^*) \geq 2^{-1}\pi \tag{213}$$

with probability converging to 1 under the asymptotic regime $p\pi \gg r(\log(n))^2$. According to Lemma 47,

$$\max_{i\in[n]}\|\mathrm{diag}(\mathbf{\Omega}_{i\cdot})(\hat{\mathbf{A}}-\mathbf{A}^*)\|_2^2 \leq \|\hat{\mathbf{A}}-\mathbf{A}^*\|_F^2 \leq e_{A,F}^2. \tag{214}$$

According to (203), $e_{A,F} \ll \pi^{1/2}$. Thus, the above two inequalities and Lemma 25 together imply that with probability converging to 1,

$$\min_{i\in[n]} \sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}})) \geq 2^{-3}\delta_2^*\pi. \tag{215}$$

Next, we verify conditions of Lemma 16. According to Lemma 46, on the event $p_{\max} \leq 2p\pi_{\max}$, $\max_{i\in[n]}\gamma_{1,i}(\hat{\mathbf{A}}) \lesssim (p\pi(r/p)^{3/2})$. Following similar arguments as those for (133), we have with probability tending to 1,

$$\begin{aligned}&\min_{i\in[n]}\{(\gamma_{1,i}(\hat{\mathbf{A}}))^{-1}(\kappa_3(3C_1C_2))^{-1}\sigma_r^2(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))\}\\ \gtrsim&(p\pi)^{-1}(r/p)^{-3/2}(\kappa_3^*)^{-1}\pi^2(\delta_2^*)^2\\ =&(\kappa_3^*)^{-1}(\delta_2^*)^2 p^{1/2} r^{-3/2}\pi.\end{aligned} \tag{216}$$

Under the asymptotic regime $p\pi \gg (\kappa_2^*)^2(\kappa_3^*)^2(\delta_2^*)^{-4} r^5 \log(n)$, we have $\kappa_2^*\pi^{1/2} r(\log(n))^{1/2} \ll (\kappa_3^*)^{-1}(\delta_2^*)^2 p^{1/2} r^{-3/2}\pi$. Under the asymptotic regime $e_{A,F} \ll (\kappa_2^*)^{-1}(\kappa_3^*)^{-1}(\delta_2^*)^2(\log(np))^{-1} r^{(-5/2-\eta_1)\wedge(-3/2)}\pi^{1/2}$, we have $\kappa_2^*\log(np) r^{(1+\eta_1)\vee 0} p^{1/2}\pi^{1/2} e_{A,F} \ll (\kappa_3^*)^{-1}(\delta_2^*)^2 p^{1/2} r^{-3/2}\pi$. Combining the analysis, we have $\kappa_2^*\{\log^{1/2}(n) r\pi^{1/2} + \log(np) r^{(1+\eta_1)\vee 0} p^{1/2}\pi^{1/2} e_{A,F}\} \ll (\kappa_3^*)^{-1}(\delta_2^*)^2 p^{1/2} r^{-3/2}\pi$. This, together with (216) implies with probability tending to 1,

$$\max_{i\in[n]}\{\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|+\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|+\beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*\} \ll \min_{i\in[n]}\{(\gamma_{1,i}(\hat{\mathbf{A}}))^{-1}(\kappa_3(3C_1C_2))^{-1}\sigma_r^2(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))\}. \tag{217}$$

According to (215) and note that $C_1 \gtrsim r^{1/2+\eta_2}p^{1/2}$, we have

$$\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))C_2 \gtrsim \delta_2^*\pi r^{1/2+\eta_2}p^{1/2}. \tag{218}$$

According to (201), $p\pi \gg (\kappa_2^*)^2(\delta_2^*)^{-2}\log(n)r^{1-2\eta_2}$, which implies $\kappa_2^*\log^{1/2}(n)r\pi^{1/2} \ll \delta_2^*\pi r^{1/2+\eta_2}p^{1/2}$. According to (203), $e_{\mathbf{A},F} \ll \delta_2^*(\kappa_2^*)^{-1}(\log(np))^{-1}r^{(-1/2-\eta_1+\eta_2)\wedge(1/2+\eta_2)}\pi^{1/2}$, which implies $\kappa_2^*\log(np)r^{(1+\eta_1)\vee 0}p^{1/2}\pi^{1/2}e_{A,F} \ll \delta_2^*\pi r^{1/2+\eta_2}p^{1/2}$. Combining the analysis with (201) and (212), we obtain with probability tending to 1,

$$\max_{i\in[n]}\{\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|+\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|+\beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*\} \ll \min_{i\in[n]}\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}}))C_2. \tag{219}$$

Thus, conditions of Lemma 16 are satisfied. According to Lemma 16 with $\mathbf{A}$ replaced by $\hat{\mathbf{A}}$ and according to (212) and (215), we have $\|\tilde{\mathbf{\Theta}}-\mathbf{\Theta}^*\|_{2\to\infty}\leq C_1$ and

$$\begin{aligned}&\|\tilde{\mathbf{\Theta}}-\mathbf{\Theta}^*\|_{2\to\infty}\\ \leq&\max_{i\in[n]}\left[(\sigma_r(\mathcal{I}_{1,i}(\hat{\mathbf{A}})))^{-1}\{\|\mathbf{Z}_{i\cdot}\mathrm{diag}(\mathbf{\Omega}_{i\cdot})\hat{\mathbf{A}}\|+\|\mathbf{B}_{1,i}(\hat{\mathbf{A}})\|+\beta_{1,i}(\hat{\mathbf{A}})\kappa_3^*\}\right]\\ \lesssim&(\delta_2^*\pi)^{-1}\kappa_2^*\{r\log^{1/2}(n)\pi^{1/2}+\log(np)r^{(1+\eta_1)\vee 0}p^{1/2}\pi^{1/2}e_{A,F}\}\\ =&\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}\{r(\log(n))^{1/2}+\log(np)r^{(1+\eta_1)\vee 0}p^{1/2}e_{\mathbf{A},F}\}\end{aligned} \tag{220}$$

with probability converging to 1. Moreover, from (215) the optimization problem $\max_{\boldsymbol{\theta}_i\in\mathbb{R}^r}\sum_{j\in[p]}\omega_{ij}\{y_{ij}\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j - b(\boldsymbol{\theta}_i^T\hat{\mathbf{a}}_j)\}$ is strictly convex. Thus, $\tilde{\boldsymbol{\theta}}_i$ is the unique solution to this optimization problem.

■

B.3 Additional Theoretical Results for Algorithm 1 without Data Splitting

Lemma 49. *Let* $\tilde{\mathbf{M}}$ *be obtained by Algorithm 1. Assume that* $\lim_{n,p\to\infty}\mathbb{P}(\|\hat{\mathbf{M}}-\mathbf{M}^*\|_F\leq e_{\mathbf{M},F}) = 1$*, and the following asymptotic regime holds:*

1. $\phi\sim 1$, $\pi_{\min}\sim\pi_{\max}\sim\pi$;
2. $\|\mathbf{U}_r^*\|_{2\to\infty}\lesssim(r/n)^{1/2}$, $\|\mathbf{V}_r^*\|_{2\to\infty}\lesssim(r/p)^{1/2}$, $C_2\sim(r/p)^{1/2}$;
3. $(np)^{1/2}r^{\eta_2}\lesssim\sigma_r(\mathbf{M}^*)\leq\sigma_1(\mathbf{M}^*)\lesssim(np)^{1/2}r^{\eta_1}$ *for some constants* η_1 *and* η_2;
4. $p\pi \gg (\kappa_2^*)^4(\delta_2^*)^{-6}(\log(np))^3\cdot\max\left[r^{(1+2\eta_1)\vee(3+2\eta_1-4\eta_2)\vee(1-4\eta_2)},(\kappa_3^*)^2r^{\{7+8(\eta_1-\eta_2)\}\vee(5+6\eta_1-8\eta_2)}\right]$;
5. $n\pi \gg (\kappa_2^*)^2(\delta_2^*)^{-4}(\log(np))^2\max\{r^{(1+2\eta_1-2\eta_2)\vee(1+2\eta_1-4\eta_2)},(\kappa_3^*)^2r^{5+8\eta_1-8\eta_2}\}$;
6.

$$\begin{aligned}&(np)^{-1/2}e_{\mathbf{M},F}\\ \ll&(\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-2}\pi^{1/2}\cdot\\ &\min\left[r^{(1/2+2\eta_2)\wedge(-3/2-2\eta_1+3\eta_2)\wedge(-1/2-\eta_1+3\eta_2)\wedge(1/2+3\eta_2)},\right.\\ &\left.\qquad(\kappa_3^*)^{-1}r^{(-7/2-5\eta_1+5\eta_2)\wedge(-5/2-4\eta_1+5\eta_2)\wedge(-3/2-3\eta_1+5\eta_2)}\right].\end{aligned} \tag{221}$$

Then, with probability converging to 1*, estimating equations in steps 3 and 4 of Algorithm 1 have a unique solution and*

$$\begin{aligned}&\|\tilde{\mathbf{M}}-\mathbf{M}^*\|_{\max}\\ \lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2(\log(np))^2\Big[r^{(5/2+2\eta_1-2\eta_2)\vee(3/2+\eta_1-2\eta_2)}\{(n\wedge p)\pi\}^{-1/2}\\ &\qquad+r^{(5/2+3\eta_1-3\eta_2)\vee(3/2+2\eta_1-3\eta_2)\vee(1/2+\eta_1-3\eta_2)}(np\pi)^{-1/2}e_{\mathbf{M},F}\Big].\end{aligned}\tag{222}$$

Proof First, we analyze the asymptotic regime assumption. The 4-th condition of the asymptotic regime, i.e.,

$$p\pi\gg(\kappa_2^*)^4(\delta_2^*)^{-6}(\log(np))^3\cdot\max\Big[r^{(1+2\eta_1)\vee(3+2\eta_1-4\eta_2)\vee(1-4\eta_2)},(\kappa_3^*)^2r^{\{7+8(\eta_1-\eta_2)\}\vee(5+6\eta_1-8\eta_2)}\Big]\tag{223}$$

implies

$$p\pi\gg\begin{cases}(\delta_2^*)^{-4}(\kappa_2^*)^2\log^2(n)\max\{r^{1\vee(1-2\eta_2)},(\kappa_3^*)^2r^5\},\\ (\delta_2^*)^{-6}(\kappa_2^*)^4(\log(np))^3r^{(3+2\eta_1-4\eta_2)\vee(1-4\eta_2)},\\ (\delta_2^*)^{-6}(\kappa_2^*)^4(\kappa_3^*)^2(\log(np))^3r^{\{7+8(\eta_1-\eta_2)\}\vee(5+6\eta_1-8\eta_2)},\end{cases}\tag{224}$$

where we used the fact that $7+8(\eta_1-\eta_2)\geq 7>5$, $(1+2\eta_1)\vee(2-2\eta_2)\geq 1$, and $2-2\eta_2<3+2\eta_1-4\eta_2$.

The 6-th condition of the asymptotic regime, i.e.,

$$\begin{aligned}&(np)^{-1/2}e_{\mathbf{M},F}\\ \ll&(\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-2}\pi^{1/2}\cdot\\ &\min\big[r^{(1/2+2\eta_2)\wedge(-3/2-2\eta_1+3\eta_2)\wedge(-1/2-\eta_1+3\eta_2)\wedge(1/2+3\eta_2)},\\ &\qquad(\kappa_3^*)^{-1}r^{(-7/2-5\eta_1+5\eta_2)\wedge(-5/2-4\eta_1+5\eta_2)\wedge(-3/2-3\eta_1+5\eta_2)}\big]\end{aligned}\tag{225}$$

implies

$$(np)^{-1/2}e_{\mathbf{M},F}\ll\begin{cases}r^{\eta_2}(\kappa_2^*)^{-1}(\delta_2^*)^2(\log(np))^{-1}r^{0\wedge(-1/2-\eta_1+\eta_2)\wedge(1/2+\eta_2)}\pi^{1/2},\\ r^{\eta_2}(\kappa_2^*)^{-1}(\delta_2^*)^2(\log(np))^{-1}(\kappa_3^*)^{-1}r^{(-5/2-\eta_1)\wedge(-3/2)}\pi^{1/2},\\ (\delta_2^*)^3(\kappa_2^*)^{-2}(\log(np))^{-2}r^{(-3/2-2\eta_1+3\eta_2)\wedge(-1/2-\eta_1+3\eta_2)\wedge(1/2+3\eta_2)}\pi^{1/2},\\ (\delta_2^*)^3(\kappa_2^*)^{-2}(\kappa_3^*)^{-1}(\log(np))^{-2}r^{(-7/2-5\eta_1+5\eta_2)\wedge(-5/2-4\eta_1+5\eta_2)\wedge(-3/2-3\eta_1+5\eta_2)}\pi^{1/2},\end{cases}\tag{226}$$

where we used the fact that $\eta_2\geq-1/2-\eta_1+2\eta_2$, $-1/2-\eta_1+2\eta_2>-3/2-2\eta_1+3\eta_2$, $-5/2-\eta_1+\eta_2>-7/2-5\eta_1+5\eta_2$, and $-3/2+\eta_2>-5/2-4\eta_1+5\eta_2$.

According to (226),

$$\begin{aligned}&e_{\mathbf{M},F}\\ \ll&(np)^{1/2}r^{\eta_2}(\kappa_2^*)^{-1}(\delta_2^*)^2(\log(np))^{-1}\min\{r^{0\wedge(-1/2-\eta_1+\eta_2)\wedge(1/2+\eta_2)},(\kappa_3^*)^{-1}r^{(-5/2-\eta_1)\wedge(-3/2)}\}\pi^{1/2},\end{aligned}\tag{227}$$

which, together with Lemma 42, implies

$$
\begin{aligned}
&e_{\mathbf{A},F}\\
\ll&(\kappa_2^*)^{-1}(\delta_2^*)^2(\log(np))^{-1}\min\{r^{0\wedge(-1/2-\eta_1+\eta_2)\wedge(1/2+\eta_2)},(\kappa_3^*)^{-1}r^{(-5/2-\eta_1)\wedge(-3/2)}\}\pi^{1/2}.
\end{aligned} \tag{228}
$$

Also, according to the lemma's assumption, $p\pi\gg(\delta_2^*)^{-4}(\kappa_2^*)^2\log^2(n)\max\{r^{1\vee(1-2\eta_2)},(\kappa_3^*)^2r^5\}$. Thus, the conditions of Lemma 48 are satisfied. According to Lemma 48, $\|\hat{\mathbf{\Theta}}-\mathbf{\Theta}^*\|_{2\to\infty}\leq e_{\mathbf{\Theta},2\to\infty}$, with probability converging to 1, for $e_{\mathbf{\Theta},2\to\infty}$ satisfying

$$
\begin{aligned}
&e_{\mathbf{\Theta},2\to\infty}\\
\sim&\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}\{r(\log(n))^{1/2}+\log(np)r^{(1+\eta_1)\vee 0}p^{1/2}e_{\mathbf{A},F}\}\\
\lesssim&\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}\{r(\log(n))^{1/2}+\log(np)r^{(1+\eta_1)\vee 0}p^{1/2}\cdot r^{-\eta_2}(np)^{-1/2}e_{\mathbf{M},F}\}\\
\sim&\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}\{r(\log(n))^{1/2}+\log(np)r^{(1+\eta_1-\eta_2)\vee(-\eta_2)}n^{-1/2}e_{\mathbf{M},F}\}.
\end{aligned} \tag{229}
$$

Note that the proof of Lemma 39 does not require the independence between $\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}$ and the missing pattern $\mathbf{\Omega}$. Thus, following similar arguments, Lemma 39 still applies with $\tilde{\mathbf{\Theta}}_{\mathcal{N}_2}$ replaced with $\tilde{\mathbf{\Theta}}$ and $\mathcal{N}_2$ replaced with $[n]$. Next, we verify that the asymptotic regime of Lemma 39 is satisfied.

According to (224), $p\pi\gg(\delta_2^*)^{-6}(\kappa_2^*)^4(\log(np))^3r^{(3+2\eta_1-4\eta_2)\vee(1-4\eta_2)}$, which implies

$$
\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}r(\log(n))^{1/2}\ll(\delta_2^*)^2(\kappa_2^*)^{-1}p^{1/2}(\log(np))^{-1}r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)}. \tag{230}
$$

According to (224), $p\pi\gg(\delta_2^*)^{-6}(\kappa_2^*)^4(\kappa_3^*)^2(\log(np))^3r^{\{7+8(\eta_1-\eta_2)\}\vee(5+6\eta_1-8\eta_2)}$, which implies

$$
\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}r(\log(n))^{1/2}\ll(\delta_2^*)^2(\kappa_2^*)^{-1}p^{1/2}(\log(np))^{-1}(\kappa_3^*)^{-1}r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}. \tag{231}
$$

According to (226),
$(np)^{-1/2}e_{\mathbf{M},F}\ll(\delta_2^*)^3(\kappa_2^*)^{-2}(\log(np))^{-2}r^{(-3/2-2\eta_1+3\eta_2)\wedge(-1/2-\eta_1+3\eta_2)\wedge(1/2+3\eta_2)}\pi^{1/2}$, which implies

$$
\begin{aligned}
&\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}\log(np)r^{(1+\eta_1-\eta_2)\vee(-\eta_2)}n^{-1/2}e_{\mathbf{M},F}\\
\ll&(\delta_2^*)^2(\kappa_2^*)^{-1}p^{1/2}(\log(np))^{-1}r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)}.
\end{aligned} \tag{232}
$$

According to (226),
$(np)^{-1/2}e_{\mathbf{M},F}\ll(\delta_2^*)^3(\kappa_2^*)^{-2}(\kappa_3^*)^{-1}(\log(np))^{-2}r^{(-7/2-5\eta_1+5\eta_2)\wedge(-5/2-4\eta_1+5\eta_2)\wedge(-3/2-3\eta_1+5\eta_2)}\pi^{1/2}$, which implies

$$
\begin{aligned}
&\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}\log(np)r^{(1+\eta_1-\eta_2)\vee(-\eta_2)}n^{-1/2}e_{\mathbf{M},F}\\
\ll&(\delta_2^*)^2(\kappa_2^*)^{-1}p^{1/2}(\log(np))^{-1}(\kappa_3^*)^{-1}r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}.
\end{aligned} \tag{233}
$$

Combining equations (229), (230), (231), (232) and (233), we have

$$
\begin{aligned}
&e_{\mathbf{\Theta},2\to\infty}\\
\ll&\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}(\delta_2^*)^2(\kappa_2^*)^{-1}p^{1/2}(\log(np))^{-1}\cdot\min\{r^{(-1/2-\eta_1+2\eta_2)\wedge(1/2+2\eta_2)},\\
&\qquad(\kappa_3^*)^{-1}r^{(-5/2-4\eta_1+4\eta_2)\wedge(-3/2-3\eta_1+4\eta_2)}\},
\end{aligned} \tag{234}
$$

which implies $e_{\mathbf{\Theta},2\to\infty}$ satisfies the 5-th condition of the asymptotic regime of Lemma 39.

On the other hand, according to the lemma's assumption,

$$\begin{aligned}&n\pi_{\min}\\ \gg&(\kappa_2^*)^2(\delta_2^*)^{-4}(\log(np))^2\max\left\{(\pi_{\max}/\pi_{\min})r^{(1+2\eta_1-2\eta_2)\vee(1+2\eta_1-4\eta_2)},(\kappa_3^*)^2(\pi_{\max}/\pi_{\min})^3r^{5+8\eta_1-8\eta_2}\right\}.\end{aligned} \tag{235}$$

Thus, the other requirements for the asymptotic regime in Lemma 39 are also satisfied.

According to Lemma 39, we have $\|\tilde{\mathbf{A}}-\mathbf{A}^*\|_{2\to\infty}\le e_{\mathbf{A},2\to\infty}$ with probability converging to 1, where

$$e_{\mathbf{A},2\to\infty}\sim\kappa_2^*(\delta_2^*)^{-1}r^{-2\eta_2}\log(np)p^{-1/2}\Big\{r^{1+\eta_1}(n\pi)^{-1/2}+r^{(1+\eta_1)\vee 0}p^{-1/2}e_{\mathbf{\Theta},2\to\infty}\Big\}. \tag{236}$$

Combining the above display with (229), we further have

$$\begin{aligned}&e_{\mathbf{A},2\to\infty}\\ \lesssim&\kappa_2^*(\delta_2^*)^{-1}r^{-2\eta_2}\log(np)p^{-1/2}\Big[r^{1+\eta_1}(n\pi)^{-1/2}\\ &+r^{(1+\eta_1)\vee 0}p^{-1/2}\cdot\kappa_2^*(\delta_2^*)^{-1}\pi^{-1/2}\{r(\log(n))^{1/2}+\log(np)r^{(1+\eta_1-\eta_2)\vee(-\eta_2)}n^{-1/2}e_{\mathbf{M},F}\}\Big]\\ \lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2(\log(np))^2p^{-1/2}\Big[r^{(2+\eta_1-2\eta_2)\vee(1-2\eta_2)}\{(n\wedge p)\pi\}^{-1/2}\\ &\qquad+r^{(2+2\eta_1-3\eta_2)\vee(1+\eta_1-3\eta_2)\vee(-3\eta_2)}(np\pi)^{-1/2}e_{\mathbf{M},F}\Big].\end{aligned} \tag{237}$$

Next, we derive an asymptotic upper bound for $\|\tilde{\mathbf{M}}-\mathbf{M}^*\|_{\max}$. Recall that $\tilde{\mathbf{M}}=\tilde{\mathbf{\Theta}}\tilde{\mathbf{A}}^T$. Thus, for $\hat{\mathbf{P}}\in\mathcal{O}_{r\times r}$ defined in (183) and $\mathbf{\Theta}^*=(\mathbf{U}_r^*)\mathbf{D}_r^*\hat{\mathbf{P}}$, $\mathbf{A}^*=\mathbf{V}_r^*\hat{\mathbf{P}}$, we have $\tilde{\mathbf{M}}-\mathbf{M}^*=\tilde{\mathbf{\Theta}}\tilde{\mathbf{A}}^T-\mathbf{\Theta}^*(\mathbf{A}^*)^T=(\tilde{\mathbf{\Theta}}-\mathbf{\Theta}^*)(\mathbf{A}^*)^T+\tilde{\mathbf{\Theta}}(\tilde{\mathbf{A}}-\mathbf{A}^*)^T$. Thus,

$$\|\tilde{\mathbf{M}}-\mathbf{M}^*\|_{\max}\le\|\tilde{\mathbf{\Theta}}-\mathbf{\Theta}^*\|_{2\to\infty}\|\mathbf{A}^*\|_{2\to\infty}+\|\tilde{\mathbf{A}}-\mathbf{A}^*\|_{2\to\infty}\|\tilde{\mathbf{\Theta}}\|_{2\to\infty}. \tag{238}$$

According to Lemma 48 and the assumption $\|\mathbf{A}^*\|_{2\to\infty}\le C_2\lesssim(r/p)^{1/2}$, with probability converging to 1, the above display is further bounded by

$$\|\tilde{\mathbf{M}}-\mathbf{M}^*\|_{\max}\lesssim e_{\mathbf{\Theta},2\to\infty}r^{1/2}p^{-1/2}+e_{\mathbf{A},2\to\infty}r^{1/2+\eta_1}p^{1/2}. \tag{239}$$

Combining the above inequality with (229) and (237), we obtain with probability tending to 1

$$\begin{aligned}
&\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \\
\lesssim & r^{1/2} p^{-1/2} \cdot \kappa_2^* (\delta_2^*)^{-1} \pi^{-1/2} \{ r (\log(n))^{1/2} + \log(np) r^{(1+\eta_1-\eta_2)\vee(-\eta_2)} n^{-1/2} e_{\mathbf{M},F} \} \\
& + r^{1/2+\eta_1} p^{1/2} \cdot (\delta_2^*)^{-2} (\kappa_2^*)^2 (\log(np))^2 p^{-1/2} \Big[r^{(2+\eta_1-2\eta_2)\vee(1-2\eta_2)} \{(n \wedge p)\pi\}^{-1/2} \\
& \qquad + r^{(2+2\eta_1-3\eta_2)\vee(1+\eta_1-3\eta_2)\vee(-3\eta_2)} (np\pi)^{-1/2} e_{\mathbf{M},F} \Big] \\
\lesssim & (\delta_2^*)^{-2} (\kappa_2^*)^2 (\log(np))^2 \Big[r^{(5/2+2\eta_1-2\eta_2)\vee(3/2+\eta_1-2\eta_2)} \{(n \wedge p)\pi\}^{-1/2} \\
& + r^{(3/2+\eta_1-\eta_2)\vee(1/2-\eta_2)\vee(5/2+3\eta_1-3\eta_2)\vee(3/2+2\eta_1-3\eta_2)\vee(1/2+\eta_1-3\eta_2)} (np\pi)^{-1/2} e_{\mathbf{M},F} \Big] \\
\lesssim & (\delta_2^*)^{-2} (\kappa_2^*)^2 (\log(np))^2 \Big[r^{(5/2+2\eta_1-2\eta_2)\vee(3/2+\eta_1-2\eta_2)} \{(n \wedge p)\pi\}^{-1/2} \\
& \qquad\qquad + r^{(5/2+3\eta_1-3\eta_2)\vee(3/2+2\eta_1-3\eta_2)\vee(1/2+\eta_1-3\eta_2)} (np\pi)^{-1/2} e_{\mathbf{M},F} \Big],
\end{aligned} \tag{240}$$

where we used the fact that $3/2 + \eta_1 - \eta_2 < 5/2 + 3\eta_1 - 3\eta_2$ and $1/2 - \eta_2 < 3/2 + 2\eta_1 - 3\eta_2$ in the last inequality. This completes the proof. ■

B.4 Proof of Theorem 5

Proof [Proof of Theorem 5]

Note that when $\pi_{\min} \sim \pi_{\max} \sim \pi$ and $\eta_1 = \eta_2 = \eta$, the 4-th asymptotic requirement in Lemma 49 becomes

$$p\pi \gg (\kappa_2^*)^4 (\delta_2^*)^{-6} (\log(np))^3 \cdot \max \Big[r^{(1+2\eta)\vee(3-2\eta)\vee(1-4\eta)}, (\kappa_3^*)^2 r^{7\vee(5-2\eta)} \Big]. \tag{241}$$

When $\eta \geq -1$, the above requirement is implied by

$$p\pi \gg (\kappa_2^*)^4 (\delta_2^*)^{-6} (\log(np))^3 \cdot \max \Big[r^{(1+2\eta)\vee 5}, (\kappa_3^*)^2 r^7 \Big], \tag{242}$$

which is implied by the asymptotic requirement *R5*.

Similarly, the 5-th asymptotic requirement in Lemma 49 becomes

$$n\pi \gg (\kappa_2^*)^2 (\delta_2^*)^{-4} (\log(np))^2 \max \{ r^{1\vee(1-2\eta)}, (\kappa_3^*)^2 r^5 \},$$

which is implied by the asymptotic requirement *R6*: $n\pi \gg (\kappa_2^*)^2 (\delta_2^*)^{-4} (\log(np))^2 \max \{ r^3, (\kappa_3^*)^2 r^5 \}$.

The 6-th asymptotic requirement in Lemma 49 becomes

$$\begin{aligned}
(np)^{-1/2} e_{\mathbf{M},F} \ll (\kappa_2^*)^{-2} (\delta_2^*)^3 (\log(np))^{-2} \pi^{1/2} \min \big[& r^{(1/2+2\eta)\wedge(-3/2+\eta)\wedge(-1/2+2\eta)\wedge(1/2+3\eta)}, \\
& (\kappa_3^*)^{-1} r^{(-7/2)\wedge(-5/2+\eta)\wedge(-3/2+2\eta)} \big]
\end{aligned} \tag{243}$$

and is implied by *R7*: $(np)^{-1/2} e_{\mathbf{M},F} \ll (\kappa_2^*)^{-2} (\delta_2^*)^3 (\log(np))^{-2} \pi^{1/2} \min [r^{-5/2}, (\kappa_3^*)^{-1} r^{-7/2}]$ for $\eta \geq -1$.

Thus, under *R1-R7*, the conditions of Lemma 49 are satisfied, and thus with probability converging to 1,

$$\begin{aligned}
&\|\tilde{\mathbf{M}} - \mathbf{M}^*\|_{\max} \\
\lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2(\log(np))^2 \\
&\cdot \Big[r^{(5/2+2\eta_1-2\eta_2)\vee(3/2+\eta_1-2\eta_2)}\{(n\wedge p)\pi\}^{-1/2} \\
&\qquad + r^{(5/2+3\eta_1-3\eta_2)\vee(3/2+2\eta_1-3\eta_2)\vee(1/2+\eta_1-3\eta_2)}(np\pi)^{-1/2}e_{\mathbf{M},F}\Big] \\
\lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2(\log(np))^2\Big[r^{5/2\vee(3/2-\eta)}\{(n\wedge p)\pi\}^{-1/2} + r^{(5/2)\vee(3/2-\eta)\vee(1/2-2\eta_2)}(np\pi)^{-1/2}e_{\mathbf{M},F}\Big] \\
\lesssim&(\delta_2^*)^{-2}(\kappa_2^*)^2(\log(np))^2\Big[r^{5/2}\{(n\wedge p)\pi\}^{-1/2} + r^{5/2}(np\pi)^{-1/2}e_{\mathbf{M},F}\Big].
\end{aligned} \tag{244}$$

The above analysis gives the error bound of $\tilde{\mathbf{M}}$. The proof for the 'in particular' part of the theorem is similar to that of the proof of Theorem 10, and we skip the repetitive details. ■

Appendix C. Proof of Corollaries

Proof [Proof of Corollary 8] For binomial model $b''(x) = ke^x(1+e^x)^{-2}$ and $b^{(3)}(x) = ke^x(1+e^x)^{-2}\{1-2(1+e^{-x})^{-1}\}$. Thus, $\kappa_2(\alpha) \le k$, $\kappa_3(\alpha) \le k$, and $\delta_2(\alpha) \ge ke^{\alpha}(1+e^{\alpha})^{-2} \gtrsim ke^{-\alpha}$. This implies that $\kappa_2^*, \kappa_3^* \lesssim 1$ under the asymptotic regime that $k \sim 1$ (*R8B*). Also, $\delta_2^* \gtrsim ke^{-2(\rho+1)} \gtrsim e^{-2\rho} \gtrsim ke^{-2\log(n\wedge p)^{1-\epsilon_0}} \gg (n\wedge p)^{-\epsilon_1}$ for any constant $\epsilon_1 > 0$, where the third inequality is due to *R9B*. Combining the analysis above, we have $(\kappa_2^*)^4(\delta_2^*)^{-6}\log(np)^3 \ll (n\vee p)^{6\epsilon_1}\log(np)^3 \ll (n\vee p)^{7\epsilon_1}$. Similarly, $(\kappa_2^*)^2(\delta_2^*)^{-4}(\log(np))^2 \ll (n\vee p)^{5\epsilon_1}$, and $(\kappa_2^*)^{-2}(\delta_2^*)^3(\log(np))^{-2} \gg (n\wedge p)^{-4\epsilon_1}$. Combine the above analysis with *R5B – R7B*, and note that $(1+2\eta)\vee 5 \le (3+4\eta)\vee 7$ for $\eta \ge -1$, we verify that *R5 – R7* hold with $7\epsilon_1 < \epsilon_0$.

For normal model, $b''(x) = 1$ and $b^{(3)}(x) = 0$ for all x. Thus, $\kappa_2^* = \delta_2^* = 1$ and $\kappa_3^* = 0$. Part 2 of Corollary 8 then follows by simplifying Theorem 5.

In the rest of the analysis, we focus on the Poisson model. Note that $\|\mathbf{M}^*\| \le C_1C_2$ so we could choose $\rho \le C_1C_2 \lesssim r^{1+\eta}$. Under *R10P*, $r^{1+\eta} \lesssim (\log(n\wedge p))^{1-\epsilon_0}$, so $\max(\rho, C_1C_2) \lesssim (\log(n\wedge p))^{1-\epsilon_0}$.

For Poisson model, $b(x) = e^x$ so $b''(x) = b^{(3)}(x) = e^x$. Thus, $\kappa_2(\alpha), \kappa_3(\alpha) \le e^{\alpha}$ and $\delta_2(\alpha) \ge e^{-\alpha}$. This implies $\kappa_2^* \le e^{2\rho+1} \lesssim e^{2\rho} \lesssim e^{2(\log(n\wedge p))^{1-\epsilon_0}} \lesssim (n\wedge p)^{\epsilon_1}$ for any constant $\epsilon_1 > 0$. Similarly, $\delta_2^* \gtrsim e^{-2\rho} \gtrsim (n\wedge p)^{-\epsilon_1}$ and $\kappa_3^* \lesssim e^{6C_1C_2} \lesssim (n\vee p)^{\epsilon_1}$ for any constant $\epsilon_1 > 0$. The proof then follows similarly as that for the normal model. ■

Proof [Proof of Corollary 12] The proof of Corollary 12 is similar to that of Corollary 8, except that *R7B* is replaced by *R7'B* to ensure *R7'* holds. We omit the repetitive details. ■

Setting	n	p	r	π	Variable Types	Setting	n	p	r	π	Variable Type
1	400	200	3	0.6	O	13	400	200	5	0.6	O
2	800	400	3	0.6	O	14	800	400	5	0.6	O
3	1600	800	3	0.6	O	15	1600	800	5	0.6	O
4	400	200	3	0.2	O	16	400	200	5	0.2	O
5	800	400	3	0.2	O	17	800	400	5	0.2	O
6	1600	800	3	0.2	O	18	1600	800	5	0.2	O
7	400	200	3	0.6	O + C	19	400	200	5	0.6	O + C
8	800	400	3	0.6	O + C	20	800	400	5	0.6	O + C
9	1600	800	3	0.6	O + C	21	1600	800	5	0.6	O + C
10	400	200	3	0.2	O + C	22	400	200	5	0.2	O + C
11	800	400	3	0.2	O + C	23	800	400	5	0.2	O + C
12	1600	800	3	0.2	O + C	24	1600	800	5	0.2	O + C

Table 5: Simulation settings. 'Variable type = O' indicates all the variables are ordinal (with $k_j = 5$), and 'Variable type = O + C' indicates half of the variables are ordinal (with $k_j = 5$) and half are continuous. For continuous and ordinal variables, we assume the Normal and Binomial models, respectively.

Appendix D. Simulation Settings and Additional Results

D.1 Simulation Setting Details

A full list of our simulation settings is given in Table 5 below. For each setting, data are generated as follows. For each replication, we first generate $\mathbf{\Theta}^* = (\theta^*_{ik})_{n\times r}$ and $\mathbf{A}^* = (a^*_{ij})_{p\times r}$, where θ^*_{ik}s and a^*_{ij}s are independently from a uniform distribution over the interval $[-0.9, 0.9]$. Then $\mathbf{M}^*$ is given by $\mathbf{M}^* = \mathbf{\Theta}^*(\mathbf{A}^*)^T$. The missing indicators ω_{ij}s are generated independently from a Bernoulli distribution with parameter π, where $\pi = 0.6$ and 0.2 are considered in the simulation settings. When $\omega_{ij} = 1$ and for an ordinal variable j, Y_{ij} is generated from a Binomial distribution with $k_j = 5$ trials and success probability $\exp(m^*_{ij})/(1 + \exp(m^*_{ij}))$. When $\omega_{ij} = 1$ and for an continuous variable j, Y_{ij} is generated from a normal distribution $N(m^*_{ij}, 1)$. In the implementation, we set $C_2 = 2\sqrt{r/p}$ in Algorithms 1, 2, and 3. We set $\rho' = r$ in the NBE and $C = \sqrt{r}$ in the CJMLE.

D.2 Additional Simulation Results

In Figures 3 though 8 below, we give the results under Settings 7 through 24. The patterns are similar to those in Figures 1 and 2, except for few cases when n and p are relatively small.

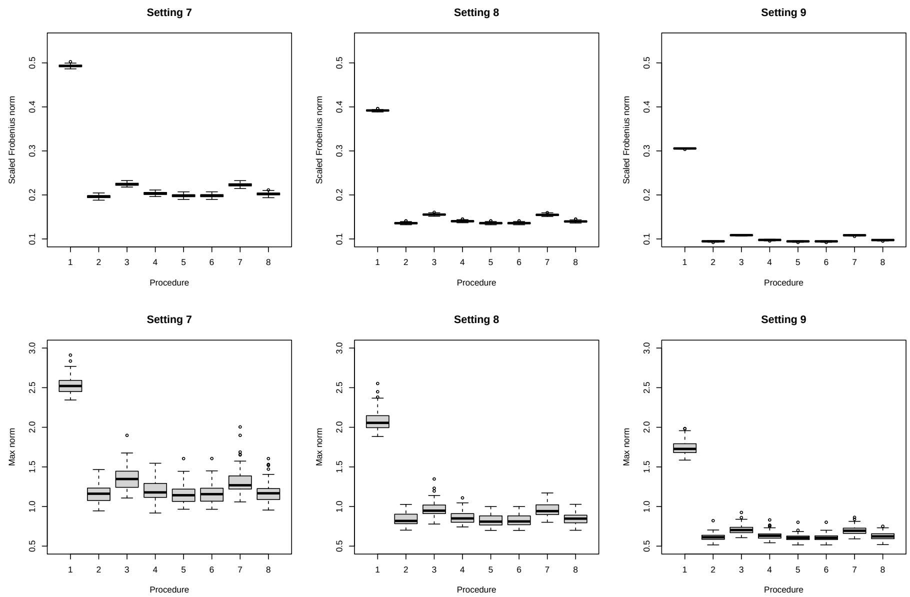

Figure 3: Results from Simulation Settings 7-9. The plots can be interpreted similarly as those in Figure 1.

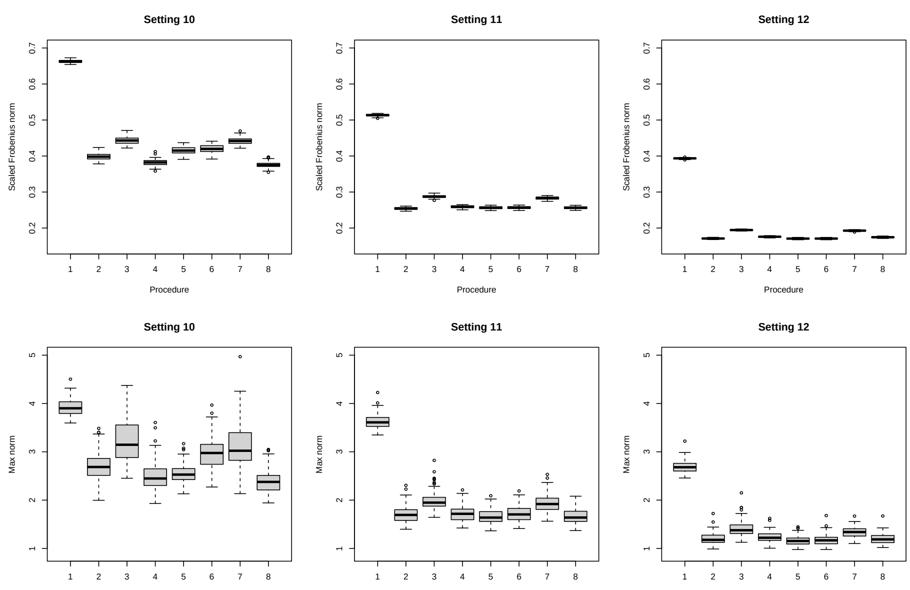

Figure 4: Results from Simulation Settings 10-12. The plots can be interpreted similarly as those in Figure 1.

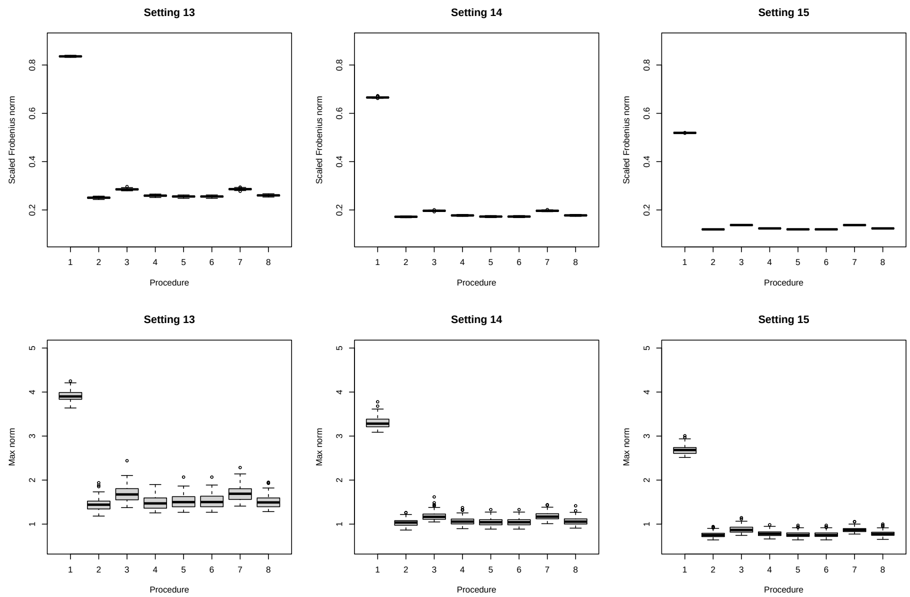

Figure 5: Results from Simulation Settings 13-15. The plots can be interpreted similarly as those in Figure 1.

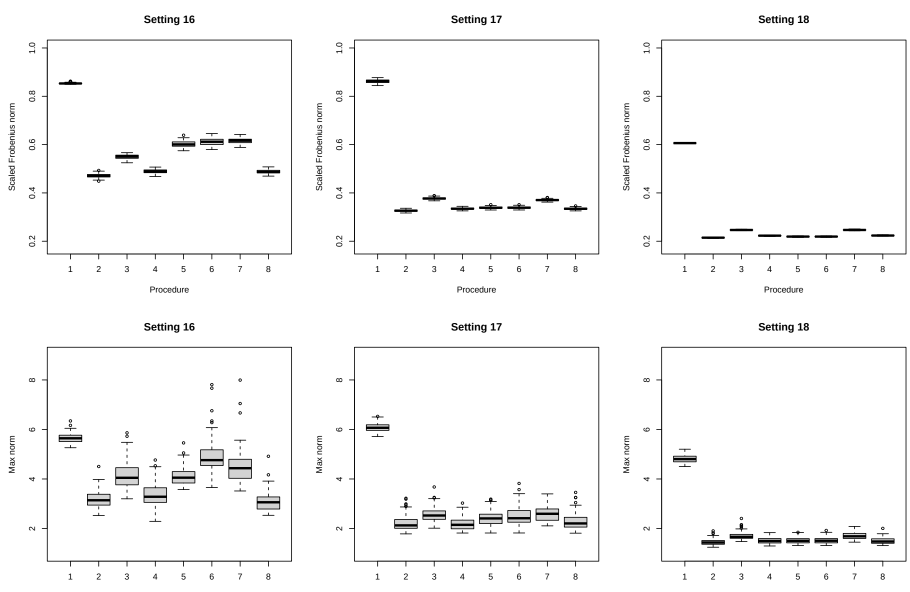

Figure 6: Results from Simulation Settings 16-18. The plots can be interpreted similarly as those in Figure 1.

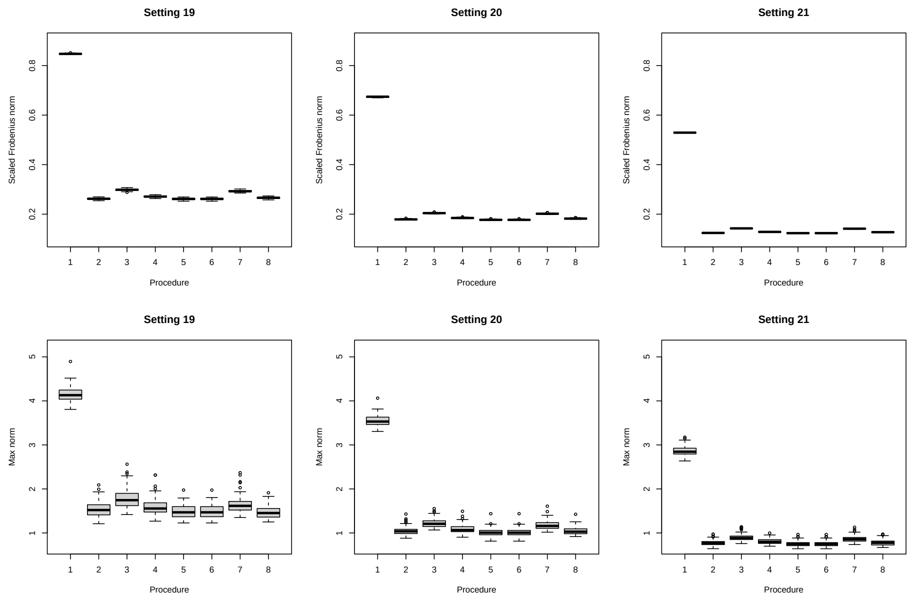

Figure 7: Results from Simulation Settings 19-21. The plots can be interpreted similarly as those in Figure 1.

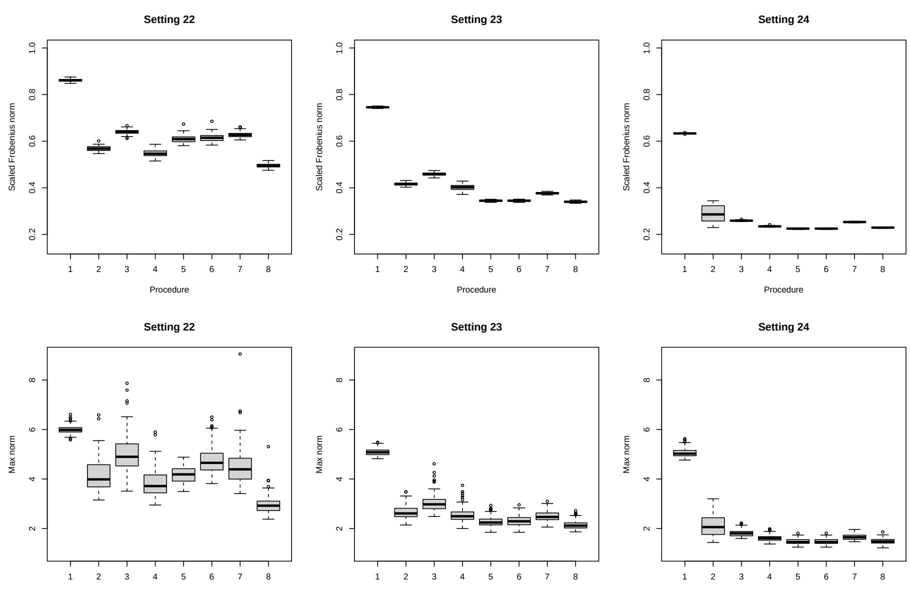

Figure 8: Results from Simulation Settings 22-24. The plots can be interpreted similarly as those in Figure 1.

References

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.

Mokhtar Z Alaya and Olga Klopp. Collective matrix completion. *Journal of Machine Learning Research*, 20:1–43, 2019.

David J Bartholomew, Fiona Steele, Irini Moustaki, and Jane I Galbraith. *Analysis of multivariate social science data*. CRC Press, Boca Raton, FL, 2008.

Yoav Bergner, Peter Halpin, and Jill-Jênn Vie. Multidimensional item response theory in the style of collaborative filtering. *Psychometrika*, 87(1):266–288, 2022.

Sonia A Bhaskar. Probabilistic low-rank matrix completion from quantized measurements. *The Journal of Machine Learning Research*, 17(1):2131–2164, 2016.

Pratik Biswas, Tzu-Chen. Lian, Ta-Chung. Wang, and Yinyu Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Transactions on Sensor Networks (TOSN)*, 2 (2):188–220, 2006.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford, England, 2013.

Tony Cai and Wen-Xin Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

Tony Cai and Wen-Xin Zhou. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1):1493–1525, 2016.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Yang Cao and Yao Xie. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2015.

Joshua Cape, Minh Tang, and Carey E Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Annals of Statistics*, 47(5):2405–2439, 2019.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

Yunxiao Chen and Xiaoou Li. Determining the number of factors in high-dimensional generalized latent factor models. *Biometrika*, 109(3):769–782, 2022.

Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84(1):124–146, 2019a.

Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115(532):1756–1770, 2020a.

Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115:1756–1770, 2020b.

Yunxiao Chen, Chengcheng Li, Jing Ouyang, and Gongjun Xu. Statistical inference for noisy incomplete binary matrix. *Journal of Machine Learning Research*, 24(95):1–66, 2023.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019b.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020c.

Victor Chernozhukov, Christian Hansen, Yuan Liao, and Yinchu Zhu. Inference for low-rank models. *The Annals of statistics*, 51(3):1309–1330, 2023.

Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3:189–223, 2014.

Andrey Feuerverger, Yu He, and Shashi Khatri. Statistical significance of the Netflix challenge. *Statistical Science*, 27:202–231, 2012.

David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

Shelby J Haberman. When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33(2):204–229, 2008.

Ruijian Han, Rougang Ye, Chunxi Tan, and Kani Chen. Asymptotic theory of sparse Bradley–Terry model. *Annals of Applied Probability*, 30:2491–2515, 2020.

Ruijian Han, Yiming Xu, and Kani Chen. A general pairwise comparison model for extremely sparse networks. *Journal of the American Statistical Association*, 118(544):2422–2432, 2023.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 5(4):1–19, 2015.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

Anura P Jayasumana, Randy Paffenroth, Gunjan Mahindre, Sridhar Ramasamy, and Kelum Gajamannage. Network topology mapping from partial virtual coordinates and graph geodesics. *IEEE/ACM Transactions on Networking*, 27(6):2405–2417, 2019.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20 (1):282–303, 2014.

Olga Klopp, Jean Lafond, Eric Moulines, and Joseph Salmon. Adaptive multinomial matrix completion. *Electronic Journal of Statistics*, 9(2):2950–2975, 2015.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Geofferey N Masters and Benjamin D Wright. The essential process in a family of measurement models. *Psychometrika*, 49(4):529–544, 1984.

Andrew D McRae and Mark A Davenport. Low-rank matrix completion and denoising under Poisson noise. *Information and Inference: A Journal of the IMA*, 10(2):697–720, 2021.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

OECD. *PISA 2018 assessment and analytical framework*. OECD Publishing, Paris, France, 2019a.

OECD. *PISA 2018 techinical report*. OECD Publishing, Paris, France, 2019b.

James M Ortega and Werner C Rheinboldt. *Iterative solution of nonlinear equations in several variables*. SIAM, Philadelphia, PA, 2000.

Mark Reckase. *Multidimensional item response theory*. Springer, New York, NY, 2009.

Geneviève Robin, Julie Josse, Éric Moulines, and Sylvain Sardy. Low-rank model with covariates for count data with missing values. *Journal of Multivariate Analysis*, 173:416–434, 2019.

Geneviève Robin, Olga Klopp, Julie Josse, Éric Moulines, and Robert Tibshirani. Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association*, 115(531):1292–1303, 2020.

Anders Skrondal and Sophia Rabe-Hesketh. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press, Boca Raton, FL, 2004.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Michel Wedel, Ulf Böckenholt, and Wagner A Kamakura. Factor models for multivariate count data. *Journal of Multivariate Analysis*, 87(2):356–369, 2003.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

Dong Xia and Ming Yuan. Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1):58–77, 2021.

Haoran Zhang, Yunxiao Chen, and Xiaoou Li. A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, 85(2):358–372, 2020.