One-Step Estimator for Permuted Sparse Recovery

Hang Zhang
Amazon
410 Terry Ave N, Seattle, WA 98109, USA
hagzhang@amazon.com

Ping Li
LinkedIn Ads
700 Bellevue Way NE, Bellevue, WA 98004, USA
pinli@linkedin.com

Abstract

This paper considers the unlabeled sparse recovery under multiple measurements, i.e., $\mathbf{Y} = \mathbf{\Pi}^{\natural}\mathbf{X}\mathbf{B}^{\natural} + \mathbf{W}$, where $\mathbf{Y} \in \mathbb{R}^{n\times m}, \mathbf{\Pi}^{\natural} \in \mathbb{R}^{n\times n}, \mathbf{X} \in \mathbb{R}^{n\times p}, \mathbf{B}^{\natural} \in \mathbb{R}^{p\times m}, \mathbf{W} \in \mathbb{R}^{n\times m}$ represents the observations, missing (or incomplete) correspondence information, sensing matrix, sparse signals, and additive sensing noise, respectively. Different from the previous works on multiple measurements ($m > 1$) which all focus on the sufficient samples regime, namely, $n > p$, we consider a sparse matrix $\mathbf{B}^{\natural}$ and investigate the insufficient samples regime (i.e., $n \ll p$) for the first time. To begin with, we establish the lower bound on the sample number and *signal-to-noise ratio* (SNR) for the correct permutation recovery. Then, we present a simple yet effective estimator. Under mild conditions, we show that our estimator can restore the correct correspondence information with high probability. Numerical experiments are presented to corroborate our theoretical claims.

1. Introduction

In recent years, linear regression with permuted correspondence has received increasing attention due to its wide applications in the field of machine learning, signal processing, and statistics. Among all these applications, two most prominent examples are (i) *linkage record*, which merges two datasets of the same objects into one comprehensive dataset; and (ii) *data de-anonymization*, which infers the hidden labels of private data with the help of public datasets. Apart from these two applications, other applications include *correspondence estimation* between pose and estimation in graphics; *time-domain sampling* in the presence of clock jitter; *multi-target tracking*; *unsupervised data alignment*, etc (Pananjady et al., 2018; Slawski and Ben-David, 2019; Zhang et al., 2018; Slawski et al., 2020; 2019; Zhang et al., 2019; Fang and Li, 2022; Zhang et al., 2022; Zhang and Li, 2023). In this paper, we consider the canonical model, i.e., a linear sensing relation with permuted labels:

$$\mathbf{Y} = \mathbf{\Pi}^{\natural}\mathbf{X}\mathbf{B}^{\natural} + \mathbf{W},$$

where $\mathbf{Y} \in \mathbb{R}^{n\times m}$ is the sensing result, $\mathbf{\Pi}^{\natural} \in \mathbb{R}^{n\times n}$ is an unknown permutation matrix, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design (sensing) matrix, $\mathbf{B}^{\natural} \in \mathbb{R}^{p\times m}$ represents the sparse signals of interest, and $\mathbf{W} \in \mathbb{R}^{n\times m}$ denotes the additive noise. Assuming the signal $\mathbf{B}^{\natural}$ is a sparse signal, to put more specifically, each column of $\mathbf{B}^{\natural}$ is k-sparse, we would like to (i) study the statistical limits of the permutation recovery under this scenario, e.g., the minimum sample number n and *signal-to-noise ratio* (SNR); and (ii) propose a practical estimator that can efficiently recover the permutation once the minimum requirements are met. To begin with, we briefly review the previous works.

Related Works. The study of permuted linear regression has a long history that can at least date back to DeGroot and Goel (1976; 1980); Goel (1975); Bai and Hsing (2005). Recent interest in this area start from Unnikrishnan et al. (2015). Focusing on the noiseless case $\mathbf{W} = \mathbf{0}$ with single measurement $(m = 1)$, Unnikrishnan et al. (2015) establish the necessary condition $n \geq 2p$ for the permutation recovery if $\mathbf{B}^{\natural}$ is an arbitrary vector residing within the linear space $\mathbb{R}^{p}$. Later, Pananjady et al. (2018) extend the analysis to the noisy scenario. They showed the minimum SNR should be at least the order of $\Omega(n^{c})$, where $c > 0$ is some positive constant. Numerical experiments suggest c is within the region $[4, 5]$. Other works such as Hsu et al. (2017); Abid et al. (2017); Slawski and Ben-David (2019); Tsakiris et al. (2020); Haghighatshoar and Caire (2018) also focus on this regime and obtain the same answer. In Emiya et al. (2014), the setting with a sparse signal $\mathbf{B}^{\natural}$ is first studied. However, only empirical investigation is conducted without rigorous theoretical analysis. In the first work with theoretical analysis (Zhang and Li, 2021), both the statistical limits and practical estimators with almost optimal performance are presented for the permutation recovery. Peng et al. (2021) study the problem from the viewpoint of algebraic geometry.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

All existing works suggest that $\mathsf{SNR} = \Omega(n^c)$ is inevitable for the permutation reconstruction if only one measurement is conducted, namely, $m = 1$.

On the other hand, numerous works suggest multiple measurements, i.e., $m > 1$, can greatly reduce the SNR requirement, even to some positive constant. This line of research starts from Zhang et al. (2019; 2022), where the information-theoretic lower bounds and the *maximum likelihood* (ML) estimator are investigated. Later, Zhang and Li (2020) study this problem from the viewpoint of non-convex optimization and propose an optimal estimator for the permutation recovery. Independently, Slawski et al. (2020) investigate this problem from the viewpoint of denoising. Putting parsimonious constraints on the number of permuted rows, they view $(\mathbf{I} - \mathbf{\Pi}^{\natural})\mathbf{X}\mathbf{B}^{\natural}$ as sparse outliers and design the permutation recovery algorithm accordingly. Notice that all these works focus on the sufficient samples regime, namely, $n = \Omega(p)$. In this paper, we focus on the insufficient samples regime. Assuming $\mathbf{B}^{\natural}$ to be sparse, we would like to show the correct permutation can be obtained with $n \ll p$ and $\mathsf{SNR} = O(1)$.

Our **contributions** are summarized as follows

- We study the lower bounds w.r.t the number n and *signal-to-noise ratio* (SNR) for the correct reconstruction of both the permutation matrix $\mathbf{\Pi}^{\natural}$ and the signal $\mathbf{B}^{\natural}$. Assuming each column $\mathbf{B}^{\natural}_{:,\ell}$ $(1 \leq \ell \leq m)$ is k-sparse, we show that the sample number n should be at least of the order $\Omega(k\log p)$; meanwhile the SNR should satisfy $\log\det(\mathbf{I} + \mathbf{B}^{\natural\top}\mathbf{B}^{\natural}/\sigma^2) > \log n + \frac{m\log\binom{p}{k}}{n}$.
- We propose a one-step estimator for the correspondence recovery, which consists of two sub-parts: one for $\mathbf{\Pi}^{\natural}$ and another for $\mathbf{B}^{\natural}$. By formulating the correspondence recovery as a *linear assignment problem* (LAP) (Kuhn, 1955; Bertsekas and Castañón, 1992; Burkard et al., 2012), we show the correct permutation matrix can be obtained when (i) SNR **is above certain positive constant** and (ii) **the minimum sample number** n **is in the linear dependence with the sparsity number** k.

On top of the above contributions, we would like to briefly mention our proof strategy, which is based on a tailored version of the **leave-one-out** technique. Compared with the previous works that adopt the leave-out-out technique (Chen et al., 2020; Sur et al., 2019; Karoui, 2013; 2018; Cai et al., 2021), our construction has the following characteristics:

- Our construction method is adaptive, which replaces multiple rows ranging from 2 to 4 depending on each permuted row. Meanwhile, previous works such as Chen et al. (2020); Sur et al. (2019); Karoui (2013; 2018); Cai et al. (2021) replace a fixed number of rows (or columns).
- We not only leave out the rows but also modify the thresholding operator operated on the perturbed samples $\widetilde{\mathbf{B}}_{(\dots)}$ from $\mathsf{thres}(\cdot)$ to $(\cdot)_{\mathsf{imax}}$ (its definition is deferred to Subsection 4.2). This step is essential in controlling the approximation error since otherwise the non-zero elements in matrices $\widehat{\mathbf{B}}$ $(\propto \mathbf{X}^{\top}\mathbf{Y})$ and $\widetilde{\mathbf{B}}_{(\cdot)}$ may not share the same position and the approximation error can be considerably large. A detailed explanation is deferred to the proof of Theorem 3.

Notations. Denote c, $c^{'}$, c_i as some positive constants, whose values are not necessarily the same even for those with the same notations. We denote $a \lesssim b$ if there exists some positive constants $c_0 > 0$ such that $a \leq c_0 b$. Similarly, we define $a \gtrsim b$ provided $a \geq c_0 b$ for some positive constant c_0. We write $a \asymp b$ when $a \lesssim b$ and $a \gtrsim b$ hold simultaneously.

For an arbitrary matrix $\mathbf{M}$, we denote $\mathbf{M}_{i,:}$ as its ith row, $\mathbf{M}_{:,i}$ as its ith column, and $\mathbf{M}_{ij}$ as its (i,j)th element. Its Frobenius norm is defined as $\|\mathbf{M}\|_{\mathsf{F}}$ while the operator norm is denoted as $\|\mathbf{M}\|_{\mathsf{OP}}$, whose definitions can be found in Section 2.3 of Golub and Van Loan (2012). In addition, we define its stable rank as $\mathsf{srank}(\cdot) \triangleq \|\cdot\|_{\mathsf{F}}^2/\|\cdot\|_{\mathsf{OP}}^2$ (Section 2.1.15 in Tropp (2015)) and its support set $\mathsf{supp}(\cdot)$ as $\{(i,j) : (\cdot)_{i,j} \neq 0\}$. The inner product between matrices is denoted as $\langle\!\langle \cdot,\cdot \rangle\!\rangle$ while the inner product between vectors is denoted as $\langle \cdot,\cdot \rangle$.

We define the set of all possible permutation matrices as $\mathcal{P}_n$, which is defined as $\{\mathbf{\Pi} \in \{0,1\}^{n\times n} : \sum_{i=1}^n \mathbf{\Pi}_{ij} = 1, \sum_{j=1}^n \mathbf{\Pi}_{ij} = 1\}$. Associate with each permutation matrix $\mathbf{\Pi}$, we define the operator $\pi(\cdot)$ that transforms index i to $\pi(i)$ under $\mathbf{\Pi}$. The Hamming distance $\mathsf{d_H}(\mathbf{\Pi}_1, \mathbf{\Pi}_2)$ between two permutation matrices $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ is defined as $\mathsf{d_H}(\mathbf{\Pi}_1, \mathbf{\Pi}_2) \triangleq \sum_{i=1}^n \mathbb{1}(\pi_1(i) \neq \pi_2(i))$. The SNR is defined as $\|\mathbf{B}^{\natural}\|_{\mathsf{F}}^2/(m\cdot\sigma^2)$.

Organizations. In Section 2, we formally state our problem setting and briefly discuss its relations with previous works. In Section 3, we investigate the mini-max lower bound. In Section 4, we present a practical estimator for the permutation and signal recovery; along with its performance analysis. We show that its performance almost matches the mini-max lower bound. In Section 5, we present the numerical experiments and draw the conclusions in Section 6.

2. Problem Formulation

We start this section with a formal restatement of the considered problem reading as

$$\mathbf{Y} = \mathbf{\Pi}^{\natural}\mathbf{X}\mathbf{B}^{\natural} + \mathbf{W}, \tag{1}$$

where $\mathbf{\Pi}^{\natural} \in \mathcal{P}_n$ denotes the (fixed but unknown) permutation matrix, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the sensing (design) matrix with its entries being i.i.d. standard normal random variables, i.e., $\mathbf{X}_{ij} \sim \mathsf{N}(0,1)$, [1] $\mathbf{B}^{\natural} \in \mathbb{R}^{p\times m}$ is a fixed sparse matrix awaiting to be reconstructed (corresponding to the sparse signal), and $\mathbf{W} \in \mathbb{R}^{n\times m}$ represents the additive sensing noise with each entry $\mathbf{W}_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{N}(0,\sigma^2)$. Here we put the separate sparse constraints on each column of $\mathbf{B}^{\natural}$, namely, $\|\mathbf{B}^{\natural}_{:,\ell}\|_0 \leq k$ $(1 \leq \ell \leq m)$. In addition, we denote h as the number of permuted rows, or equivalently, the Hamming distance between the identify matrix $\mathbf{I}$ and the permutation matrix $\mathbf{\Pi}^{\natural}$, namely, $h \triangleq \mathsf{d}_{\mathsf{H}}(\mathbf{I}, \mathbf{\Pi}^{\natural})$.

Our goal is to reconstruct both the permutation matrix $\mathbf{\Pi}^{\natural}$ and sparse signal $\mathbf{B}^{\natural}$ from (1). Note that we do not assume that different columns of $\mathbf{\Pi}^{\natural}$ share the same support set. Actually, we prefer each column to be with a different support set, since otherwise $\mathrm{rank}(\mathbf{B}^{\natural})$ will be bounded by k and will bring more difficulties to the permutation recovery. A detailed explanation is deferred to Section 4.

Before proceeding, we briefly review the prior art. In Unnikrishnan et al. (2015), where $\mathbf{B}^{\natural}$ can reside within the entire linear space $\mathbb{R}^p$, it is proved that $n \geq 2p$ is required for the correct permutation recovery. As a result, subsequent works such as Pananjady et al. (2018); Zhang et al. (2019); Slawski and Ben-David (2019); Slawski et al. (2020); Zhang and Li (2020) all focus on the sufficient sample regime, i.e., $n = \Omega(p)$. Only until Zhang and Li (2021), the insufficient sample regime, i.e., $n = o(p)$, receives its first theoretical investigation. Similar to our setting, they put sparsity assumptions on $\mathbf{B}^{\natural}$ however focus on the single measurement scenario, namely, $m = 1$. They show that the minimum SNR for correct $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\mathbf{B}^{\natural}))$ to be $\Omega(n^{c_1/\mathrm{srank}(\mathbf{B}^{\natural})} p^{c_2 p/n})$. In the following context, we will show that the minimum SNR can be reduced to some positive constant provided multiple measurements are made $(m \gg 1)$.

3. Information Theoretic Lower Bounds

This section establishes the information-theoretic lower bounds for the correct permutation recovery. Our goal is to ensure both $\mathbf{\Pi}^{\natural}$ and $\mathbf{B}^{\natural}$ can be reliably reconstructed. We investigate this problem from two perspectives: (i) the sample number n and (ii) the minimum SNR.

3.1. The minimum sample number n

We obtain the minimum sample number n such that sparse signal $\mathbf{B}^{\natural}$ can be reliably recovered with high probability. Here we consider the oracle situation where $\mathbf{\Pi}^{\natural}$ is given a prior. As each column in $\mathbf{B}^{\natural}$ does not necessarily share the same support set, we need to iteratively reconstruct each column $\mathbf{B}^{\natural}_{:,\ell}$, a k-sparse signal, from the corresponding readings $\mathbf{Y}_{:,\ell}$ $(1 \leq \ell \leq m)$. With the classical result in Donoho (2006); Candes et al. (2006); Candès et al. (2006), we obtain the lower bound on n, namely, $n \gtrsim k \log p$. Naturally, we can expect this bound to apply to the non-oracle situation as well, since a reliable estimation of $\mathbf{B}^{\natural}$ is hopeless provided it is out of reach even in the oracle situation.

3.2. The minimum SNR

Then we turn to the minimum SNR requirement for the correct permutation recovery. To begin with, we restate the previous result in Zhang et al. (2019).

Theorem 1 (Theorem 1 in Zhang et al. (2019))**.** *Consider the oracle case where $\mathbf{B}^{\natural}$ is given a prior. Then, there exists an integer n_0 such that for an arbitrary estimator $\widehat{\mathbf{\Pi}}$, we have*

$$\inf_{\widehat{\mathbf{\Pi}}} \sup_{\mathbf{\Pi}^{\natural} \in \mathcal{P}_n} \mathbb{P}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^{\natural}) \geq \frac{1}{2},$$

provided that (i) $\log\det(\mathbf{I} + {\mathbf{B}^{\natural\top}\mathbf{B}^{\natural}}/{\sigma^2}) < \log n$ and (ii) $n \geq n_0$.

One drawback of this bound is the missing role of the sparsity number k. This is because Theorem 1 assumes $\mathbf{B}^{\natural}$ to be perfectly known while sparsity number k only kicks in when $\mathbf{B}^{\natural}$ needs to be reconstructed. To handle such an issue, we take $\mathrm{supp}(\mathbf{B}^{\natural})$ into account as well. Then we have

Theorem 2. *There exists an integer $n_0 \geq 0$ such that for arbitrary estimators $\widehat{\mathbf{\Pi}}$ and $\widehat{\mathbf{B}}$, we have*

$$\inf_{\widehat{\mathbf{\Pi}}, \widehat{\mathbf{B}}} \sup_{\substack{\mathbf{\Pi} \in \mathcal{P}_n, \\ \mathbf{B} \in \mathcal{B}_{n,p,m,k}}} \mathbb{P}_{\mathbf{X},\mathbf{W}}[(\widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\mathbf{B}})) \neq (\mathbf{\Pi}, \mathrm{supp}(\mathbf{B}))] \geq \frac{1}{2},$$

hold for all $n \geq n_0$, where $\mathcal{P}_n$ denotes the set of all permutation matrices, $\mathcal{B}_{n,p,m,k}$ is defined as the set reading as $\{\mathbf{B} \in \mathbb{R}^{p\times m} : \log\det(\mathbf{I} + {\mathbf{B}^{\top}\mathbf{B}}/{\sigma^2}) \leq \log n + \frac{m \log \binom{p}{k}}{n}\}$, and $\mathrm{supp}(\cdot) \triangleq \{(i,j) : (\cdot)_{i,j} \neq 0\}$ denotes the support set.

This theorem suggests that for all possible estimators to reconstruct $\mathbf{\Pi}$ and $\mathrm{supp}(\mathbf{B})$, there exists at least one pair $(\mathbf{\Pi}, \mathbf{B})$, $\mathbf{\Pi} \in \mathcal{P}_n, \mathbf{B} \in \mathcal{B}_{n,p,m,k}$ such that the reconstruction error rate will be at least ${1}/{2}$. Hence, a reliable correspondence reconstruction requires $\mathbf{B}^{\natural}$ (a fixed but unknown matrix) to satisfy $\log\det\left(\mathbf{I} + {\mathbf{B}^{\natural\top}\mathbf{B}^{\natural}}/{\sigma^2}\right) > \log n + {m\log\binom{p}{k}}/{n}$.

We leave its rigorous proof to the supplementary material and only give an intuitive interpretation, which comes from the coding theory. First, we assume each entry in $\mathbf{B}$ to be binary, i.e., $\mathbf{B}_{ij} \in \{0,1\}$, $(1 \leq i \leq p,\ 1 \leq j \leq m)$. Thus, the information of $\mathbf{B}$ is fully incorporated in $\mathrm{supp}(\mathbf{B})$. In the following, we will use $\mathrm{supp}(\mathbf{B})$ and $\mathbf{B}$ interchangeably, as they are identical.

Afterward, we view the sensing relation in (1) as the following transmission process: (i) pair $(\mathbf{\Pi}, \mathbf{B})$ is encoded into

[1] Experiments suggest that we may relax this assumption to that $\mathbf{X}_{ij}$ are i.i.d. sub-Gaussian random variables.

the codeword $\mathbf{\Pi}\mathbf{X}\mathbf{B}$; (ii) $\mathbf{\Pi}\mathbf{X}\mathbf{B}$ passes through a Gaussian additive channel; and (iii) one observes $\mathbf{Y}$, from which one would like to obtain $(\mathbf{\Pi}, \mathbf{B})$.

Using the terminology from coding theory, we can compute the corresponding code rate and channel capacity as $\frac{\log n! + m \log \binom{p}{k}}{n}$ and $\log\det(\mathbf{I} + {}^{\mathbf{B}^\top \mathbf{B}}/_{\sigma^2})$, respectively. Due to the Shannon theorem, we can expect non-negligible decoding error if the coding rate is greater than the channel capacity, which leads to $\log\det(\mathbf{I} + {}^{\mathbf{B}^\top \mathbf{B}}/_{\sigma^2}) < \log n + \frac{m \log \binom{p}{k}}{n}$, the formula in Theorem 2.

Remark 1. *Note that the minimum* SNR *requirements can be derived from Theorem 2 as* $\log\det(\mathbf{I} + {}^{\mathbf{B}^\top \mathbf{B}}/_{\sigma^2})$ *is closely related to* SNR. *When* $\mathbf{B}$ *is of rank-one, we have* $\log\det(\mathbf{I} + {}^{\mathbf{B}^\top \mathbf{B}}/_{\sigma^2})$ *be* $\log(1 + m \cdot \mathsf{SNR})$. *When* $\mathbf{B}$ *is of full rank and with identical singular values, we have* $\log\det(\mathbf{I} + {}^{\mathbf{B}^\top \mathbf{B}}/_{\sigma^2})$ *be* $\mathsf{rank}(\mathbf{B}) \cdot \log(1 + \mathsf{SNR})$.

Having obtained the information-theoretic lower bounds, we will propose a computationally efficient estimator, whose performance matches these lower bounds to a good extent.

4. Estimator Design

This section proposes a computationally efficient estimator for the permutation recovery. Denote $\mathsf{thres}(\cdot)$ as the operator which only keeps the element with the largest magnitude in each column, we reconstruct $\mathbf{\Pi}^\natural$ with the *linear assignment problem* (LAP) (Kuhn, 1955; Bertsekas and Castañón, 1992; Burkard et al., 2012) reading as

$$\mathbf{\Pi}^{\mathrm{opt}} = \mathrm{argmax}_{\mathbf{\Pi} \in \mathcal{P}_n} \left\langle \mathbf{\Pi}, \mathbf{Y} \cdot \mathsf{thres}(\mathbf{X}^\top \mathbf{Y})^\top \cdot \mathbf{X}^\top \right\rangle,$$

Once the permutation matrix $\mathbf{\Pi}^{\mathrm{opt}}$ is obtained, we can transform (1) to the previous setting and iteratively recover each k-sparse column $\mathbf{B}^\natural_{:,i}$. A formal statement of the algorithm is in Algorithm 1. Here, we use the Lasso estimator, which can be replaced with other estimators (say Dantzig estimator, etc), to reconstruct the signal $\mathbf{B}^\natural$.

Design intuition. The design of (2) shares a similar idea of the estimator in Zhang and Li (2020): we would like to approximate the direction of $\mathbf{B}^\natural$ by the product $\mathbf{X}^\top\mathbf{Y}$. However, due to insufficient samples, product $\mathbf{X}^\top\mathbf{Y}$ is poorly aligned with $\mathbf{B}^\natural$, or equivalently, large errors in approximating $\mathbf{B}^\natural$ with $\mathbf{X}^\top\mathbf{Y}$ and weak correlation $\langle \mathbf{B}^\natural, \mathbf{X}^\top\mathbf{Y}\rangle$. To reduce the approximation errors, we apply $\mathsf{thres}(\cdot)$ and set certain entries in $\mathbf{X}^\top\mathbf{Y}$ to zero.

Note that **we always keep one nonzero entry in the operation** $\mathsf{thres}(\cdot)$ **regardless of the sparsity number** k. This operation is different from almost all previous works, which range from Blumensath and Davies (2009); Foucart (2011) in the literature on compressed sensing to Jain et al. (2013); Yuan et al. (2014); Li et al. (2016) in the literature on optimization, since all these works suggest keeping at least $O(k)$ non-zero elements for a k-sparse signal.

Algorithm 1 One-step estimator.

Input: observation $\mathbf{Y}$ and sensing matrix $\mathbf{X}$.
Output: pair $(\mathbf{\Pi}^{\mathrm{opt}},\ \mathbf{B}^{\mathrm{opt}})$, which is written as

$$\mathbf{\Pi}^{\mathrm{opt}} = \mathrm{argmax}_{\mathbf{\Pi} \in \mathcal{P}_n} \left\langle \mathbf{\Pi}, \mathbf{Y} \cdot \mathsf{thres}(\mathbf{X}^\top \mathbf{Y})^\top \cdot \mathbf{X}^\top \right\rangle, \tag{2}$$

$$\mathbf{B}^{\mathrm{opt}} = \mathrm{argmin}_{\mathbf{B}} (2n)^{-1} \left\| \mathbf{\Pi}^{\mathrm{opt}\top}\mathbf{Y} - \mathbf{X}\mathbf{B} \right\|_{\mathrm{F}}^2 + \lambda_n \|\mathbf{B}\|_1, \tag{3}$$

where $\mathsf{thres}(\cdot)$ applies to each column and thresholds all entries to zero except the one with the largest magnitude, $\mathcal{P}_n$ denotes the set of all possible permutation matrices, $\|\cdot\|_1 \triangleq \sum_{i,j} |(\cdot)_{i,j}|$ denotes the absolute sum of all entries, and $\lambda_n > 0$ is some regularizer coefficient.

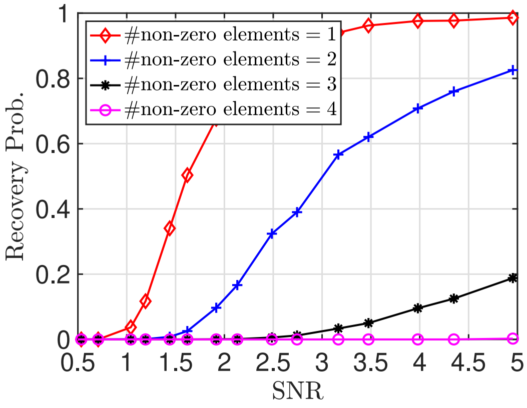

Figure 1. Keeping more non-zero elements in $\mathbf{X}^\top\mathbf{Y}$ deteriorates permutation recovery. $n = 100$, $p = 500$, $h = 25$, $k = 5$, $\mathsf{srank}(\mathbf{B}^\natural) = 100$. **#non-zero elements** refers to the number of non-zero elements kept in each column of $\mathbf{X}^\top\mathbf{Y}$.

More surprisingly, our numerical experiments suggest keeping more non-zero elements can be harmful to permutation recovery. An illustration is put in Figure 1, from which we observe the SNR required for correct correspondence recovery increases with the number of non-zero elements kept in $\mathbf{X}^\top\mathbf{Y}$.

An intuitive explanation of the operator $\mathsf{thres}(\cdot)$'s role is given as follows. Roughly speaking, we apply the operator $\mathsf{thres}(\cdot)$ to $\mathbf{X}^\top\mathbf{Y}$ for a better approximation of $\mathbf{B}^\natural$'s direction, or equivalently, increase the correlation $\langle \mathsf{thres}(\mathbf{X}^\top\mathbf{Y}), \mathbf{B}^\natural\rangle$. Due to the insufficient sample number n, $\mathbf{X}^\top\mathbf{Y}$ is poorly aligned with $\mathbf{B}^\natural$. Thus, keeping more non-zero elements in $\mathsf{thres}(\cdot)$ leads to a potential decrease of $\langle \mathsf{thres}(\mathbf{X}^\top\mathbf{Y}), \mathbf{B}^\natural\rangle$ and less satisfying performance.

4.1. Main results

This subsection formally presents our main results.

4.1.1. Results in recovering $\mathbf{\Pi}^\natural$

First, we study correspondence recovery. A formal statement is given as

Theorem 3. *Suppose that $n \geq n_0$ and $p \geq p_0$, where $n_0, p_0 > 0$ are some positive constants. Provided that (i) $n \gg k(\log n)(\log^2 mnp)$, (ii) $\mathsf{srank}(\mathbf{B}^\natural) \gg k^2 \log^{2(1+\varepsilon_0)} n$, (iii) $h \leq c_0 \cdot n$, and (iv) $\mathsf{SNR} \geq c_1$, we have $\{\mathbf{\Pi}^{\text{opt}} = \mathbf{\Pi}^\natural\}$ with probability at least $1 - c_2 \cdot n^{-c_3}$, where $\varepsilon_0 > 0$ is an arbitrary positive constant and $h \triangleq \mathsf{d_H}(\mathbf{I}, \mathbf{\Pi}^\natural)$ denotes the number of permuted rows.*

Discussion. Table 1 compares our result with the previous works. Our work gives the first affirmative answer such that $\mathsf{SNR} \geq \Omega(1)$ is sufficient to obtain the correct permutation matrix with insufficient samples, i.e., $n \ll p$.

In addition, we would like to compare it with the lower bound. To begin with, we discuss the SNR requirement, which is the top priority of our analysis. From Theorem 3, we can see that the correct permutation matrix can be obtained provided that SNR is above some positive constant; meanwhile Theorem 2 requires $\mathsf{SNR} > 0$. This means our SNR requirement has at most a gap of some positive constant with the statistical lower bound.

For the sample number n, the lower bound requires n to be at least of order $\Omega(k \log p)$; while Theorem 3 requires n to be $\Omega(k(\log n)(\log^2 mnp))$, which means the lower bound is matched up to some multiplicative logarithmic terms. We conjecture that the required n in Theorem 3 can be further optimized with a more delicate analysis.

Moreover, our estimator allows the maximum number of permuted rows to be linearly proportional to the sample number, i.e., $h_{\text{max}} \asymp n$, which is order-optimal.

In the end, we would like to discuss the assumption on $\mathsf{srank}(\mathbf{B}^\natural)$. When the eigenvalues of $\mathbf{B}^\natural$ is almost uniformly distributed, we have $\mathsf{srank}(\mathbf{B}^\natural) \asymp \mathsf{rank}(\mathbf{B}^\natural)$, which is of order $O(\min(p, m))$. Thus, we conclude that the constraint $\mathsf{srank}(\mathbf{B}^\natural) \gg k^2 \log^{2(1+\varepsilon_0)} n$ will reduce to the following requirement, i.e.,

$$\min(p, m) \gg k^2 \log^{2(1+\varepsilon_0)} n. \tag{4}$$

Notice the assumption $p \gg n$, we conclude that the first requirement in (4), i.e., $p \gg k^2 \log^{2(1+\varepsilon_0)} n$, is not very strict. For the second constraint, $m \gg k^2 \log^{2(1+\varepsilon_0)} n$, it suggests more multiple measurements are desired. [2] In addition, if we assume that for each column its maximum entry's energy is at least a constant proportion of the total energy, i.e., $\inf_j {}^{\max_i |\mathbf{B}^\natural_{i,j}|}/_{\|\mathbf{B}^\natural_{:,j}\|_2} \geq \varepsilon_1$ $(1 \leq i \leq p, 1 \leq j \leq m$, we can further relax the requirements on stable rank $\mathsf{srank}(\mathbf{B}^\natural)$, to put it more specifically, $\mathsf{srank}(\mathbf{B}^\natural) \gg \log^2 n$. Here $\varepsilon_1 > 0$ is an arbitrarily small positive constant. A formal statement is given as

Corollary 1. *Consider the large-system limit where n and p are sufficiently large. Under the conditions (i) $\inf_j {}^{\max_i |\mathbf{B}^\natural_{i,j}|}/_{\|\mathbf{B}^\natural_{:,j}\|_2} \geq \varepsilon_1$, (ii) $n \gg k(\log n)(\log^2 mnp)$, (iii) $\mathsf{srank}(\mathbf{B}^\natural) \gg \log^{2(1+\varepsilon_0)} n$, (iv) $h \leq c_0 \cdot n$, and (v) $\mathsf{SNR} \geq c_1$, we have $\{\mathbf{\Pi}^{\text{opt}} = \mathbf{\Pi}^\natural\}$ with probability at least $1 - c_2 \cdot n^{-c_3}$, where $\varepsilon_0, \varepsilon_1 > 0$ are some arbitrarily small positive constants; and $h \triangleq \mathsf{d_H}(\mathbf{I}, \mathbf{\Pi}^\natural)$ is the number of permuted rows.*

In this scenario, the requirements in (4) reduce to

$$p \gg \log^2 n \ \text{ and } \ m \gg \log^2 n.$$

Notice that the first requirement is almost trivially satisfied, as $p \gg n \gg \log^2 n$.

Remark 2. *Compared with Zhang and Li (2020) which only requires $\mathsf{srank}(\mathbf{B}^\natural)$ to be above certain positive constant, our estimator needs a larger stable rank $\mathsf{srank}(\mathbf{B}^\natural)$. Although we cannot claim that $\mathsf{srank}(\mathbf{B}^\natural)$ must be lower bounded by some non-decreasing functions of $\log n$, we have a numerical evidence such that $\mathsf{srank}(\mathbf{B}^\natural)$ may have to increase with sample number n, in other words, its lower bound may not be reduced to be some positive constant. Fixing the parameters p, k, h, and $\mathsf{srank}(\mathbf{B}^\natural)$, we study the impact of sample number n on the permutation recovery and put the results in Figure 2. We observe that a larger n has a negative impact on the permutation reconstruction once n exceeds a certain threshold. One possible reason is that the stable rank $\mathsf{srank}(\mathbf{B}^\natural)$ is fixed as a constant and violates the requirement $\mathsf{srank}(\mathbf{B}^\natural) \gg \log^2 n$.*

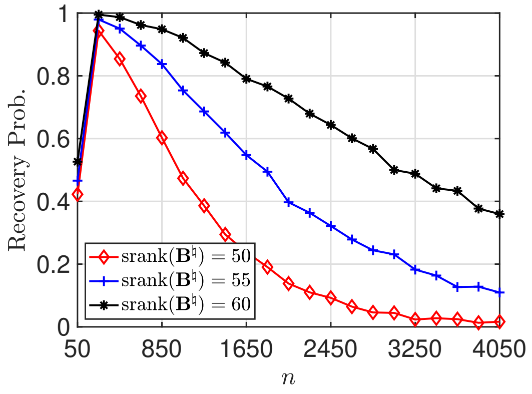

Figure 2. Illustration of the dual role of sample number n, for the noiseless case (infinite SNR) with $p = 200$ (signal length), $k = 5$ (sparsity number), and $h = 25$ (number of permuted rows).

[2] Specifically, we should mention the stable rank assumption is almost independent of the sparsity assumption, in other words, it does not confine matrix $\mathbf{B}^\natural$ to be an extremely sparse matrix.

Table 1. Comparison with previous works. All results are presented in their best order, which only holds in certain regimes. Here, $\mathsf{SNR}_{\mathrm{min}}$, n_{min} and h_{max} denotes the minimum SNR required for correct permutation recovery, the minimum required sample number and maximum allowed number of permuted rows, respectively. We represent the matrix rank as $\mathrm{rank}(\cdot)$. In addition, we use **NA** to denote that the analysis is **not applied** to the corresponding work. Moreover, the logarithmic term is omitted in $\widetilde{\Omega}(\cdot)$.

	$\mathsf{SNR}_{\mathrm{min}}\ (\geq)$		${}^{n_{\mathrm{min}}}\!/_{p}\ (\geq)$		${}^{h_{\mathrm{max}}}\!/_{n}\ (\leq)$	
	$m=1$	$m \gg 1$	$m=1$	$m \gg 1$	$m=1$	$m \gg 1$
Pananjady et al. (2018)	$\widetilde{\Omega}(n^c)$	**NA**	$\widetilde{\Omega}(1)$	**NA**	$\widetilde{\Omega}(1)$	**NA**
Slawski and Ben-David (2019)	$\widetilde{\Omega}(n^c)$	**NA**	$\widetilde{\Omega}(1)$	**NA**	$\widetilde{\Omega}(\log^{-1} n)$	**NA**
Zhang et al. (2019)	**NA**	$\widetilde{\Omega}(1)$	**NA**	$\widetilde{\Omega}(1)$	**NA**	$\widetilde{\Omega}\left(\log^{-1} r(\mathbf{B}^\natural)\right)$
Slawski et al. (2020)	**NA**	$\widetilde{\Omega}(1)$	**NA**	$\widetilde{\Omega}(p)$	**NA**	$\widetilde{\Omega}(\log^{-1} n)$
Zhang and Li (2020)	$\widetilde{\Omega}(n^c)$	$\widetilde{\Omega}(1)$	$\widetilde{\Omega}(1)$	$\widetilde{\Omega}(\sqrt{p})$	$\widetilde{\Omega}(1)$	$\widetilde{\Omega}(1)$
Zhang and Li (2021)	$\widetilde{\Omega}(n^c)$	**NA**	$o(1)$	**NA**	$\widetilde{\Omega}(1)$	**NA**
Our Estimator	**NA**	$\widetilde{\Omega}(1)$	**NA**	$o(1)$	**NA**	$\widetilde{\Omega}(1)$

4.1.2. RESULTS IN RECOVERING $\mathbf{B}^\natural$

Once the ground truth $\mathbf{\Pi}^\natural$ is obtained, we can reduce (1) to the traditional model in compressed sensing/sparse recovery (Candès et al., 2006; Candes et al., 2006; Donoho, 2006; Wainwright, 2019) and reconstruct the signal $\mathbf{B}^\natural$ with a Lasso estimator. Concerning the reconstruction error $|||\mathbf{B}^\natural - \mathbf{B}^{\mathrm{opt}}|||_{\mathrm{F}}$, we have the following corollary.

Corollary 2. *Suppose that $n \geq n_0$ and $p \geq p_0$, where $n_0, p_0 > 0$ are some positive constants. Provided that (i) $n \gg k(\log n)(\log^2 mnp)$, (ii) $\mathsf{srank}(\mathbf{B}^\natural) \gg k^2 \log^{2(1+\varepsilon)} n$, (iii) $h \leq c_0 \cdot n$, and (iv) $\mathsf{SNR} \geq c_1$. Setting λ_n in (3) as $c_2\sigma\sqrt{\log p/n}$, we conclude*

$$|||\mathbf{B}^\natural - \mathbf{B}^{\mathrm{opt}}|||_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{mk\log p}{n}}$$

holds with probability exceeding $1 - c_3 \cdot n^{-c_4} - c_5 \cdot p^{-c_6}$.

Its proof is a simple combination of Theorem 3 and the previous results in Candès et al. (2006); Candes et al. (2006); Donoho (2006); Wainwright (2019). However, we observe a new phenomenon: the reconstruction error $|||\mathbf{B}^{\mathrm{opt}} - \mathbf{B}^\natural|||_{\mathrm{F}}$ is affected by the signal energy $|||\mathbf{B}^\natural|||_{\mathrm{F}}$ on top of the sensing noise σ^2. [3] Moreover, we should mention that the above difference will still exist even when we replace (3) with other estimators, say, Dantzig estimator.

At last, we can relax the constraints on $\mathsf{srank}(\mathbf{B}^\natural)$ and n with an additional constraint on $\inf_j {}^{\max_i |\mathbf{B}^\natural_{i,j}|}\!/_{\|\mathbf{B}^\natural_{:,j}\|_2}$, which is the same as that in Corollary 1. A formal statement is given as the following.

Corollary 3. *Suppose that $n \geq n_0$ and $p \geq p_0$, where $n_0, p_0 > 0$ are some positive constants. Provided that (i) $\inf_j {}^{\max_i |\mathbf{B}^\natural_{i,j}|}\!/_{\|\mathbf{B}^\natural_{:,j}\|_2} \geq \varepsilon_1$, (ii) $n \gg k(\log n)(\log^2 mnp)$, (iii) $\mathsf{srank}(\mathbf{B}^\natural) \gg \log^{2(1+\varepsilon_0)} n$, (iv) $h \leq c_0 \cdot n$, and (v) $\mathsf{SNR} \geq c_1$. Setting λ_n in (3) as $c_2\sigma\sqrt{\log p/n}$, we conclude*

$$|||\mathbf{B}^\natural - \mathbf{B}^{\mathrm{opt}}|||_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{mk\log p}{n}}$$

holds with probability at least $1 - c_3 \cdot n^{-c_4} - c_5 \cdot p^{-c_6}$.

4.2. Proof outline

Due to the space limit, we only give a sketch of our proof ideas and leave the technical details to the supplementary material. Denote $\widehat{\mathbf{B}} = (n-h)^{-1}\mathbf{X}^\top\mathbf{Y}$, our goal is to show

$$\langle \mathbf{Y}, \mathbf{\Pi}^\natural\mathbf{X} \cdot \mathsf{thres}(\widehat{\mathbf{B}})\rangle > \langle \mathbf{Y}, \mathbf{\Pi}\mathbf{X} \cdot \mathsf{thres}(\widehat{\mathbf{B}})\rangle, \quad \forall\, \mathbf{\Pi} \neq \mathbf{\Pi}^\natural \tag{5}$$

holds with high probability under the settings in Theorem 3.

Same as Zhang and Li (2020), our analysis faces the difficulties brought by (i) combinatorial nature of the problem and (ii) high-order moments of sub-Gaussian random variables. On top of these challenges, we are subject to insufficient samples, i.e., $n \ll p$. These issues are tackled by a combination of relaxation and tailored leave-one-out analysis, which can be roughly divided into the following three stages.

Stage I. We consider the sufficient condition of (5), which can be written as

[3] This corollary has one requirement on SNR, which is affected by both $|||\mathbf{B}^\natural|||_{\mathrm{F}}$ and σ^2. In addition, we may have $|||\mathbf{B}^{\mathrm{opt}} - \mathbf{B}^\natural|||_{\mathrm{F}}$ be directly linked to $|||\mathbf{B}^\natural|||_{\mathrm{F}}$, as a wrong correspondence can be obtained with low SNR.

$$\langle \mathbf{Y}_{i,:}, \mathsf{thres}(\widehat{\mathbf{B}})^{\top}\mathbf{X}_{\pi^{\natural}(i),:}\rangle \geq \langle \mathbf{Y}_{i,:}, \mathsf{thres}(\widehat{\mathbf{B}})^{\top}\mathbf{X}_{j,:}\rangle, \quad \forall\ j \neq \pi^{\natural}(i).$$

Re-arranging the terms, we obtain an equivalent form reading as

$$\begin{aligned} &\langle \mathbf{B}^{\natural\top}\mathbf{X}_{\pi^{\natural}(i),:}, \mathsf{thres}(\widehat{\mathbf{B}})^{\top}\mathbf{X}_{\pi^{\natural}(i),:}\rangle \\ \geq\ &\langle \mathbf{B}^{\natural\top}\mathbf{X}_{\pi^{\natural}(i),:}, \mathsf{thres}(\widehat{\mathbf{B}})^{\top}\mathbf{X}_{j,:}\rangle \\ &+ \langle \mathbf{W}_{i,:}, \mathsf{thres}(\widehat{\mathbf{B}})^{\top}\left(\mathbf{X}_{\pi^{\natural}(i),:} - \mathbf{X}_{j,:}\right)\rangle. \end{aligned} \tag{6}$$

Informally speaking, we first assume that $\mathsf{thres}(\widehat{\mathbf{B}})$ is almost parallel to $\mathbf{B}^{\natural}$; and the dependence of $\mathsf{thres}(\widehat{\mathbf{B}})$ on $\mathbf{X}_{\pi^{\natural}(i),:}, \mathbf{X}_{j,:}$ is negligible. Then, we can approximate (6)'s left-hand side as $\left|\!\left|\!\left|\mathbf{B}^{\natural}\right|\!\right|\!\right|_{\mathrm{F}}^{2}$ and its right-hand side as a sum of $\boldsymbol{z}_1\mathbf{B}^{\natural\top}\mathbf{B}^{\natural}\boldsymbol{z}_2$ and $\sqrt{2}\boldsymbol{z}\left|\!\left|\!\left|\mathbf{B}^{\natural}\right|\!\right|\!\right|_{\mathrm{F}}$ up to some normalization constant, where $\boldsymbol{z}_1, \boldsymbol{z}_2 \overset{\text{i.i.d}}{\sim} \mathsf{N}(\mathbf{0}, \mathbf{I}_{p\times p})$ and $\boldsymbol{z} \overset{\text{i.i.d}}{\sim} \mathsf{N}(\mathbf{0}, \mathbf{I}_{m\times m})$ are Gaussian random vectors. Easily, we can see that (6) holds with high probability provided the SNR is sufficiently large. In the following two stages, we will verify the above two assumptions, that is, (i) $\angle(\mathsf{thres}(\widehat{\mathbf{B}}), \mathbf{B}^{\natural})$ is small and (ii) the dependence between $\mathsf{thres}(\widehat{\mathbf{B}})$ and $\mathbf{X}_{\pi^{\natural}(i),:}, \mathbf{X}_{j,:}$ is negligible.

Stage II. We would like to lower-bound the inner product $\langle \mathbf{B}^{\natural}, \mathsf{thres}(\widehat{\mathbf{B}})\rangle$. Denote $\widehat{\boldsymbol{\beta}}$ as the corresponding column in $\widehat{\mathbf{B}}$, we can express $\langle \mathbf{B}^{\natural}, \mathsf{thres}(\widehat{\mathbf{B}})\rangle$ as $\sum_{\boldsymbol{\beta}^{\natural}\in\{\mathbf{B}^{\natural}_{:,\ell}\}_{1\leq\ell\leq m}}\langle \boldsymbol{\beta}^{\natural}, \mathsf{thres}(\widehat{\boldsymbol{\beta}})\rangle$. From the definition of $\mathsf{thres}(\cdot)$, we notice that $\mathsf{thres}(\widehat{\boldsymbol{\beta}})$ only has one non-zero entry. W.l.o.g. we assume its index is one and hence have

$$\begin{aligned} \langle \boldsymbol{\beta}^{\natural}, \mathsf{thres}(\widehat{\boldsymbol{\beta}})\rangle =\ &\boldsymbol{\beta}^{\natural}_1\widehat{\boldsymbol{\beta}}_1 = (\boldsymbol{\beta}^{\natural}_1)^2 + \boldsymbol{\beta}^{\natural}_1(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^{\natural}_1) \\ \geq\ &(\boldsymbol{\beta}^{\natural}_1)^2 - \max_i|\boldsymbol{\beta}^{\natural}_i| \cdot \|\boldsymbol{\beta}^{\natural} - \widehat{\boldsymbol{\beta}}\|_{\infty}. \end{aligned} \tag{7}$$

We (i) upper-bound $\max_i|\boldsymbol{\beta}^{\natural}_i|$ and $\|\boldsymbol{\beta}^{\natural} - \widehat{\boldsymbol{\beta}}\|_{\infty}$; and (ii) lower-bound $|\boldsymbol{\beta}^{\natural}_1|$. Part (i) is quite standard and part (ii) lies in analyzing the event $|\widehat{\boldsymbol{\beta}}_1| \geq \max_j|\widehat{\boldsymbol{\beta}}_j|$, which is due to the definition of $\mathsf{thres}(\cdot)$.

Stage III. We would like to show the dependence between $\mathsf{thres}(\widehat{\mathbf{B}})$ and rows $\mathbf{X}_{\pi^{\natural}(i),:}$ and $\mathbf{X}_{j,:}$ is negligible. This is accomplished by a tailored leave-one-out analysis. For each row indices pair $(\pi^{\natural}(i), j)$, we construct a perturbed matrix $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$ by replacing the rows $\mathbf{X}_{\pi^{\natural}(i),:}, \mathbf{X}_{j,:}$ with their i.i.d. substitutes. A more detailed description of the construction process is given as follows.

First, we draw i.i.d. samples for each rows of $\mathbf{X}_{i,:}$ and denote it as $\widetilde{\mathbf{X}}_{i,:}$ $(1 \leq i \leq n)$. Similarly we draw samples $\widetilde{\mathbf{W}}_{j,:}$ for each row in $\mathbf{W}$, $1 \leq j \leq n$. For arbitrary indices $\pi^{\natural}(i)$ and j such that $j \neq \pi^{\natural}(i)$, we create the samples $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$ as (assume $i \neq \pi^{\natural}(i)$ and $j \neq \pi^{\natural}(j)$)

$$\begin{aligned} \widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)} = (n-h)^{-1}&\Bigg(\sum_{\ell,\pi^{\natural}(\ell)\neq\pi^{\natural}(i),j}\mathbf{X}_{\ell,:}\mathbf{X}^{\top}_{\pi^{\natural}(\ell),:}\Bigg)\mathbf{B}^{\natural} \\ &+ (n-h)^{-1}\Big(\widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{X}}^{\top}_{\pi^{\natural}(i),:} + \widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}\widetilde{\mathbf{X}}^{\top}_{\pi^{\natural}(\pi^{\natural}(i)),:})\mathbf{B}^{\natural} \\ &+ (n-h)^{-1}\Big(\widetilde{\mathbf{X}}_{j,:}\widetilde{\mathbf{X}}^{\top}_{\pi^{\natural}(j),:} + \widetilde{\mathbf{X}}_{\pi^{\natural-1}(j),:}\widetilde{\mathbf{X}}^{\top}_{j,:}\Big)\mathbf{B}^{\natural} \\ &+ (n-h)^{-1}\sum_{\ell\neq\pi^{\natural}(i),i}\mathbf{X}_{\ell,:}\mathbf{W}^{\top}_{\ell,:} \\ &+ (n-h)^{-1}\Big(\widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}\widetilde{\mathbf{W}}^{\top}_{\pi^{\natural}(i),:} + \widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{W}}^{\top}_{i,:}\Big). \end{aligned}$$

Provided that $i = \pi^{\natural}(i)$, we can simplify the summaries $\widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{X}}^{\top}_{\pi^{\natural}(i),:} + \widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}\widetilde{\mathbf{X}}^{\top}_{\pi^{\natural}(\pi^{\natural}(i)),:}$ and $\widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}\widetilde{\mathbf{W}}^{\top}_{\pi^{\natural}(i),:} + \widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{W}}^{\top}_{i,:}$ in the above construction as the terms $\widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{X}}^{\top}_{i,:}$ and $\widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{W}}^{\top}_{i,:}$, respectively. Similarly, we will simplify $\widetilde{\mathbf{X}}_{j,:}\widetilde{\mathbf{X}}^{\top}_{\pi^{\natural}(j),:} + \widetilde{\mathbf{X}}_{\pi^{\natural-1}(j),:}\widetilde{\mathbf{X}}^{\top}_{j,:}$ as $\mathbf{X}_{j,:}\mathbf{X}^{\top}_{j,:}$ when $j = \pi^{\natural}(j)$.

Easily, we can verify that $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$ is independent from the rows $\mathbf{X}_{\pi^{\natural}(i),:}, \mathbf{X}_{j,:}$ as the latter are not involved in $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$. Meanwhile, we have $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$ exhibit a similar behavior as $\widehat{\mathbf{B}}$ as they share almost identical components. In fact, this is the basic idea of leave-one-out technique (Chen et al., 2020; Sur et al., 2019; Karoui, 2013; 2018; Cai et al., 2021). Compared with these works, our construction method has the following characteristics

- The number of replaced rows in our method varies for different pairs of row indices. As we can see from the above example, we need to substitute the rows $\mathbf{X}_{\pi^{\natural}(i),:}$ and $\mathbf{X}_{j,:}$ in $\widehat{\mathbf{B}}$ with their i.i.d. replicas. Since the number of impacted terms depends on the relations between index $\pi^{\natural}(i)$ and index j, we need to replace a varying number of terms in the perturbed matrix $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$. Meanwhile, the number of replaced terms is fixed in the works thereof.

- We modify the operator $\mathsf{thres}(\cdot)$ in approximating $\mathsf{thres}(\widehat{\mathbf{B}})$. While the previous works usually keep the operator $\mathsf{thres}(\cdot)$ intact, we approximate it with the operator $(\cdot)_{\mathsf{imax}}$, which denotes the positions of non-zero elements in $\mathsf{thres}(\widehat{\mathbf{B}})$. In other words, the positions of the non-zero elements we keep in the *leave-one-out* samples $\widehat{\mathbf{B}}^{(\cdot)}$ are determined by $\mathsf{thres}(\widehat{\mathbf{B}})$ rather than $\mathsf{thres}(\widehat{\mathbf{B}}^{(\cdot)})$. In our analysis, we can see this step is essential in controlling the approximation errors. Otherwise, the approximation error can be considerably large, since $\mathsf{thres}(\widehat{\mathbf{B}})$ may not share the same support set with $\mathsf{thres}(\widehat{\mathbf{B}}^{(\cdot)})$, let alone their ℓ_2 differences.

Note that the above explanation is a simplified version of our proof technique. The technical details, which are deferred to

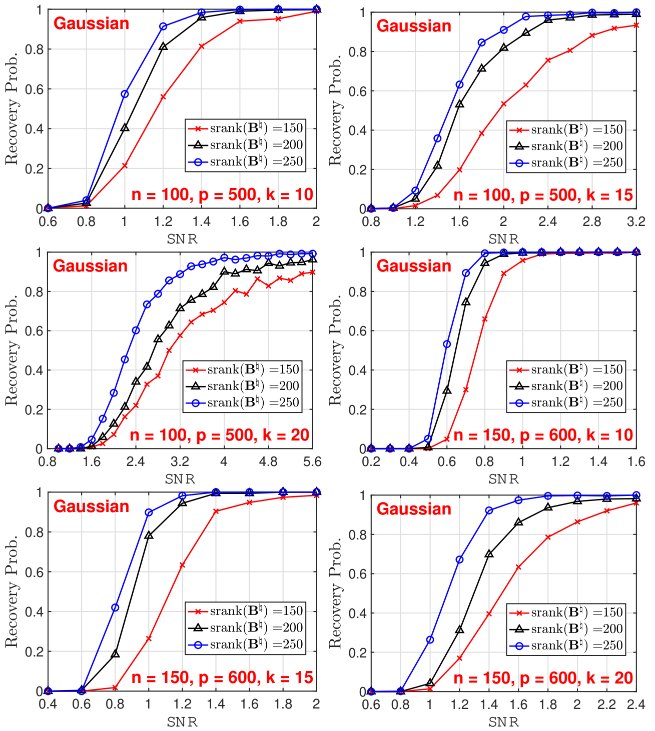

Figure 3. Simulated recovery rate $\mathbb{P}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^\natural)$ with $n = \{100, 150\}$, $p \in \{500, 600\}$, $h \in \{25, 37\}$, and $\mathbf{X}_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{N}(0,1)$, w.r.t. SNR.

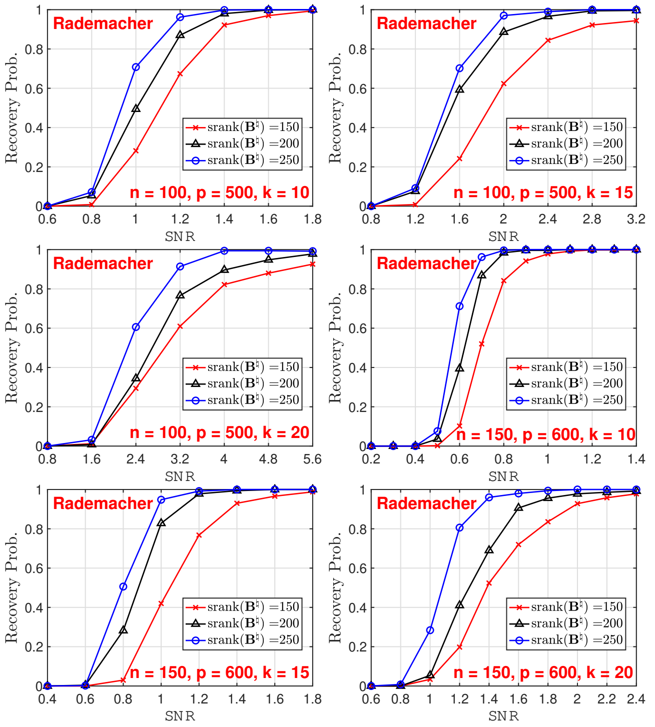

Figure 4. Simulated recovery rate $\mathbb{P}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^\natural)$ with $n = \{100, 150\}$, $p \in \{500, 600\}$, $h \in \{25, 37\}$, and $\mathbf{X}_{ij} \overset{\text{i.i.d}}{\sim}$ Rademacher $\mathbb{P}(\mathbf{X}_{ij} = -1) = \mathbb{P}(\mathbf{X}_{ij} = 1) = {}^1\!/_2$, w.r.t. SNR.

the supplementary material, can be different from the above description but do follow the same spirit.

Moreover, we would like to briefly discuss our algorithm's computational complexity: in the first step for permutation recovery, our estimator only requires one matrix multiplication and thresholding operation on top of the operations in the oracle estimator; in the second step for sparse signal recovery, our estimator needs one additional matrix multiplication when compared with the work without permutation.

5. Numerical Results

This section presents the numerical experiments to verify our main theorem, to put it more specifically, Theorem 3: we would like to prove the correct permutation can be obtained, i.e., $\{\mathbf{\Pi}^{\text{opt}} = \mathbf{\Pi}^\natural\}$, with $n \ll p$ and SNR $\geq c$.

5.1. Numerical results on synthetic data

To begin with, we present the numerical results with synthetic data. First, we adopt the Gaussian sensing matrix $\mathbf{X}$, i.e., $\mathbf{X}_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{N}(0,1)$, which is the same setting of Theorem 3. Then, we separately investigate the impact of the sampling ratio ${}^n\!/_p$ and the sparsity number k on the permutation recovery. Second, we extend our sensing matrix setting from Gaussian to sub-Gaussian. Assuming $\mathbf{X}_{ij}$ to be i.i.d. Rademacher random variable, namely, $\mathbb{P}(\mathbf{X}_{ij} = \pm 1) = {}^1\!/_2$, we show that our reconstruction algorithm in Algorithm 1 can successfully reconstruct the permutation matrix under the sub-Gaussian setting as well.

Experiment setting with Gaussian distribution. We let $\mathbf{X}_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{N}(0,1)$ and pick the sample number n to be $\{100, 150\}$ and set $h = {}^n\!/_4$. We vary the signal length p to be $\{500, 600\}$. Then we set the sparsity number k within the region $\{10, 15, 20\}$. And the stable rank srank$(\mathbf{B}^\natural)$ is within the range $\{150, 200, 250\}$. The corresponding simulation results can be found in Figure 3.

Discussion w.r.t ${}^n\!/_p$. First, we confirm our theory such that correct permutation can be obtained with insufficient samples, i.e., $n \ll p$. In addition, we notice that the permutation recovery becomes easy with a larger ${}^n\!/_p$: the first row in Figure 3 is with ${}^n\!/_p = 0.2$ while the second row is with ${}^n\!/_p = 0.25$. We can verify the SNR for the correct permutation recovery is smaller in the second row than that for the first row. However, we should stress that this conclusion may not hold provided that srank$(\mathbf{B}^\natural)$ is not sufficiently large. More details are referred to Figure 2.

Discussion w.r.t. sparsity number k. We vary the sparsity number k to be within $\{10, 15, 20\}$. We conclude a

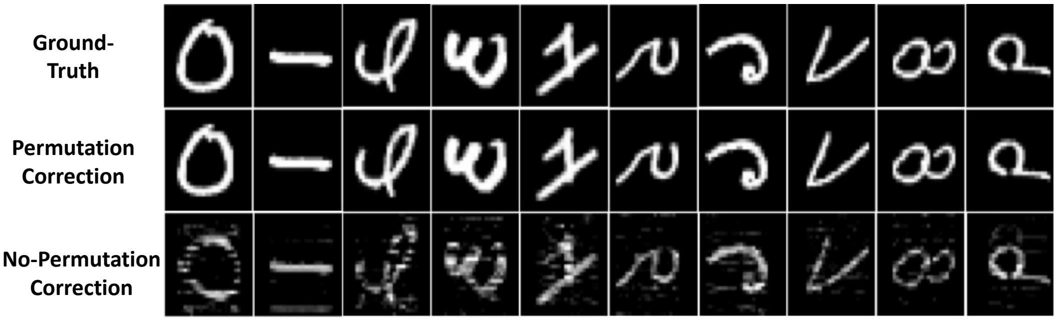

Figure 5. We set the sample number n as 100 and the sensing noise variance σ^2 as 1. (**Top panel.**) The ground-truth images. (**Middle panel.**) The reconstructed images with our algorithm using both (2) and (3). (**Bottom panel.**) The reconstructed images with only (3), where $\mathbf{\Pi}^{\mathrm{opt}}$ is set to be the identity matrix.

large sparsity number k can make the permutation recovery more difficult. For example, consider the case when $(n, p, \mathsf{srank}(\mathbf{B}^\natural)) = (100, 500, 200)$. When $k = 10$, correct permutation requires $\mathsf{SNR} \geq 1.4$; when $k = 15$, correct permutation needs $\mathsf{SNR} \geq 2.2$; and when $k = 20$, correct permutation requires $\mathsf{SNR} \geq 4$. The same conclusion holds for other cases as well. [4]

Experiment setting with the sub-Gaussian distribution. In addition to the Gaussian setting, we also evaluate our estimator's performance when $\mathbf{X}_{ij}$ is sub-Gaussian. Here, we pick $\mathbf{X}_{ij}$ to be i.i.d. Rademacher random variables such that $\mathbb{P}(\mathbf{X}_{ij} = \pm 1) = 1/2$. The corresponding results are put in Figure 4, from which we can observe a similar pattern as that of the Gaussian setting.

5.2. Numerical results with real-world data

This subsection evaluates our algorithm on the real-world dataset, namely, MNIST dataset (LeCun et al., 1998). An illustration of the reconstructed images is put in Figure 5.

First, we create the sparse matrix $\mathbf{B}^\natural$ as a block-diagonal matrix $\operatorname{diag}\left(\mathrm{image}_0^{\mathrm{train}}, \mathrm{image}_0^{\mathrm{test}}, \mathrm{image}_1^{\mathrm{train}}, \cdots \mathrm{image}_9^{\mathrm{test}}\right)$, where $\mathrm{image}_i^{\mathrm{train}} \in \mathbb{R}^{28\times 28}$ (resp. $\mathrm{image}_i^{\mathrm{test}} \in \mathbb{R}^{28\times 28}$) represents an arbitrary image of digit i $(0 \leq i \leq 9)$ in the training (resp. test) set. These images under the testing test are put in the "**Ground-Truth**" panel.

Then, we create a Gaussian sensing matrix $\mathbf{X}$ and a noise matrix $\mathbf{W}$. Afterward, we permute the sensing results and apply our algorithm in Algorithm 1 to reconstruct the images. The reconstructed images are put in the panel under the name "**Permutation Correction**".

As the benchmark, we pretend there is no missing information of correspondence between $\mathbf{B}^\natural$ and $\mathbf{Y}$. We set $\mathbf{\Pi}^{\mathrm{opt}}$ in (3) be the identity matrix and reconstruct $\mathbf{B}^\natural$ with the Lasso estimator. Since there is no permutation recovery in the algorithm, we present the corresponding reconstructed images in the "**No-Permutation Correction**" panel. From Figure 5, we observe that our permutation recovery algorithm can significantly improve the reconstruction performance.

6. Conclusion

In this paper, we studied the unlabeled sparse recovery with multiple measurements (i.e., $m > 1$) for the first time. To begin with, we investigated the lower bounds on the sample number n and the SNR. Furthermore, we proposed a simple yet effective estimator, which restores the permutation matrix via a linear assignment problem. We proved that our estimator can obtain the correct correspondence information when SNR **is above certain positive constant** and the **required sample number n is in linear dependence with sparsity number k.** In addition, we discovered multiple phenomena that are seldom encountered before: (i) keeping more non-zero elements in $\mathsf{thres}(\cdot)$ deteriorates permutation recovery; and (ii) increasing sample number n plays a dual role in reconstructing the permutation. In the course of analyzing our estimator's performance and explaining the above phenomena, we developed a tailored version of the leave-one-out technique, which involves an adaptive number of replaced elements and simultaneous modification of the threshold operator. Moreover, we provided numerical experiments to corroborate our claims and showed our estimator can reliably reconstruct the permutation matrix even when the entries $\mathbf{X}_{ij}$ are sub-Gaussian random variables.

[4] The result is consistent with Theorem 2 as $\mathcal{B}_{n,p,m,k}$ covers a broader class of matrices $\mathbf{B}$ when k increases from 15 to 20, which implies a larger SNR is required for permutation recovery.

References

Abubakar Abid, Ada Poon, and James Zou. Linear regression with shuffled labels. *arXiv preprint arXiv:1705.01342*, 2017.

Zhidong Bai and Tailen Hsing. The broken sample problem. *Probability Theory and Related Fields*, 131(4):528–552, 2005.

Dimitri P. Bertsekas and David A. Castañón. A forward/reverse auction algorithm for asymmetric assignment problems. *Comput. Optim. Appl.*, 1(3):277–297, 1992.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment problems, revised reprint*, volume 106. SIAM, 2012.

Changxiao Cai, Gen Li, Yuejie Chi, H Vincent Poor, and Yuxin Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics*, 49(2):944–967, 2021.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

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

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2nd edition, 2012.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

Morris H. DeGroot and Prem K. Goel. The matching problem for multivariate normal data. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 38(1):14–29, 1976.

Morris H. DeGroot and Prem K. Goel. Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, 8(2):264–278, 03 1980.

David L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

Valentin Emiya, Antoine Bonnefoy, Laurent Daudet, and Rémi Gribonval. Compressed sensing with unknown sensor permutation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1040–1044, Florence, Italy, 2014.

Guanhua Fang and Ping Li. Regression with label permutation in generalized linear model. *arXiv preprint arXiv:2206.11775*, 2022.

Simon Foucart. Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM J. Numer. Anal.*, 49(6): 2543–2563, 2011.

Prem K. Goel. On re-pairing observations in a broken random sample. *The Annals of Statistics*, 3(6):1364–1369, 11 1975.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

Saeid Haghighatshoar and Giuseppe Caire. Signal recovery from unlabeled samples. *IEEE Trans. Signal Process.*, 66 (5):1242–1257, 2018.

Daniel J. Hsu, Kevin Shi, and Xiaorui Sun. Linear regression without correspondence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1531–1540, Long Beach, CA, 2017.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Symposium on Theory of Computing Conference (STOC)*, pages 665–674, Palo Alto, CA, 2013.

Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

Noureddine EL Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Jarvis D. Haupt. Stochastic variance reduced optimization for nonconvex sparse learning. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 917–925, New York City, NY, 2016.

Ashwin Pananjady, Martin J. Wainwright, and Thomas A. Courtade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Trans. Inf. Theory*, 64(5):3286–3300, 2018.

Liangzu Peng, Boshi Wang, and Manolis C. Tsakiris. Homomorphic sensing: Sparsity and noise. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8464–8475, Virtual Event, 2021.

Martin Slawski and Emanuel Ben-David. Linear regression with sparsely permuted data. *Electronic Journal of Statistics*, 1:1–36, 2019.

Martin Slawski, Mostafa Rahmani, and Ping Li. A sparse representation-based approach to linear regression with partially shuffled labels. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 38–48, Tel Aviv, Israel, 2019.

Martin Slawski, Emanuel Ben-David, and Ping Li. Two-stage approach to multivariate linear regression with sparsely mismatched data. *J. Mach. Learn. Res.*, 21(204): 1–42, 2020.

Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1):487–558, 2019.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230, 2015.

Manolis C. Tsakiris, Liangzu Peng, Aldo Conca, Laurent Kneip, Yuanming Shi, and Hayoung Choi. An algebraic-geometric approach for linear regression without correspondences. *IEEE Trans. Inf. Theory*, 66(8):5130–5144, 2020.

Jayakrishnan Unnikrishnan, Saeid Haghighatshoar, and Martin Vetterli. Unlabeled sensing: Solving a linear system with unordered measurements. In *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 786–793, Monticello, IL, 2015.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Xiaotong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pages 127–135, Beijing, China, 2014.

Hang Zhang and Ping Li. Optimal estimator for unlabeled linear regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 11153–11162, Virtual Event, 2020.

Hang Zhang and Ping Li. Sparse recovery with shuffled labels: Statistical limits and practical estimators. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 1760–1765, Melbourne, Australia, 2021.

Hang Zhang and Ping Li. Greed is good: correspondence recovery for unlabeled linear regression. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, Pittsburgh, PA, 2023.

Hang Zhang, Martin Slawski, and Ping Li. Permutation recovery from multiple measurement vectors in unlabeled sensing. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 1857–1861, Paris, France, 2019.

Hang Zhang, Martin Slawski, and Ping Li. The benefits of diversity: Permutation recovery in unlabeled sensing from multiple measurement vectors. *IEEE Trans. Inf. Theory*, 68(4):2509–2529, 2022.

Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Median-truncated nonconvex approach for phase retrieval with outliers. *IEEE Trans. Inf. Theory*, 64(11):7287–7310, 2018.

A. Proof of Theorem 2

Proof. The proof technique is a combination of that in Zhang et al. (2019) and Zhang and Li (2021). First, we put uniform distribution as the prior of $\mathbf{\Pi}$, i.e., $\mathbb{P}(\mathbf{\Pi}^{\natural} = \mathbf{\Pi}_{\text{samp}}) = |\mathcal{P}_n|^{-1}$, where $\mathbf{\Pi}_{\text{samp}}$ is an arbitrary fixed permutation matrix and $\mathcal{P}_n$ denotes the set of all possible permutation matrices. In addition, we introduce distributions for the support set of $\mathbf{B}$. For an arbitrary column $\mathbf{B}_{:,\ell}$, we assume its support set to be uniformly distributed among $\binom{p}{k}$ possible patterns. Easily, we can verify the relation

$$\sup_{\mathbf{\Pi},\mathbf{B}} \mathbb{P}_{\mathbf{X},\mathbf{W}}\left[(\mathbf{\Pi}, \text{supp}(\mathbf{B})) \neq (\widehat{\mathbf{\Pi}}, \text{supp}(\widehat{\mathbf{B}}))\right] \geq \mathbb{P}_{\mathbf{X},\mathbf{W},\mathbf{\Pi},\mathbf{B}}\left[(\mathbf{\Pi}, \text{supp}(\mathbf{B})) \neq (\widehat{\mathbf{\Pi}}, \text{supp}(\widehat{\mathbf{B}}))\right]. \tag{8}$$

Since $\inf_{\widehat{\mathbf{\Pi}},\widehat{\mathbf{B}}}$ can be safely added to the left-hand side of (8), our goal becomes lower-bounding $\mathbb{P}_{\mathbf{X},\mathbf{W},\mathbf{\Pi},\mathbf{B}}\left[(\mathbf{\Pi}, \text{supp}(\mathbf{B})) \neq (\widehat{\mathbf{\Pi}}, \text{supp}(\widehat{\mathbf{B}}))\right]$. Adopting the proof technique used in Theorem 2.10.1 in Cover and Thomas (2012), we consider the entropy $\mathsf{H}(\mathbf{\Pi}, \text{supp}(\mathbf{B}))$, which can be computed as

$$\mathsf{H}(\mathbf{\Pi}, \text{supp}(\mathbf{B})) \overset{\textcircled{1}}{=} \mathsf{H}(\mathbf{\Pi}) + \mathsf{H}(\text{supp}(\mathbf{B})) \overset{\textcircled{2}}{=} \log n! + m \cdot \log \binom{p}{k}, \tag{9}$$

where in ① we use the independent among $\mathbf{\Pi}$ and $\text{supp}(\mathbf{B})$, and in ② we use the fact $|\mathcal{P}_n| = n!$ and $|\text{supp}(\mathbf{B})| = \binom{p}{k}^m$. Meanwhile, we have the relation

$$\begin{aligned} \mathsf{H}(\mathbf{\Pi}, \text{supp}(\mathbf{B})) &\overset{\textcircled{3}}{=} \mathsf{H}(\mathbf{\Pi}, \text{supp}(\mathbf{B}) \mid \mathbf{X}) \\ &\overset{\textcircled{4}}{=} \underbrace{\mathsf{H}(\mathbf{\Pi}, \text{supp}(\mathbf{B}) \mid \mathbf{X}, \widehat{\mathbf{\Pi}}, \text{supp}(\widehat{\mathbf{B}}))}_{\triangleq \zeta_1} + \underbrace{\mathsf{I}(\mathbf{\Pi}, \text{supp}(\mathbf{B}); \widehat{\mathbf{\Pi}}, \text{supp}(\widehat{\mathbf{B}}) \mid \mathbf{X})}_{\triangleq \zeta_2}. \end{aligned} \tag{10}$$

where ③ is due to the independence between $\mathbf{X}$ and $\mathbf{\Pi}, \mathbf{B}$; and ④ is because of the definition of conditional entropy and mutual information. The proof is thus complete by separately bounding η_1 and η_2.

Analysis of ζ_1. We upper-bound ζ_1 with Fano's inequality (Cover and Thomas, 2012, Theorem 2.10.1), which proceeds as

$$\zeta_1 \leq \mathsf{H}(\mathbf{\Pi}, \text{supp}(\mathbf{B}) \mid \widehat{\mathbf{\Pi}}, \text{supp}(\widehat{\mathbf{B}})) \leq 1 + \log\left(|(\mathbf{\Pi}, \text{supp}(\mathbf{B}))|\right) \cdot \mathbb{P}_{\mathbf{X},\mathbf{W},\mathbf{\Pi},\mathbf{B}}\left[(\mathbf{\Pi}, \text{supp}(\mathbf{B})) \neq (\widehat{\mathbf{\Pi}}, \text{supp}(\widehat{\mathbf{B}}))\right]. \tag{11}$$

Analysis of ζ_2. Due to the Markov property of $(\mathbf{\Pi}, \text{supp}(\mathbf{B})) \to \mathbf{Y} \to (\widehat{\mathbf{\Pi}}, \text{supp}(\widehat{\mathbf{B}}))$, we invoke the data-processing inequality (Cover and Thomas, 2012, Theorem 2.8.1) and conclude

$$\zeta_2 \leq \mathsf{I}(\mathbf{\Pi}, \text{supp}(\mathbf{B}); \mathbf{Y} \mid \mathbf{X}).$$

Invoking the definition of conditional mutual information, we have

$$\begin{aligned} \mathsf{I}(\mathbf{\Pi}, \text{supp}(\mathbf{B}); \mathbf{Y} \mid \mathbf{X}) &= \mathbb{E}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}}\left[\mathsf{h}(\mathbf{Y} \mid \mathbf{X} = \boldsymbol{x})) - \mathsf{h}(\mathbf{Y} \mid \mathbf{\Pi}, \text{supp}(\mathbf{B}), \mathbf{X} = \boldsymbol{x})\right] \\ &\overset{\textcircled{5}}{\leq} \frac{1}{2} \log\det\left(\mathbb{E}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}} \mathbf{Y}\mathbf{Y}^{\top}\right) - \frac{mn}{2} \log \sigma^2, \end{aligned}$$

where in ⑤ we use the property (Cover and Thomas, 2012, Theorem 8.6.5)

$$\mathsf{h}(\mathbf{Z}) \leq \frac{1}{2} \log\det \text{Cov}(\mathbf{Z}) \leq \frac{1}{2} \log\det \mathbb{E}\left(\mathbf{Z}\mathbf{Z}^{\top}\right),$$

for a random vector $\mathbf{Z}$ with finite covariance matrix $\text{Cov}(\mathbf{Z})$, and the entropy for a Gaussian random vector. Following the same procedure as in Zhang et al. (2019) (Lemma 11), we have

$$\log\det \mathbb{E}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}} \mathbf{Y}\mathbf{Y}^{\top} = nm \cdot \log \sigma^2 + n \cdot \log\det\left(\mathbf{I} + {}^{\mathbf{B}^{\top}\mathbf{B}}/_{\sigma^2}\right),$$

which further yields to

$$\mathsf{I}(\mathbf{\Pi}, \text{supp}(\mathbf{B}); \mathbf{Y} \mid \mathbf{X}) \leq \frac{n}{2} \log\det\left(\mathbf{I} + {}^{\mathbf{B}^{\top}\mathbf{B}}/_{\sigma^2}\right).$$

Summary. Combing (9), (10), and (11) then leads to a lower-bound on $\mathbb{P}_{\mathbf{X},\mathbf{W},\mathbf{\Pi},\mathbf{B}}\left[(\mathbf{\Pi}, \mathrm{supp}(\mathbf{B})) \neq (\widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\mathbf{B}}))\right]$ reading as

$$\mathbb{P}_{\mathbf{X},\mathbf{W},\mathbf{\Pi},\mathbf{B}}\left[(\mathbf{\Pi}, \mathrm{supp}(\mathbf{B})) \neq (\widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\mathbf{B}}))\right] \geq \frac{\log n! + m\dot{\log}\binom{p}{k} - 1 - (n/2) \log\det\left(\mathbf{I} + \mathbf{B}^\top\mathbf{B}/\sigma^2\right)}{\log\left(|(\mathbf{\Pi}, \mathrm{supp}(\mathbf{B}))|\right)}.$$

Easily, we can verify $\mathbb{P}_{\mathbf{X},\mathbf{W},\mathbf{\Pi},\mathbf{B}}\left[(\mathbf{\Pi}, \mathrm{supp}(\mathbf{B})) \neq (\widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\mathbf{B}}))\right]$ is lower bounded by $1/2$ given the assumptions in Theorem 2 and thus complete the proof.

□

B. Proof of Theorem 3

We define $\widehat{\mathbf{B}}$ as $(n-h)^{-1}\mathbf{X}^\top\mathbf{Y}$ and define operator $\mathsf{imax}(i)$ as $\mathrm{argmax}_j|\widehat{\mathbf{B}}_{j,i}|$ $(1 \leq i \leq m)$ for each column in $\widehat{\mathbf{B}}$, which returns the index of the entry with the largest magnitude. With a slight abuse of notation, we denote $\widehat{\mathbf{B}}_{\mathsf{imax}} = \mathsf{thres}(\widehat{\mathbf{B}})$. The benefits of this notation will be seen shortly.

In addition, we define the error $\mathcal{E}_{\text{err}}$ as

$$\mathcal{E}_{\text{err}} = \left\{\exists\, \mathbf{\Pi},\ \text{s.t.}\ \ \left\langle \mathbf{Y}, \mathbf{\Pi}\mathbf{X}\cdot\widehat{\mathbf{B}}_{\mathsf{imax}}\right\rangle \geq \left\langle \mathbf{Y}, \mathbf{\Pi}^\natural\mathbf{X}\cdot\widehat{\mathbf{B}}_{\mathsf{imax}}\right\rangle\right\}. \tag{12}$$

According to the sensing relation such that $\mathbf{Y} = \mathbf{\Pi}^\natural\mathbf{X}\mathbf{B}^\natural + \mathbf{W}$, we rewrite (12) as

$$\left\langle \mathbf{\Pi}^\natural\mathbf{X}\mathbf{B}^\natural + \mathbf{W}, \mathbf{\Pi}\mathbf{X}\cdot\widehat{\mathbf{B}}_{\mathsf{imax}}\right\rangle \leq \left\langle \mathbf{\Pi}^\natural\mathbf{X}\mathbf{B}^\natural + \mathbf{W}, \mathbf{\Pi}^\natural\mathbf{X}\cdot\widehat{\mathbf{B}}_{\mathsf{imax}}\right\rangle,$$

and would like to show it holds with a probability near zero. The major technical difficulties come from the fact that $\widehat{\mathbf{B}}$ is correlated with sensing matrix $\mathbf{X}$, which introduces high-order moments. The solution to such a challenge is broadly divided into the following two parts.

Part I: Relaxation of the error event. We first relax the error event $\mathcal{E}_{\text{err}}$

$$\mathcal{E}_{\text{err}} \subseteq \underbrace{\left\{\left\langle \mathbf{Y}_{i,:}, \widehat{\mathbf{B}}_{\mathsf{imax}}^\top\mathbf{X}_{\pi^\natural(i),:}\right\rangle \leq \left\langle \mathbf{Y}_{i,:}, \widehat{\mathbf{B}}_{\mathsf{imax}}^\top\mathbf{X}_{j,:}\right\rangle,\ \exists\ j \neq \pi^\natural(i)\right\}}_{\triangleq\ \mathcal{E}_{\text{err-relax}}}, \tag{13}$$

which means $\mathbb{P}(\mathcal{E}_{\text{err}}) \leq \mathbb{P}(\mathcal{E}_{\text{err-relax}})$.

Part II: Decoupling dependence via a modified leave-one-out technique. To decompose the dependence between $\widehat{\mathbf{B}}$ and the rows $\mathbf{X}_{\pi^\natural(i),:}$ and $\mathbf{X}_{j,:}$, we modify the leave-one-out technique and construct a perturbed matrix $\widehat{\mathbf{B}}^{(\pi^\natural(i),j)}$, which shares almost identical statistical behaviors as $\widehat{\mathbf{B}}$. Before delving into the technical details, we first provide a glimpse of the construction idea. Recalling the definition of $\widehat{\mathbf{B}}$, which is written as

$$\widehat{\mathbf{B}} = \frac{1}{n-h}\left(\sum_{\ell=1}^{n}\mathbf{X}_{\ell,:}\mathbf{X}_{\pi^\natural(\ell),:}^\top\right)\mathbf{B}^\natural + \frac{\mathbf{X}^\top\mathbf{W}}{n-h},$$

we construct the perturbed matrix $\widehat{\mathbf{B}}^{(\pi^\natural(i),j)}$ by replacing the corresponding rows with their i.i.d. samples. The detailed construction method is stated as follows.

To begin with, we draw i.i.d. samples for each rows of $\mathbf{X}_{i,:}$ and denote it as $\widetilde{\mathbf{X}}_{i,:}$ $(1 \leq i \leq n)$. Similarly we draw samples $\widetilde{\mathbf{W}}_{j,:}$ for each row in $\mathbf{W}$, $1 \leq j \leq n$. For arbitrary indices $\pi^\natural(i)$ and j such that $j \neq \pi^\natural(i)$, we create the samples $\widehat{\mathbf{B}}^{(\pi^\natural(i),j)}$ as (assume $i \neq \pi^\natural(i)$ and $j \neq \pi^\natural(j)$)

$$\begin{aligned}\widehat{\mathbf{B}}^{(\pi^\natural(i),j)} =\ & (n-h)^{-1}\left(\sum_{\ell,\pi^\natural(\ell)\neq\pi^\natural(i),j}\mathbf{X}_{\ell,:}\mathbf{X}_{\pi^\natural(\ell),:}^\top\right)\mathbf{B}^\natural \\ & + (n-h)^{-1}\left(\widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{X}}_{\pi^\natural(i),:}^\top + \widetilde{\mathbf{X}}_{\pi^\natural(i),:}\widetilde{\mathbf{X}}_{\pi^\natural(\pi^\natural(i)),:}^\top + \widetilde{\mathbf{X}}_{j,:}\widetilde{\mathbf{X}}_{\pi^\natural(j),:}^\top + \widetilde{\mathbf{X}}_{\pi^{\natural-1}(j),:}\widetilde{\mathbf{X}}_{j,:}^\top\right)\mathbf{B}^\natural\end{aligned}$$

$$+ (n-h)^{-1}\left(\sum_{\ell\neq \pi^{\natural}(i), i} \mathbf{X}_{\ell,:}\mathbf{W}_{\ell,:}^{\top} + \widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}\widetilde{\mathbf{W}}_{\pi^{\natural}(i),:}^{\top} + \widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{W}}_{i,:}^{\top}\right). \tag{14}$$

Provided that $i = \pi^{\natural}(i)$, we can simplify the summaries $\widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}^{\top} + \widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}\widetilde{\mathbf{X}}_{\pi^{\natural}(\pi^{\natural}(i)),:}^{\top}$ and $\widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}\widetilde{\mathbf{W}}_{\pi^{\natural}(i),:}^{\top} + \widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{W}}_{i,:}^{\top}$ in the above construction as the terms $\widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{X}}_{i,:}^{\top}$ and $\widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{W}}_{i,:}^{\top}$, respectively. Similarly, we will simplify $\widetilde{\mathbf{X}}_{j,:}\widetilde{\mathbf{X}}_{\pi^{\natural}(j),:}^{\top} + \widetilde{\mathbf{X}}_{\pi^{\natural -1}(j),:}\widetilde{\mathbf{X}}_{j,:}^{\top}$ as $\mathbf{X}_{j,:}\mathbf{X}_{j,:}^{\top}$ when $j = \pi^{\natural}(j)$.

With the above construction method, easily we can verify that $\widehat{\mathbf{B}}^{(\pi^{\natural},j)}$ is independent of the rows $\mathbf{X}_{\pi^{\natural}(i),:}$, $\mathbf{X}_{j,:}$ and $\mathbf{W}_{i,:}$ as they are not involved in $\widehat{\mathbf{B}}^{(\pi^{\natural},j)}$. Before delving into the technical details, we first collect all required notations.

B.1. Notations

Define the following events

$$\mathcal{E}_1 \triangleq \left\{\left|\frac{\sum_{\ell=\pi^{\natural}(\ell)}\mathbf{X}_{\ell,i}^2}{n-h} - 1\right| \leq c_0\sqrt{\frac{\log(np)}{n-h}},\ \forall\, 1\leq i\leq p\right\};$$

$$\mathcal{E}_2(\boldsymbol{\beta}) \triangleq \left\{\frac{\sum_{\ell=\pi^{\natural}(\ell)}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\ell,\backslash i}\boldsymbol{\beta}_{\backslash i}\right\rangle}{n-h} \lesssim \sqrt{\frac{\log(mnp)}{n-h}}\|\boldsymbol{\beta}_{\backslash i}\|_2,\ \forall\, 1\leq i\leq p\right\};$$

$$\mathcal{E}_3(\boldsymbol{\beta}) \triangleq \left\{\frac{\sum_{\ell\neq\pi^{\natural}(\ell)}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}\right\rangle}{n-h} \lesssim \frac{\sqrt{h}\log(mnp)}{n-h}\|\boldsymbol{\beta}\|_2,\ \forall\, 1\leq i\leq p\right\};$$

$$\mathcal{E}_4 \triangleq \left\{\frac{\sum_{\ell=1}^{n}\mathbf{X}_{\ell,i}^{\top}\mathbf{W}_{\ell,j}}{n-h} \lesssim \frac{\sigma\sqrt{n}\log(mnp)}{n-h},\ \forall\, 1\leq i\leq p, 1\leq j\leq m\right\};$$

$$\mathcal{E}_5 \triangleq \left\{\left\|\left(\widehat{\mathbf{B}}_{\mathsf{imax}} - \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}\right)^{\top}\boldsymbol{x}\right\|_2 \lesssim \frac{\log^{3/2}(np)}{n-h}\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}} + \frac{\sigma(\log np)\sqrt{m(\log mn)}}{n-h},\ \ \forall\, 1\leq \pi^{\natural}(i)\neq j\leq p\right\},$$

where $\boldsymbol{\beta}\in\mathbb{R}^p$ is an arbitrary column of $\mathbf{B}_{\ell,:}^{\natural}$ $(1\leq \ell\leq m)$, $\boldsymbol{\beta}_{\backslash i}\in\mathbb{R}^p$ denotes its copy with the ith entry being set to be zero, and $\boldsymbol{x}\in\mathbb{R}^p$ denotes an arbitrary row in matrix $\mathbf{X}$, which follows Gaussian distribution $\mathsf{N}(\mathbf{0}, \mathbf{I}_{p\times p})$. Note that $\boldsymbol{x}$ is not necessarily independent from $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$.

For an arbitrary event E, we denote its complement as $\overline{E}$. In addition, we define matrix $\mathbf{M}^{(\pi^{\natural}(i),j)}$ as $\mathbf{B}^{\natural\top}\widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}$. For the notation simplicity, we drop the superscript $\pi^{\natural}(i)$ and j in $\mathbf{M}^{(\pi^{\natural}(i),j)}$ when there is no ambiguity. The following context provides the technical details and a diagram representing the dependence among all lemmas is put in Figure 6.

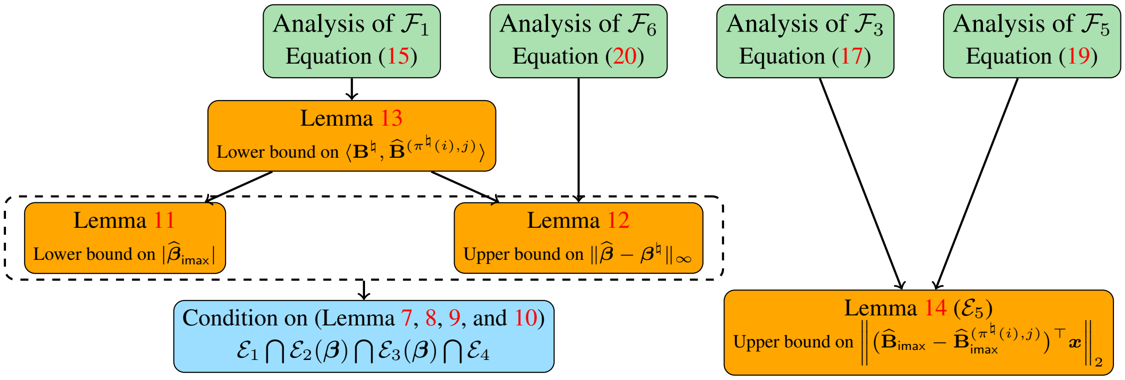
Analysis of $\mathcal{F}_1$ Equation (15)
Analysis of $\mathcal{F}_6$ Equation (20)
Analysis of $\mathcal{F}_3$ Equation (17)
Analysis of $\mathcal{F}_5$ Equation (19)
Lemma 13 Lower bound on $\langle \mathbf{B}^{\natural}, \widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}\rangle$
Lemma 11 Lower bound on $|\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}|$
Lemma 12 Upper bound on $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\natural}\|_{\infty}$
Condition on (Lemma 7, 8, 9, and 10) $\mathcal{E}_1\bigcap\mathcal{E}_2(\boldsymbol{\beta})\bigcap\mathcal{E}_3(\boldsymbol{\beta})\bigcap\mathcal{E}_4$
Lemma 14 ($\mathcal{E}_5$) Upper bound on $\left\|\left(\widehat{\mathbf{B}}_{\mathsf{imax}} - \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}\right)^{\top}\boldsymbol{x}\right\|_2$

Figure 6. Dependence diagram of lemmas.

B.2. Proof of Theorem 3

Proof. The proof can be broadly divided into three stages.

Stage I. To begin with, we prove the relation $\mathcal{E}_{\text{err}} \subseteq \mathcal{E}_{\text{err-relax}}$, whose definition can be found in (13). Conditional on $\overline{\mathcal{E}}_{\text{err-relax}}$, we have

$$\left\langle \mathbf{Y}_{i,:}, \widehat{\mathbf{B}}_{\mathsf{imax}}^{\top}\mathbf{X}_{\pi^{\natural}(i),:}\right\rangle > \left\langle \mathbf{Y}_{i,:}, \widehat{\mathbf{B}}_{\mathsf{imax}}^{\top}\mathbf{X}_{j,:}\right\rangle, \ \ \forall\ j \neq \pi^{\natural}(i).$$

Thus we conclude

$$\left\langle \mathbf{Y}, \mathbf{\Pi}^{\natural}\mathbf{X}\cdot \widehat{\mathbf{B}}_{\mathsf{imax}}\right\rangle > \left\langle \mathbf{Y}, \mathbf{\Pi}\mathbf{X}\cdot \widehat{\mathbf{B}}_{\mathsf{imax}}\right\rangle, \ \ \forall\, \mathbf{\Pi} \neq \mathbf{\Pi}^{\natural},$$

which suggests $\overline{\mathcal{E}}_{\text{err-relax}}$ will automatically lead to $\overline{\mathcal{E}}_{\text{err}}$, in other words, $\overline{\mathcal{E}}_{\text{err-relax}} \subseteq \overline{\mathcal{E}}_{\text{err}}$. Hence, we could upper bound $\mathbb{P}\left(\mathcal{E}_{\text{err}}\right)$ by $\mathbb{P}\left(\mathcal{E}_{\text{err-relax}}\right)$.

Stage II. Regarding the relation $\langle \mathbf{Y}_{i,:}, \widehat{\mathbf{B}}_{\mathsf{imax}}^{\top}\mathbf{X}_{\pi^{\natural}(i),:}\rangle \leq \langle \mathbf{Y}_{i,:}, \widehat{\mathbf{B}}_{\mathsf{imax}}^{\top}\mathbf{X}_{j,:}\rangle$, we can recast it as

$$\begin{aligned}
\underbrace{\langle \mathbf{B}^{\natural\top}\mathbf{X}_{\pi^{\natural}(i),:}, \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)\top}\mathbf{X}_{\pi^{\natural}(i),:}\rangle}_{\eta_1^{(\pi^{\natural}(i),j)}} \leq\ & \underbrace{\langle \mathbf{B}^{\natural\top}\mathbf{X}_{\pi^{\natural}(i),:}, \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)\top}\mathbf{X}_{j,:}\rangle}_{\eta_2^{(\pi^{\natural}(i),j)}} \\
&+ \underbrace{\langle \mathbf{B}^{\natural\top}\mathbf{X}_{\pi^{\natural}(i),:}, \left(\widehat{\mathbf{B}}_{\mathsf{imax}} - \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}\right)^{\top}\left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^{\natural}(i),:}\right)\rangle}_{\eta_3^{(\pi^{\natural}(i),j)}} \\
&+ \underbrace{\left\langle \mathbf{W}_{i,:}, \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)\top}\left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^{\natural}(i),:}\right)\right\rangle}_{\eta_4^{(\pi^{\natural}(i),j)}} \\
&+ \underbrace{\left\langle \mathbf{W}_{i,:}, \left(\widehat{\mathbf{B}}_{\mathsf{imax}} - \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}\right)^{\top}\left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^{\natural}(i),:}\right)\right\rangle}_{\eta_5^{(\pi^{\natural}(i),j)}}.
\end{aligned}$$

We should emphasize that the subscript imax are solely determined by $\widehat{\mathbf{B}}$ rather than its perturbed partner $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$. This is to ensure $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$ share the same support set. In the following analysis, we will see this property plays an important role in bounding the difference between $\widehat{\mathbf{B}}_{\mathsf{imax}}$ and $\widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}$, which is contained in $\eta_3^{(\pi^{\natural}(i),j)}$ and $\eta_5^{(\pi^{\natural}(i),j)}$. The following context separately studies each term $\eta_{\ell}^{(\pi^{\natural}(i),j)}$ $(1 \leq \ell \leq 5)$. First we define quantities $\Delta_{\ell}^{(\pi^{\natural}(i),j)}$ $(1 \leq \ell \leq 5)$ as

$$\begin{aligned}
\Delta_1^{(\pi^{\natural}(i),j)} &\triangleq \frac{\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^2}{k} - \frac{m\sigma^2\left(\log mnp\right)^2}{n} - \frac{\sqrt{m}\sigma\log(mnp)}{\sqrt{n}}\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}} - \frac{c_0\log n}{\sqrt{\operatorname{srank}(\mathbf{B}^{\natural})}}\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}\left\|\widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}\right\|_{\mathrm{F}}; \\
\Delta_2^{(\pi^{\natural}(i),j)} &\triangleq \log n\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}\left\|\widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}\right\|_{\mathrm{F}}/\sqrt{\operatorname{srank}(\mathbf{B}^{\natural})}; \\
\Delta_3^{(\pi^{\natural}(i),j)} &\triangleq \frac{\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^2\sqrt{\log n}\left(\log np\right)^{3/2}}{n} + \frac{\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}\sigma\sqrt{m\log n}(\log mn)^{1/2}(\log np)}{n}; \\
\Delta_4^{(\pi^{\natural}(i),j)} &\triangleq \sigma(\log n)\left\|\widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}\right\|_{\mathrm{F}}; \\
\Delta_5^{(\pi^{\natural}(i),j)} &\triangleq \frac{\sqrt{m\log n}\left(\log^{3/2} np\right)}{n-h}\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}\sigma + \frac{m\sigma^2(\log np)\sqrt{(\log n)(\log mn)}}{n-h}.
\end{aligned}$$

In addition, we define the events $\mathcal{F}_{\ell}$ $(1 \leq \ell \leq 6)$ as

$$\mathcal{F}_1 \triangleq \left\{\eta_1^{(\pi^{\natural}(i),j)} \gtrsim \Delta_1^{(\pi^{\natural}(i),j)},\ \forall\, 1 \leq \pi^{\natural}(i) \neq j \leq p\right\}; \tag{15}$$

$$\mathcal{F}_2 \triangleq \left\{|\eta_2^{(\pi^\natural(i),j)}| \lesssim \Delta_2^{(\pi^\natural(i),j)},\ \forall\, 1 \le \pi^\natural(i) \neq j \le p\right\}; \tag{16}$$

$$\mathcal{F}_3 \triangleq \left\{|\eta_3^{(\pi^\natural(i),j)}| \lesssim \Delta_3^{(\pi^\natural(i),j)},\ \forall\, 1 \le \pi^\natural(i) \neq j \le p\right\}; \tag{17}$$

$$\mathcal{F}_4 \triangleq \left\{|\eta_4^{(\pi^\natural(i),j)}| \lesssim \Delta_4^{(\pi^\natural(i),j)},\ \forall\, 1 \le \pi^\natural(i) \neq j \le p\right\}; \tag{18}$$

$$\mathcal{F}_5 \triangleq \left\{|\eta_5^{(\pi^\natural(i),j)}| \lesssim \Delta_5^{(\pi^\natural(i),j)},\ \forall\, 1 \le \pi^\natural(i) \neq j \le p\right\}, \tag{19}$$

$$\mathcal{F}_6 \triangleq \left\{\left\|\widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^\natural(i),j)} - \mathbf{B}_{\mathsf{imax}}^\natural\right\|_{\mathrm{F}} \lesssim \frac{\log(mnp)}{\sqrt{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} + \frac{\sigma\log(mnp)\sqrt{m}}{\sqrt{n}},\ \forall\, 1 \le \pi^\natural(i) \neq j \le p\right\}. \tag{20}$$

In Lemma 1, Lemma 2, Lemma 3, Lemma 4, Lemma 5, and Lemma 12, we will show all the above events, namely, $\mathcal{F}_\ell$ $(1 \le \ell \le 6)$, hold with probability approaching one.

Stage III. Given the assumptions in Theorem 3, we will verify the relation $\bigcap_{\ell=1}^6 \mathcal{F}_\ell \subseteq \mathcal{E}_{\text{err-relax}}$. Considering the difference $\Delta_1^{(\pi^\natural(i),j)} - \sum_{\ell=2}^5 \Delta_\ell^{(\pi^\natural(i),j)}$, we can lower-bound it as

$$\begin{aligned}\frac{\Delta_1^{(\pi^\natural(i),j)}}{m\sigma^2} - \sum_{\ell=2}^5 \frac{\Delta_\ell^{(\pi^\natural(i),j)}}{m\sigma^2} \ge\ & \mathsf{SNR}\underbrace{\left(\frac{1}{k} - \frac{\sqrt{\log n}\log^{3/2}(np)}{n} - \frac{2\log n}{\sqrt{\mathsf{srank}(\mathbf{B}^\natural)}}\right)}_{\zeta_1} \\ & - \sqrt{\mathsf{SNR}}\underbrace{\left(\frac{\sqrt{\log n}(\log np)\left(\sqrt{\log mn} + \sqrt{\log np}\right)}{n} + \frac{\log mnp}{\sqrt{n}} + \frac{\log n}{\sqrt{m}} + \frac{2(\log n)(\log mnp)}{\sqrt{n\cdot \mathsf{srank}(\mathbf{B}^\natural)}}\right)}_{\zeta_{1/2}} \\ & - \underbrace{\left(\frac{\log n(\log mnp)}{\sqrt{mn}} + \frac{\log^2(mnp) + \sqrt{\log n(\log mn)}(\log np)}{n}\right)}_{\zeta_0}.\end{aligned}$$

Under the assumptions in Theorem 3, we have

$$\begin{aligned}\zeta_1 &\asymp k^{-1}; \\ \zeta_{1/2} &\asymp \frac{\log mnp}{\sqrt{n}} + \frac{\log n}{\sqrt{m}}; \\ \zeta_0 &\asymp \frac{\log n(\log mnp)}{\sqrt{mn}} + \frac{\log^2(mnp)}{n},\end{aligned}$$

which leads to

$$\begin{aligned}\frac{\Delta_1^{(\pi^\natural(i),j)}}{m\sigma^2} - \sum_{\ell=2}^5 \frac{\Delta_\ell^{(\pi^\natural(i),j)}}{m\sigma^2} &\ge \zeta_1\cdot\mathsf{SNR} - \zeta_{1/2}\sqrt{\mathsf{SNR}} - \zeta_0 \\ &\overset{①}{\gtrsim} \frac{1}{k} - \frac{1}{k\sqrt{\log n}} - \frac{1}{k\log^{\varepsilon} n} - \frac{1}{k^2\log^{\varepsilon+1/2} n} - \frac{1}{k^2\log n} > 0,\end{aligned}$$

where in ① we use the relation $m \ge \mathrm{rank}(\mathbf{B}^\natural) \ge \mathrm{srank}(\mathbf{B}^\natural) \gg k^2\log^{2(1+\varepsilon)} n$. Then we conclude

$$\sum_{\ell=2}^5 |\eta_\ell^{(\pi^\natural(i),j)}| \le \sum_{\ell=2}^5 \Delta_\ell^{(\pi^\natural(i),j)} \le \Delta_1^{(\pi^\natural(i),j)} \le \eta_1^{(\pi^\natural(i),j)},\ \ \forall\, 1 \le \pi^\natural(i) \neq j \le p,$$

which means we will automatically obtain $\mathcal{E}_{\text{err-relax}}$ when assuming $\bigcap_{\ell=1}^6 \mathcal{F}_\ell$, in other words, $\bigcap_{\ell=1}^6 \mathcal{F}_\ell \subseteq \mathcal{E}_{\text{err-relax}}$. Hence, we can complete the proof by $\mathbb{P}\left(\mathcal{E}_{\text{err}}\right) \le \mathbb{P}\left(\mathcal{E}_{\text{err-relax}}\right) \le \sum_{\ell=1}^6 \mathbb{P}(\overline{\mathcal{F}}_\ell)$.

□

Lemma 1. *Conditional on the intersection of events $\mathcal{E}_1 \bigcap_{\ell=1}^m \mathcal{E}_2(\mathbf{B}^\natural_{\ell,:}) \bigcap_{\ell=1}^m \mathcal{E}_3(\mathbf{B}^\natural_{\ell,:}) \bigcap \mathcal{E}_4$, we have $\mathbb{P}(\mathcal{F}_1) \geq 1 - c_0 n^{-c_1}$ provided $n \gg k \log^2(mnp)$.*

Proof. Recalling the definition of $\mathbf{M}$, i.e., $\mathbf{M} \triangleq \mathbf{B}^{\natural\top} \widehat{\mathbf{B}}^{(\pi^\natural(i),j)}_{\mathsf{imax}}$, we divide the proof procedure as

- **Step I.** Condition on $\mathbf{M}$, we have

  $$\eta_1^{(\pi^\natural(i),j)} \geq \mathrm{Tr}(\mathbf{M}) - \frac{c_0 \log n}{\sqrt{\mathrm{srank}(\mathbf{B}^\natural)}} \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} \left\|\widehat{\mathbf{B}}^{(\pi^\natural(i),j)}_{\mathsf{imax}}\right\|_{\mathrm{F}}, \quad \forall\, 1 \leq \pi^\natural(i) \neq j \leq n,$$

  hold with probability exceeding $1 - n^{-c}$.

- **Step II.** Provided $n \gg k \log^2(mnp)$, we have

  $$\mathrm{Tr}(\mathbf{M}) \gtrsim \frac{1}{k}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2 - \frac{m\sigma^2\left(\log mnp\right)^2}{n} - \frac{\sqrt{m}\sigma \log(mnp)}{\sqrt{n}} \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}},$$

  condition on $\mathcal{E}_1 \bigcap_{\ell=1}^m \mathcal{E}_2(\mathbf{B}^\natural_{\ell,:}) \bigcap_{\ell=1}^m \mathcal{E}_3(\mathbf{B}^\natural_{\ell,:}) \bigcap \mathcal{E}_4$.

For the clarify of presentation, we defer the proof of **Step II** to Lemma 13 and focus on **Step I**. Due to the construction of $\widehat{\mathbf{B}}^{(\pi^\natural,j)}$ in (14), we conclude $\widehat{\mathbf{B}}^{(\pi^\natural,j)}$ is independent with row $\mathbf{X}_{\pi^\natural(i),:}$. Hence we can first condition on $\mathbf{M}$ and rewrite term η_1 in terms of a quadratic product $\boldsymbol{x}^\top \mathbf{M} \boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^p$ is a random vector satisfying $\boldsymbol{x} \in \mathsf{N}(\mathbf{0}, \mathbf{I}_{p\times p})$.

With Hanson-Wright inequality (c.f. Theorem 6.2.1 in Vershynin (2018)), we have

$$\begin{aligned}
&\mathbb{P}\left(|\eta_1^{(\pi^\natural(i),j)} - \mathbb{E}\eta_1^{(\pi^\natural(i),j)}| \geq t,\ \exists\, 1 \leq \pi^\natural(i) \neq j \leq n\right) \\
\leq\ & n^2 \cdot \mathbb{P}\left(\left|\boldsymbol{x}^\top \mathbf{M}\boldsymbol{x} - \mathrm{Tr}(\mathbf{M})\right| \geq t\right) \leq n^2 \cdot 2\exp\left(-c_0\left(\frac{t}{\|\mathbf{M}\|_{\mathrm{OP}}} \wedge \frac{t^2}{\|\mathbf{M}\|_{\mathrm{F}}^2}\right)\right).
\end{aligned}$$

Setting $t = \log n \|\mathbf{M}\|_{\mathrm{F}}$, we have

$$\eta_1^{(\pi^\natural(i),j)} \geq \mathbb{E}\eta_1^{(\pi^\natural(i),j)} - c_0 \log n \|\mathbf{M}\|_{\mathrm{F}} = \mathrm{Tr}(\mathbf{M}) - c_0 \log n \|\mathbf{M}\|_{\mathrm{F}},\ \forall\, 1 \leq \pi^\natural(i) \neq j \leq n,$$

holds with probability $1 - 2n^{-c}$. Then we complete the proof by showing

$$\|\mathbf{M}\|_{\mathrm{F}} \leq \left\|\mathbf{B}^\natural\right\|_{\mathrm{OP}} \left\|\widehat{\mathbf{B}}^{(\pi^\natural(i),j)}_{\mathsf{imax}}\right\|_{\mathrm{F}} \overset{\textcircled{1}}{=} \frac{1}{\sqrt{\mathrm{srank}(\mathbf{B}^\natural)}} \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} \left\|\widehat{\mathbf{B}}^{(\pi^\natural(i),j)}_{\mathsf{imax}}\right\|_{\mathrm{F}},$$

where ① is due to the definition of stable rank.

□

Lemma 2. *We have $\mathbb{P}(\mathcal{F}_2) \geq 1 - 4n^{-c}$ where event $\mathcal{F}_2$ is defined in (16).*

Proof. First, we fix the indices $\pi^\natural(i)$ and j such that $\pi^\natural(i) \neq j$. Due to the independence across the rows of $\mathbf{X}$, we conclude

$$\begin{aligned}
\mathbb{P}\left(|\boldsymbol{x}^\top \mathbf{M}\boldsymbol{y}| \gtrsim \log n \|\mathbf{M}\|_{\mathrm{F}}\right) \leq\ & \mathbb{P}\left(|\boldsymbol{x}^\top \mathbf{M}\boldsymbol{y}| \gtrsim \log n \|\mathbf{M}\|_{\mathrm{F}},\ \|\mathbf{M}\boldsymbol{y}\|_2 \lesssim \sqrt{\log n}\|\mathbf{M}\|_{\mathrm{F}}\right) \\
&+ \mathbb{P}\left(\|\mathbf{M}\boldsymbol{y}\|_2 \gtrsim \sqrt{\log n}\|\mathbf{M}\|_{\mathrm{F}}\right) \\
\overset{\textcircled{1}}{\leq}\ & \mathbb{P}\left(|\boldsymbol{x}^\top \mathbf{M}\boldsymbol{y}| \gtrsim \sqrt{\log n}\|\mathbf{M}\boldsymbol{y}\|_2\right) + 2n^{-c} \leq 4n^{-c},
\end{aligned}$$

where $\mathbf{M}$ is defined as $\mathbf{B}^{\natural\top}\widehat{\mathbf{B}}^{(\pi^\natural(i),j)}_{\mathsf{imax}}$ and in ① we have the rotation invariance of Gaussian random vector. Regarding the event $\mathcal{F}_2$, we invoke the union bound and complete the proof as $\mathbb{P}(\mathcal{F}_2) \leq n^2 \cdot 4n^{-c} = 4n^{-c^{'}}$. □

Lemma 3. *Condition on $\mathcal{E}_5$, we conclude $\mathbb{P}\left(\mathcal{F}_3 | \mathcal{E}_5\right) \geq 1 - 2n^{-c}$, where $\mathcal{F}_3$ is defined in (17).*

Proof. First we recall the definition of η_3, which is written as

$$\eta_3 \triangleq \langle \mathbf{B}^{\natural\top}\mathbf{X}_{i,:}, \left(\widehat{\mathbf{B}}_{\mathsf{imax}} - \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}\right)^{\top} (\mathbf{X}_{j,:} - \mathbf{X}_{i,:})\rangle.$$

We begin the proof as

$$\mathbb{P}\left(\overline{\mathcal{F}}_3 | \mathcal{E}_5\right) \leq \mathbb{P}\Bigg(\|\mathbf{B}^{\natural\top}\mathbf{X}_{i,:}\|_2 \gtrsim \log n \left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^2, \ \exists\, 1 \leq i \leq n | \mathcal{E}_5 \Bigg) + \mathbb{P}\left(\overline{\mathcal{F}}_3, \ \|\mathbf{B}^{\natural\top}\mathbf{X}_{i,:}\|_2 \lesssim \log n \left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^2, \ \forall\, 1 \leq i \leq n | \mathcal{E}_5\right).$$

For the first term, we invoke the Hanson-Wright inequality (c.f. Theorem 6.2.1 in Vershynin (2018)), which leads to

$$\begin{aligned}
\mathbb{P}\Bigg(\|\mathbf{B}^{\natural\top}\mathbf{X}_{i,:}\|_2 \gtrsim \log n \left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^2, \ \exists\, 1 \leq i \leq n | \mathcal{E}_5 \Bigg) &\leq \ n \cdot \mathbb{P}\left(\left| \|\mathbf{B}^{\natural\top}\mathbf{X}_{i,:}\|_2^2 - \mathbb{E}\|\mathbf{B}^{\natural\top}\mathbf{X}_{i,:}\|_2^2 \right| \gtrsim \log n \left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^2 | \mathcal{E}_5 \right) \\
&\leq \ 2n \cdot \exp\left(-c \left(\frac{\log^2 n \left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^4}{\|\mathbf{B}^{\natural\top}\mathbf{B}^{\natural}\|_{\mathrm{F}}^2} \wedge \frac{\log n \left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^2}{\|\mathbf{B}^{\natural\top}\mathbf{B}^{\natural}\|_{\mathrm{OP}}} \right) \right) \leq 2n^{-c}.
\end{aligned}$$

For the second term, we will prove it to be zero. This is because

$$\begin{aligned}
|\eta_3^{(\pi^{\natural}(i),j)}| &\leq \ \|\mathbf{B}^{\natural\top}\mathbf{X}_{\pi^{\natural}(i),:}\|_2 \cdot \left\| \left(\widehat{\mathbf{B}}_{\mathsf{imax}} - \widehat{\mathbf{B}}_{\mathsf{imax}}^{(i,j)}\right)^{\top} \left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^{\natural}(i),:}\right) \right\|_2 \\
&\overset{\textcircled{1}}{\lesssim} \ \sqrt{\log n} \left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}} \cdot \frac{\log^{3/2}(np)}{n-h} \left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}} + \frac{\sigma (\log np) \sqrt{m (\log mn)}}{n-h},
\end{aligned}$$

for all $1 \leq \pi^{\natural}(i) \neq j \leq n$, where in ① we condition on $\mathcal{E}_5$ and $\left\|\mathbf{B}^{\natural\top}\mathbf{X}_{\pi^{\natural}(i),:}\right\|_2 \lesssim \sqrt{\log n}\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}$. □

Lemma 4. *We have $\mathbb{P}(\mathcal{F}_4) \geq 1 - 4n^{-c}$.*

Proof. First we fix the indices $\pi^{\natural}(i)$ and j such that $\pi^{\natural}(i) \neq j$. Invoke the definition of $\widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}$, we conclude $\mathbf{W}_{i,:}$, $\widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)}$, $\mathbf{X}_{j,:}$, and $\mathbf{X}_{\pi^{\natural}(i),:}$ are independent with each other. Hence we can complete the proof as

$$\begin{aligned}
& \mathbb{P}\left(|\eta_4^{(\pi^{\natural}(i),j)}| \gtrsim \sigma (\log n) \left\| \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)} \right\|_{\mathrm{F}} \right) \\
\leq \ & \mathbb{P}\left(\left\| \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)\top} \left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^{\natural}(i),:}\right) \right\|_2 \gtrsim \sqrt{\log n} \left\| \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)} \right\|_{\mathrm{F}} \right) \\
+ \ & \mathbb{P}\left(|\eta_4| \gtrsim \sigma (\log n) \left\| \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)} \right\|_{\mathrm{F}}, \left\| \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)\top} \left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^{\natural}(i),:}\right) \right\|_2 \lesssim \sqrt{\log n} \left\| \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)} \right\|_{\mathrm{F}} \right) \\
\leq \ & 2n^{-c} + \mathbb{P}\left(|\eta_4| \gtrsim \sigma \sqrt{\log n} \left\| \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)\top} \left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^{\natural}(i),:}\right) \right\|_2 \right) \overset{\textcircled{1}}{\leq} 4n^{-c},
\end{aligned}$$

where in ① we use the tail bound for the Gaussian random variable. The proof is then completed with the union bound such that

$$\mathbb{P}\left(\overline{\mathcal{F}}_4\right) \leq n^2 \mathbb{P}\left(|\eta_4^{(\pi^{\natural}(i),j)}| \gtrsim \sigma (\log n) \left\| \widehat{\mathbf{B}}_{\mathsf{imax}}^{(\pi^{\natural}(i),j)} \right\|_{\mathrm{F}} \right) \leq 4n^2 \cdot n^{-c} = 1 - 4n^{-c^{\prime}}.$$

□

Lemma 5. *Condition on $\mathcal{E}_5$, we conclude $\mathbb{P}(\mathcal{F}_5 | \mathcal{E}_5) \geq 1 - 2n^{-c}$.*

Proof. The proof is in a similar form to that for Lemma 3. First we decompose the probability $\mathbb{P}(\overline{\mathcal{F}}_5 | \mathcal{E}_5)$ as

$$\mathbb{P}(\overline{\mathcal{F}}_5 | \mathcal{E}_5) \leq \ \mathbb{P}(\|\mathbf{W}_{i,:}\|_2 \gtrsim \sigma\sqrt{m \log n}, \ \exists\, 1 \leq i \leq n | \mathcal{E}_5) + \mathbb{P}\left(\overline{\mathcal{F}}_5, \ \|\mathbf{W}_{i,:}\|_2 \lesssim \sigma\sqrt{m \log n}, \ \forall\, 1 \leq i \leq n | \mathcal{E}_5\right)$$

$$\leq\ 2n\cdot n^{-c} + \mathbb{P}\left(\overline{\mathcal{F}}_5,\ \|\mathbf{W}_{i,:}\|_2 \lesssim \sigma\sqrt{m\log n},\ \forall\, 1\leq i\leq n|\mathcal{E}_5\right),$$

where the last inequality is due to the tail bound for the Gaussian random variable. We complete the proof by showing the second probability is zero. This is because

$$\begin{aligned}|\eta_5^{(\pi^\natural(i),j)}| &\leq\ \|\mathbf{W}_{i,:}\|_2\cdot\left\|\left(\widehat{\mathbf{B}}_{\mathsf{imax}}-\widehat{\mathbf{B}}^{(i,j)}_{\mathsf{imax}}\right)^\top(\mathbf{X}_{j,:}-\mathbf{X}_{i,:})\right\|_2\\ &\lesssim\ \frac{\sqrt{m\log n}\left(\log^{3/2} np\right)}{n-h}\vert\vert\vert\mathbf{B}^\natural\vert\vert\vert_{\mathrm{F}}\sigma+\frac{m\sigma^2(\log np)\sqrt{(\log n)(\log mn)}}{n-h}.\end{aligned}$$

□

Lemma 6. *Conditional on the intersection of events $\mathcal{E}_1\bigcap_{\ell=1}^m\mathcal{E}_2(\mathbf{B}^\natural_{\ell,:})\bigcap_{\ell=1}^m\mathcal{E}_3(\mathbf{B}^\natural_{\ell,:})\bigcap\mathcal{E}_4$, we have $\mathbb{P}(\mathcal{F}_5)=1$.*

Proof. We complete the proof by

$$\begin{aligned}\vert\vert\vert\widehat{\mathbf{B}}^{(\pi^\natural(i),j)}_{\mathsf{imax}}-\mathbf{B}^\natural_{\mathsf{imax}}\vert\vert\vert_{\mathrm{F}}^2 &= \sum_{\boldsymbol{\beta}^\natural}\left\|\widehat{\boldsymbol{\beta}}^{(\pi^\natural(i),j)}_{\mathsf{imax}}-\boldsymbol{\beta}^\natural_{\mathsf{imax}}\right\|_2^2 \overset{\textcircled{1}}{=} \sum_{\boldsymbol{\beta}^\natural}\left\|\widehat{\boldsymbol{\beta}}^{(\pi^\natural(i),j)}_{\mathsf{imax}}-\boldsymbol{\beta}^\natural_{\mathsf{imax}}\right\|_\infty^2\\ &\overset{\textcircled{2}}{\lesssim} \frac{\log^2(mnp)}{n}\vert\vert\vert\mathbf{B}^\natural\vert\vert\vert_{\mathrm{F}}^2+\frac{m\sigma^2\log^2(mnp)}{n},\end{aligned}$$

where ① is due to the fact such that $\widehat{\boldsymbol{\beta}}^{(\pi^\natural(i),j)}_{\mathsf{imax}}-\boldsymbol{\beta}^\natural$ has only one non-zero element, and ② is due to Lemma 12.

□

B.3. Supporting Lemmas

Lemma 7. *We have $\mathbb{P}(\mathcal{E}_1)\geq 1-2n^{-c_1}p^{-c_2}$ when $n-h\gg\log np$, and n,p are sufficiently large.*

Proof. Due to the Hanson-Wright inequality (c.f. Theorem 6.2.1 in Vershynin (2018)), we have

$$\mathbb{P}\left(\mathcal{E}_1\right)\leq 2p\exp\left[-c\left(\frac{(n-h)\log(np)}{\vert\vert\vert\mathbf{I}_{(n-h)\times(n-h)}\vert\vert\vert_{\mathrm{F}}^2}\wedge\frac{\sqrt{(n-h)\log(np)}}{\vert\vert\vert\mathbf{I}_{(n-h)\times(n-h)}\vert\vert\vert_{\mathrm{OP}}}\right)\right]\overset{\textcircled{1}}{\leq}2p\cdot n^{-c}p^{-c}=2n^{-c_1}p^{-c_2},$$

where in ① we use the fact $n-h\gg\log(np)$. □

Lemma 8. *For a fixed $\boldsymbol{\beta}\in\mathbb{R}^p$, we have $\mathbb{P}(\mathcal{E}_2(\boldsymbol{\beta}))\geq 1-pe^{-c_0(n-h)}-n^{-c_1}m^{-c_2}p^{-c_3}$.*

Proof. To begin with, we construct the sensing matrix $\mathbf{X}_\mathcal{S}$ by concatenating all rows $\mathbf{X}_{\ell,:}$ such that $\ell=\pi^\natural(\ell)$. With union bound, we can upper bound $\mathbb{P}\left(\overline{\mathcal{E}}_2(\boldsymbol{\beta})\right)$ as

$$\begin{aligned}\mathbb{P}\left(\overline{\mathcal{E}}_2(\boldsymbol{\beta})\right)\leq\ &\underbrace{p\mathbb{P}\left(\left\|\mathbf{X}_\mathcal{S}\boldsymbol{\beta}^\natural_{\setminus i}\right\|_2\geq\sqrt{2(n-h)}\|\boldsymbol{\beta}^\natural_{\setminus i}\|_2\right)}_{\zeta_1}\\ &+\underbrace{p\mathbb{P}\left(\sum_{\ell=1}^{n-h}\mathbf{X}_{\ell,i}\left\langle\mathbf{X}_{\ell,:},\boldsymbol{\beta}^\natural_{\setminus i}\right\rangle\gtrsim\sqrt{(n-h)\log(mnp)}\|\boldsymbol{\beta}^\natural_{\setminus i}\|_2,\left\|\mathbf{X}_\mathcal{S}\boldsymbol{\beta}^\natural_{\setminus i}\right\|_2<\sqrt{2(n-h)}\|\boldsymbol{\beta}^\natural_{\setminus i}\|_2\right)}_{\zeta_2}.\end{aligned}$$

Since $\mathbf{X}$ are with i.i.d Gaussian entries, we have each row in $\mathbf{X}_\mathcal{S}\boldsymbol{\beta}^\natural_{\setminus i}$ be a Gaussian random vector with zero mean and variance $\|\boldsymbol{\beta}^\natural_{\setminus i}\|_2^2$. Hence, we have $\|\mathbf{X}_\mathcal{S}\boldsymbol{\beta}^\natural_{\setminus i}\|_2^2/\|\boldsymbol{\beta}^\natural_{\setminus i}\|_2^2$ be a χ^2 random variable with freedom $n-h$, which leads to

$$\zeta_1 \overset{\textcircled{1}}{\leq} p\exp\left(\frac{n-h}{2}\left(\log 2 - 1\right)\right) \leq pe^{-0.65(n-h)},$$

where ① is due to Lemma 15.

For ζ_2, we notice that $\mathbf{X}_{\ell,i}$ is independent of the inner product $\langle \mathbf{X}_{\ell,:}, \boldsymbol{\beta}^{\natural}_{\setminus i}\rangle$. Hence, we can view the product $\sum_{\ell=1}^{n-h}\mathbf{X}_{\ell,i}\langle \mathbf{X}_{\ell,:}, \boldsymbol{\beta}^{\natural}_{\setminus i}\rangle$ as a Gaussian random variable $\mathsf{N}(0, \|\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^{\natural}_{\setminus i}\|_2^2)$, which leads to

$$\zeta_2 \leq \ p\mathbb{P}\left(\sum_{\ell=1}^{n-h}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\ell,:}, \boldsymbol{\beta}^{\natural}_{\setminus i}\right\rangle \gtrsim \sqrt{\log(mnp)}\|\mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}^{\natural}_{\setminus i}\|_2\right) \overset{\textcircled{2}}{\leq} 2n^{-c}m^{-c}p^{-c},$$

where in ② we use the tail bound of Gaussian random vectors. □

Lemma 9. *For a fixed $\boldsymbol{\beta} \in \mathbb{R}^p$, we have $\mathbb{P}(\mathcal{E}_3(\boldsymbol{\beta})) \geq 1 - c_0 n^{-c_1}m^{-c_2}p^{-c_3}$.*

Proof. According to Lemma 16, we can decompose the index set $\left\{\ell :\ \ell \neq \pi^{\natural}(\ell)\right\}$ into three disjoint sets $\mathcal{I}_j$ such that (i) indices ℓ and $\pi^{\natural}(\ell)$ do not fall into the same set $\mathcal{I}_j$; and (ii) the cardinality of each set satisfies $h_j \triangleq |\mathcal{I}_j| \geq \lfloor \frac{h}{3} \rfloor$, $(1 \leq j \leq 3)$.

Then we can decompose product $\sum_{\ell \neq \pi^{\natural}(\ell)}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}\right\rangle$ as

$$\sum_{\ell \neq \pi^{\natural}(\ell)}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}\right\rangle = \sum_{j=1}^{3}\sum_{\ell \in \mathcal{I}_j}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}^{\natural}\right\rangle.$$

With the union bound, we have

$$\mathbb{P}(\overline{\mathcal{E}}_3(\boldsymbol{\beta})) \leq \ p \cdot \sum_{j=1}^{3}\mathbb{P}\left(\sum_{\ell \in \mathcal{I}_j}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}^{\natural}\right\rangle \gtrsim (\log mp)\sqrt{h_j}\|\boldsymbol{\beta}^{\natural}\|_2\right). \tag{21}$$

Due to the properties of $\mathcal{I}_j$, we have $\mathbf{X}_{\ell,i}$ and $\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}^{\natural}\rangle$ be independent and hence

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\ell \in \mathcal{I}_j}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}^{\natural}\right\rangle \gtrsim \log(mnp)\sqrt{h_j}\|\boldsymbol{\beta}^{\natural}\|_2\right)\\
=\ &\mathbb{P}\left(\left\|\mathbf{X}_{\ell \in \mathcal{I}_j,:}\boldsymbol{\beta}^{\natural}\right\|_2^2 \gtrsim \log(mnp)|h_j|\|\boldsymbol{\beta}^{\natural}\|_2^2\right)\\
+\ &\mathbb{P}\left(\left\|\mathbf{X}_{\ell \in \mathcal{I}_j,:}\boldsymbol{\beta}^{\natural}\right\|_2^2 \lesssim \log(mnp)|h_j|\|\boldsymbol{\beta}^{\natural}\|_2^2,\ \sum_{\ell \in \mathcal{I}_j}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}^{\natural}\right\rangle \gtrsim \log(mnp)\sqrt{h_j}\|\boldsymbol{\beta}^{\natural}\|_2\right)\\
\overset{\textcircled{1}}{\leq}\ &\exp\left[\frac{h_j}{2}\left(\log(c_0\log mnp) - c_0\log mnp + 1\right)\right] + \mathbb{P}\left(\sum_{\ell \in \mathcal{I}_j}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}^{\natural}\right\rangle \gtrsim \sqrt{\log mnp}\|\mathbf{X}_{\ell \in \mathcal{I}_j,:}\boldsymbol{\beta}^{\natural}\|_2\right)\\
\overset{\textcircled{2}}{\lesssim}\ &n^{-c_0}m^{-c_0}p^{-c_0} + n^{-c_1}m^{-c_1}p^{-c_1} \asymp n^{-c}m^{-c}p^{-c},
\end{aligned}$$

where ① is due to Lemma 15 and ② is due to the Gaussian tail bound. Plugging it in (21) then completes the proof. □

Lemma 10. *We have $\mathbb{P}(\mathcal{E}_4) \geq 1 - cn^{-c_0}m^{-c}p^{-c}$.*

Proof. We complete the proof as

$$\mathbb{P}(\overline{\mathcal{E}}_4) \overset{\textcircled{1}}{\leq}\ mp \cdot \mathbb{P}\left(\sum_{\ell=1}^{n}\mathbf{X}_{\ell,i}^{\top}\mathbf{W}_{\ell,j} \gtrsim \sigma\sqrt{n}\log(mnp)\right)$$

$$
\begin{aligned}
&\overset{\textcircled{2}}{\le} mp\left[\mathbb{P}\left(\|\mathbf{X}_{:,i}\|_2 \gtrsim \sqrt{n\log(mnp)}\right) + \mathbb{P}\left(\sum_{\ell=1}^{n}\mathbf{X}_{\ell,i}^{\top}\mathbf{W}_{\ell,j} \gtrsim \sigma\sqrt{n}\log(mnp),~\|\mathbf{X}_{:,i}\|_2 \lesssim \sqrt{n\log(mnp)}\right)\right]\\
&\overset{\textcircled{3}}{\le} mp\left[\exp\left(\frac{n}{2}\left(\log(c_0\log(mnp)) - c_0\log(mnp)+1\right)\right) + \mathbb{P}\left(\sum_{\ell=1}^{n}\mathbf{X}_{\ell,i}^{\top}\mathbf{W}_{\ell,j} \gtrsim \sigma\sqrt{\log mnp}\,\|\mathbf{X}_{:,i}\|_2\right)\right]\\
&\overset{\textcircled{4}}{\le} mp\left(n^{-c_0}m^{-c_0}p^{-c_0} + n^{-c_1}m^{-c_1}p^{-c_1}\right) \asymp n^{-c_0}m^{-c}p^{-c},
\end{aligned}
$$

where ① and ② are due to the union bound, ③ is due to Lemma 15, and ④ is due to the tail bound for the Gaussian random variable. □

Lemma 11. *Conditional on the intersection of events $\mathcal{E}_1\bigcap\mathcal{E}_2(\boldsymbol{\beta}^{\natural})\bigcap\mathcal{E}_3(\boldsymbol{\beta}^{\natural})\bigcap\mathcal{E}_4$, we have*

$$
|\boldsymbol{\beta}^{\natural}_{\mathsf{imax}}| \gtrsim |\boldsymbol{\beta}^{\natural}_{\max}| - \frac{\sqrt{\log(mnp)}}{\sqrt{n}}\|\boldsymbol{\beta}^{\natural}\|_2 - \frac{\sigma\log(mnp)}{\sqrt{n}},
$$

where imax and $\max$ are defined as the indices of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{\natural}$ with the largest magnitude, i.e., $\mathsf{imax} \triangleq \operatorname{argmax}_i|\widehat{\boldsymbol{\beta}}_i|$ and $\max \triangleq \operatorname{argmax}_i|\boldsymbol{\beta}^{\natural}_i|$, respectively.

Proof. To begin with, we define $\zeta_1^{(i)}$, $\zeta_2^{(i)}$, and $\zeta_3^{(i)}$ as

$$
\begin{aligned}
\zeta_1^{(i)} &= \frac{1}{n-h}\left(\sum_{\ell=1}^{n-h}\mathbf{X}_{\ell,i}^2\right)\boldsymbol{\beta}^{\natural}_i;\\
\zeta_2^{(i)} &= \frac{1}{n-h}\sum_{\ell=1}^{n-h}\mathbf{X}_{\ell,i}\left\langle\mathbf{X}_{\ell,:},\boldsymbol{\beta}^{\natural}_{\setminus i}\right\rangle;\\
\zeta_3^{(i)} &= \frac{1}{n-h}\sum_{\ell=n-h+1}^{n}\mathbf{X}_{\ell,i}\left\langle\mathbf{X}_{\pi^{\natural}(\ell),:},\boldsymbol{\beta}^{\natural}\right\rangle;\\
\zeta_4^{(i)} &= \frac{1}{n-h}\sum_{\ell=1}^{n}\mathbf{X}_{\ell,i}^{\top}\mathbf{W}_{\ell,j},
\end{aligned}
$$

respectively. Then we can write $\widehat{\boldsymbol{\beta}}_i$ as $\sum_{j=1}^4\zeta_j^{(i)}$. Due to the definition of imax, we conclude $|\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}| \ge |\widehat{\boldsymbol{\beta}}_{\max}|$, where $\max$ is defined as the index of $\boldsymbol{\beta}^{\natural}$ with the largest magnitude, i.e., $\max \triangleq \operatorname{argmax}_i|\boldsymbol{\beta}^{\natural}_i|$. With triangle inequality, we obtain

$$
\sum_{j=1}^{4}|\zeta_j^{(\mathsf{imax})}| \ge |\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}| \ge |\widehat{\boldsymbol{\beta}}_{\max}| \ge |\zeta_1^{(\max)}| - \sum_{j=2}^{4}|\zeta_j^{(\max)}|. \tag{22}
$$

The following context separately discusses each term. First, we consider $|\zeta_1^{(\mathsf{imax})}|$ and $|\zeta_1^{(\max)}|$. Conditional on $\mathcal{E}_1$, we have

$$
\begin{aligned}
|\zeta_1^{(\mathsf{imax})}| &\le \left(1+c_0\sqrt{\frac{\log(np)}{n-h}}\right)|\boldsymbol{\beta}^{\natural}_{\mathsf{imax}}| \lesssim |\boldsymbol{\beta}^{\natural}_{\mathsf{imax}}|;\\
|\zeta_1^{(\max)}| &\ge \left(1-c_0\sqrt{\frac{\log(np)}{n-h}}\right)|\boldsymbol{\beta}^{\natural}_{\max}| \gtrsim |\boldsymbol{\beta}^{\natural}_{\max}|.
\end{aligned} \tag{23}
$$

Then we turn to study the rest of the terms. Conditional on $\mathcal{E}_2(\boldsymbol{\beta}^{\natural})\bigcap\mathcal{E}_3(\boldsymbol{\beta}^{\natural})\bigcap\mathcal{E}_4$, we have

$$
\begin{aligned}
|\zeta_2^{(\mathsf{imax})}|, |\zeta_2^{(\max)}| &\lesssim \sqrt{\frac{\log(mnp)}{n-h}}\|\boldsymbol{\beta}^{\natural}\|_2;\\
|\zeta_3^{(\mathsf{imax})}|, |\zeta_3^{(\max)}| &\lesssim \frac{\log(mnp)\sqrt{h}}{n-h}\|\boldsymbol{\beta}^{\natural}\|_2;
\end{aligned}
$$

$$|\zeta_4^{(\mathsf{imax})}|, |\zeta_4^{(\max)}| \lesssim \frac{\sigma\sqrt{n}\log(mnp)}{n-h}. \tag{24}$$

Combining (22), (23), and (24) then yields the lower-bound for $|\boldsymbol{\beta}^{\natural}_{\mathsf{imax}}|$

$$\begin{aligned} |\boldsymbol{\beta}^{\natural}_{\mathsf{imax}}| &\gtrsim |\zeta_1^{(\mathsf{imax})}| \geq |\zeta_1^{(\max)}| - \sum_{j=2}^{4}\left(|\zeta_j^{(\mathsf{imax})}| + |\zeta_j^{(\max)}|\right) \\ &\gtrsim |\boldsymbol{\beta}^{\natural}_{\max}| - c_2\sqrt{\frac{\log(mnp)}{n-h}}\|\boldsymbol{\beta}^{\natural}\|_2 - \frac{c_3\log(mnp)\sqrt{h}}{n-h}\|\boldsymbol{\beta}^{\natural}\|_2 - \frac{c_4\sigma\sqrt{n}\log(mnp)}{n-h}, \end{aligned}$$

which concludes the proof as $n \gtrsim h$.

□

Lemma 12. *Conditional on the intersection of events $\mathcal{E}_1 \bigcap \mathcal{E}_2(\boldsymbol{\beta}^{\natural}) \bigcap \mathcal{E}_3(\boldsymbol{\beta}^{\natural}) \bigcap \mathcal{E}_4$, we have*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\natural}\|_{\infty} \lesssim \frac{\log(mnp)}{\sqrt{n}}\|\boldsymbol{\beta}^{\natural}\|_2 + \frac{\sigma\log(mnp)}{\sqrt{n}}.$$

Proof. For an arbitrary index i, we consider the difference $\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}^{\natural}_i$, which can be written as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}^{\natural}_i = &\underbrace{\frac{1}{n-h}\sum_{\ell=\pi^{\natural}(i)}\left(\mathbf{X}_{\ell,i}^2 - 1\right)|\boldsymbol{\beta}^{\natural}_i|}_{\zeta_1} + \underbrace{\frac{1}{n-h}\sum_{\ell=\pi^{\natural}(\ell)}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\ell,:}, \boldsymbol{\beta}^{\natural}_{\backslash i}\right\rangle}_{\zeta_2} \\ &+ \underbrace{\frac{1}{n-h}\sum_{\ell=n-h+1}^{n}\mathbf{X}_{\ell,i}\left\langle \mathbf{X}_{\pi^{\natural}(\ell),:}, \boldsymbol{\beta}^{\natural}\right\rangle}_{\zeta_3} + \underbrace{\frac{1}{n-h}\sum_{\ell=1}^{n}\mathbf{X}_{\ell,i}^{\top}\mathbf{W}_{\ell,j}}_{\zeta_4}. \end{aligned}$$

Conditional on the intersection of events $\mathcal{E}_1 \bigcap \mathcal{E}_2(\boldsymbol{\beta}^{\natural}) \bigcap \mathcal{E}_3(\boldsymbol{\beta}^{\natural}) \bigcap \mathcal{E}_4$, we can bound ϑ_i $(1 \leq i \leq 4)$, as

$$\begin{aligned} |\zeta_1| &\lesssim \sqrt{\frac{\log(np)}{n-h}}|\boldsymbol{\beta}^{\natural}_i|; \\ |\zeta_2| &\lesssim \sqrt{\frac{\log(mnp)}{n-h}}\|\boldsymbol{\beta}^{\natural}_{\backslash i}\|_2; \\ |\zeta_3| &\lesssim \frac{\log(mnp)\sqrt{h}}{n-h}\|\boldsymbol{\beta}^{\natural}\|_2; \\ |\zeta_4| &\lesssim \frac{\sigma\sqrt{n}\log(mnp)}{n-h}, \end{aligned}$$

respectively, and complete the proof as $|\zeta_1| + |\zeta_2| \lesssim \sqrt{\frac{\log(mnp)}{n-h}}\|\boldsymbol{\beta}^{\natural}\|_2$ and $h \leq n$. □

Lemma 13. *Conditional on the intersection of events $\mathcal{E}_1 \bigcap_{\ell=1}^{m} \mathcal{E}_2(\mathbf{B}^{\natural}_{\ell,:}) \bigcap_{\ell=1}^{m} \mathcal{E}_3(\mathbf{B}^{\natural}_{\ell,:}) \bigcap \mathcal{E}_4$, we have*

$$\langle \widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}_{\mathsf{imax}}, \mathbf{B}^{\natural}\rangle \geq \frac{|||\mathbf{B}^{\natural}|||_{\mathrm{F}}^2}{k} - \frac{m\sigma^2\left(\log mnp\right)^2}{n} - \frac{\sqrt{m}\sigma\log(mnp)}{\sqrt{n}}|||\mathbf{B}^{\natural}|||_{\mathrm{F}},$$

provided $n \gg k(\log mnp)^2$.

Proof. First, we pick one arbitrary column $\boldsymbol{\beta}^{\natural}$ of $\mathbf{B}^{\natural}$. W.l.o.g. we assume that $\boldsymbol{\beta}^{\natural}_{\mathsf{imax}} \geq 0$. Then we obtain

$$\widehat{\boldsymbol{\beta}}_{\mathsf{imax}} \geq \boldsymbol{\beta}^{\natural}_{\mathsf{imax}} - \|\widehat{\boldsymbol{\beta}}^{(\pi^{\natural}(i),j)} - \boldsymbol{\beta}^{\natural}\|_{\infty},$$

where $\widehat{\boldsymbol{\beta}}^{(\pi^{\natural}(i),j)}$ denotes the corresponding column in $\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}$. Then we obtain the following lower bound on $\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}\boldsymbol{\beta}^{\natural}_{\mathsf{imax}}$

$$\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}\boldsymbol{\beta}^{\natural}_{\mathsf{imax}} \geq (\boldsymbol{\beta}^{\natural}_{\mathsf{imax}})^2 - |\boldsymbol{\beta}^{\natural}_{\mathsf{imax}}| \|\widehat{\boldsymbol{\beta}}^{(\pi^{\natural}(i),j)} - \boldsymbol{\beta}^{\natural}\|_{\infty}. \tag{25}$$

Similarly, we can show (25) holds as well when $\boldsymbol{\beta}^{\natural}_{\mathsf{imax}} < 0$. Recalling the definition of $\boldsymbol{\beta}^{\natural}_{\max}$, to put more specifically $|\boldsymbol{\beta}^{\natural}_{\max}| \geq |\boldsymbol{\beta}^{\natural}_{i}|$ $(1 \leq i \leq p)$, we can further lower bound $\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}\boldsymbol{\beta}^{\natural}_{\mathsf{imax}}$ by $(\boldsymbol{\beta}^{\natural}_{\mathsf{imax}})^2 - |\boldsymbol{\beta}^{\natural}_{\max}| \|\widehat{\boldsymbol{\beta}}^{(\pi^{\natural}(i),j)} - \boldsymbol{\beta}^{\natural}\|_{\infty}$.

For $|\boldsymbol{\beta}^{\natural}_{\mathsf{imax}}|$, we can lower bounded it by Lemma 11. While for $\|\widehat{\boldsymbol{\beta}}^{(\pi^{\natural}(i),j)} - \boldsymbol{\beta}^{\natural}\|_{\infty}$, we cannot directly use Lemma 12 since, strictly speaking, it concerns $\mathbf{X}$ with rows $\mathbf{X}_{\pi^{\natural}(i),:}, \mathbf{X}_{j,:}$ rather than $\widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}, \widetilde{\mathbf{X}}_{j,:}$ However since they follow the same distributions, we can follow the same procedure and show

$$\|\widehat{\boldsymbol{\beta}}^{(\pi^{\natural}(i),j)} - \boldsymbol{\beta}^{\natural}\|_{\infty} \lesssim \frac{\log(mnp)}{\sqrt{n}}\|\boldsymbol{\beta}^{\natural}\|_2 + \frac{\sigma\log(mnp)}{\sqrt{n}}.$$

Then we obtain

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}\boldsymbol{\beta}^{\natural}_{\mathsf{imax}} &\gtrsim \left(|\boldsymbol{\beta}^{\natural}_{\max}| - \frac{\sqrt{\log(mnp)}}{\sqrt{n}}\|\boldsymbol{\beta}^{\natural}\|_2 - \frac{\sigma\log(mnp)}{\sqrt{n}}\right)^2 - |\boldsymbol{\beta}^{\natural}_{\max}|\left(\frac{\log(mnp)}{\sqrt{n}}\|\boldsymbol{\beta}^{\natural}\|_2 + \frac{\sigma\log(mnp)}{\sqrt{n}}\right) \\ &\overset{①}{\gtrsim} |\boldsymbol{\beta}^{\natural}_{\max}|^2 - |\boldsymbol{\beta}^{\natural}_{\max}|\left(\frac{\log(mnp)}{\sqrt{n}}\|\boldsymbol{\beta}^{\natural}\|_2 + \frac{\sigma\log(mnp)}{\sqrt{n}}\right) - \frac{\log(mnp)}{n}\left(\|\boldsymbol{\beta}^{\natural}\|_2^2 + \sigma^2\cdot\log(mnp)\right),\end{aligned}$$

where in ① we use the relation $(a-b)^2 \geq {}^{a^2}\!/2 - b^2$. Under the assumption $n \geq k(\log mnp)^2$, we have $|\boldsymbol{\beta}^{\natural}_{\max}| \geq {}^{\|\boldsymbol{\beta}^{\natural}\|_2}\!/\sqrt{k} \gg {}^{\log(mnp)\|\boldsymbol{\beta}^{\natural}\|_2}\!/(2\cdot\sqrt{n})$ and thus

$$|\boldsymbol{\beta}^{\natural}_{\max}|^2 - |\boldsymbol{\beta}^{\natural}_{\max}|\cdot\frac{\log(mnp)}{\sqrt{n}}\|\boldsymbol{\beta}^{\natural}\|_2 \overset{②}{\geq} \left(\frac{1}{k} - \frac{\log(mnp)}{\sqrt{nk}}\right)\|\boldsymbol{\beta}^{\natural}\|_2^2,$$

where in ② we use the fact that $x^2 - 2ax$ is monotonically increasing in the region $[a, \infty)$; and the equality is achieved when $|\boldsymbol{\beta}^{\natural}_{\max}| = {}^{\|\boldsymbol{\beta}^{\natural}\|_2}\!/\sqrt{k}$. Hence, we obtain

$$\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}\boldsymbol{\beta}^{\natural}_{\mathsf{imax}} \gtrsim \left(\frac{1}{k} - \frac{\log(mnp)}{\sqrt{nk}} - \frac{\log(mnp)}{n}\right)\|\boldsymbol{\beta}^{\natural}\|_2^2 - \frac{\sigma\log(mnp)}{\sqrt{n}}\|\boldsymbol{\beta}^{\natural}\|_2 - \frac{\sigma^2(\log mnp)^2}{n}.$$

Having obtained the lower bound for one single column of $\mathbf{B}^{\natural}$, we complete the proof as

$$\begin{aligned}\langle\widehat{\mathbf{B}}^{(\pi^{\natural}(i),j)}_{\mathsf{imax}}, \mathbf{B}^{\natural}\rangle &= \sum_{\boldsymbol{\beta}^{\natural}}\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}\boldsymbol{\beta}^{\natural}_{\mathsf{imax}} \\ &\overset{③}{\gtrsim} \underbrace{\left(\frac{1}{k} - \frac{\log(mnp)}{\sqrt{nk}} - \frac{\log(mnp)}{n}\right)}_{\asymp\, k^{-1} \text{ since } n \gg k\log^2(mnp)}\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^2 - \frac{\sqrt{m}\sigma\log(mnp)}{\sqrt{n}}\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}} - \frac{m\sigma^2(\log mnp)^2}{n}\end{aligned}$$

where ③ comes from a reorganization of the terms and the inequality $\sum_{\boldsymbol{\beta}^{\natural}}\|\boldsymbol{\beta}^{\natural}\|_2 \leq \sqrt{m}\left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}$. □

Lemma 14. *We have* $\mathbb{P}(\mathcal{E}_5) \geq 1 - 4n^{-c}p^{-c} - 2n^{-c'}m^{-c'}$.

Proof. We begin the proof with the union bound

$$\mathbb{P}\left(\overline{\mathcal{E}}_5\right) \leq \underbrace{\mathbb{P}\left(|x_i| \gtrsim \sqrt{\log np},\ \exists\, 1 \leq i \leq p\right)}_{\leq\, 2np\cdot n^{-c}p^{-c} = 2n^{-c'}p^{-c'}} + \underbrace{\mathbb{P}\left(|\langle\boldsymbol{x}, \boldsymbol{\beta}^{\natural}\rangle| \gtrsim \sqrt{\log np}\|\boldsymbol{\beta}^{\natural}\|_2\right)}_{\leq\, 2n\cdot n^{-c}p^{-c} = 2n^{-c'}p^{-c''}}$$

$$
+ \underbrace{\mathbb{P}\left(|\boldsymbol{w}_i| \gtrsim \sigma\sqrt{\log nm},\ 1 \le i \le m\right)}_{\le 2mn\cdot n^{-c}m^{-c} = 2n^{-c'}m^{-c'}}
+ \underbrace{\mathbb{P}\left(\overline{\mathcal{E}}_5,\ \|\boldsymbol{x}\|_{\infty} \lesssim \sqrt{\log np},\ |\langle \boldsymbol{x}, \boldsymbol{\beta}^{\natural}\rangle| \lesssim \sqrt{\log np}\|\boldsymbol{\beta}^{\natural}\|_2,\ \|\boldsymbol{w}\|_{\infty} \lesssim \sigma\sqrt{\log nm}\right)}_{\triangleq \vartheta},
$$

where $\boldsymbol{w}_{(\cdot)}$ and $\widetilde{\boldsymbol{w}}_{(\cdot)}$ denote the corresponding entries from $\mathbf{W}$ and $\widetilde{\mathbf{W}}$.

Then we would prove that ϑ is zero. The technical details are attached in the following. Due to the fact that $\widehat{\mathbf{B}}^{(i,j)}_{\mathsf{imax}}$ shares the same support set as $\widehat{\mathbf{B}}_{\mathsf{imax}}$, we have

$$
\begin{aligned}
\left\|\left(\widehat{\mathbf{B}}_{\mathsf{imax}} - \widehat{\mathbf{B}}^{(i,j)}_{\mathsf{imax}}\right)^{\top}\boldsymbol{x}\right\|_2^2 &= \sum_{\boldsymbol{\beta}^{\natural}}\left[\boldsymbol{x}_{\mathsf{imax}}\left(\widehat{\boldsymbol{\beta}}_{\mathsf{imax}} - \widehat{\boldsymbol{\beta}}^{\backslash(i,j)}_{\mathsf{imax}}\right)\right]^2 \overset{①}{\lesssim} (\log np)\sum_{\boldsymbol{\beta}^{\natural}}\left\|\widehat{\boldsymbol{\beta}}_{\mathsf{imax}} - \widehat{\boldsymbol{\beta}}^{\backslash(i,j)}_{\mathsf{imax}}\right\|_2^2 \\
&= (\log np)\sum_{\boldsymbol{\beta}^{\natural}}\left\|\widehat{\boldsymbol{\beta}}_{\mathsf{imax}} - \widehat{\boldsymbol{\beta}}^{\backslash(i,j)}_{\mathsf{imax}}\right\|_{\infty}^2,
\end{aligned}
$$

where in ① we condition on the relation $\|\boldsymbol{x}\|_{\infty} \lesssim \sqrt{\log np}$.

Regarding $\|\widehat{\boldsymbol{\beta}}_{\mathsf{imax}} - \widehat{\boldsymbol{\beta}}^{\backslash(i,j)}_{\mathsf{imax}}\|_{\infty}$, we have $\|\widehat{\boldsymbol{\beta}}_{\mathsf{imax}} - \widehat{\boldsymbol{\beta}}^{\backslash(i,j)}_{\mathsf{imax}}\|_{\infty} \le \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{\backslash(i,j)}\|_{\infty}$. Again, this fact relies heavily on the fact such that $\widehat{\mathbf{B}}^{(i,j)}_{\mathsf{imax}}$ shares the same support set as $\widehat{\mathbf{B}}_{\mathsf{imax}}$. Otherwise, the best result we can have is $\|\widehat{\boldsymbol{\beta}}_{\mathsf{imax}} - \widehat{\boldsymbol{\beta}}^{\backslash(i,j)}_{\mathsf{imax}}\|_{\infty} \le \max\left(\|\widehat{\boldsymbol{\beta}}_{\mathsf{imax}}\|_{\infty}, \|\widehat{\boldsymbol{\beta}}^{\backslash(i,j)}_{\mathsf{imax}}\|_{\infty}\right)$. Afterward, we obtain

$$
\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_{\mathsf{imax}} - \widehat{\boldsymbol{\beta}}^{\backslash(\pi^{\natural}(i),j)}_{\mathsf{imax}}\|_{\infty} \le\ & (n-h)^{-1}\left\|\left(\mathbf{X}_{i,:}\mathbf{X}_{i,:}^{\top} + \mathbf{X}_{j,:}\mathbf{X}_{j,:}^{\top} - \widetilde{\mathbf{X}}_{i,:}\widetilde{\mathbf{X}}_{i,:}^{\top} - \widetilde{\mathbf{X}}_{j,:}\widetilde{\mathbf{X}}_{j,:}^{\top}\right)\boldsymbol{\beta}^{\natural}\right\|_{\infty} \\
& + (n-h)^{-1}\left\|\mathbf{X}_{\pi^{\natural}(i),:}\boldsymbol{w}_{\pi^{\natural}(i)} + \mathbf{X}_{i,:}\boldsymbol{w}_i - \widetilde{\mathbf{X}}_{\pi^{\natural}(i),:}\widetilde{\boldsymbol{w}}_{\pi^{\natural}(i)} - \widetilde{\mathbf{X}}_{i,:}\widetilde{\boldsymbol{w}}_i\right\|_{\infty} \\
\lesssim\ & (n-h)^{-1}\|\boldsymbol{x}\langle\boldsymbol{x}, \boldsymbol{\beta}^{\natural}\rangle\|_{\infty} + (n-h)^{-1}\|\boldsymbol{x}\boldsymbol{w}\|_{\infty} \\
\lesssim\ & (n-h)^{-1}\|\boldsymbol{x}\|_{\infty}|\langle\boldsymbol{x}, \boldsymbol{\beta}^{\natural}\rangle| + (n-h)^{-1}\|\boldsymbol{x}\|_{\infty}\|\boldsymbol{w}\|_{\infty} \\
\overset{②}{\lesssim}\ & (n-h)^{-1}(\log np)\|\boldsymbol{\beta}^{\natural}\|_2 + (n-h)^{-1}\sigma\sqrt{(\log np)(\log mn)},
\end{aligned}
$$

where in ② we use the condition $\|\boldsymbol{x}\|_{\infty} \lesssim \sqrt{\log np}$, $|\langle\boldsymbol{x}, \boldsymbol{\beta}^{\natural}\rangle| \lesssim \sqrt{\log np}\|\boldsymbol{\beta}^{\natural}\|_2$, and $\|\boldsymbol{w}\|_{\infty} \lesssim \sigma\sqrt{\log nm}$.

Iterating over all columns $\mathbf{B}^{\natural}$, we can show $\mathcal{E}_5$ holds with probability one provided $\|\boldsymbol{x}\|_{\infty} \lesssim \sqrt{\log np}$, $|\langle\boldsymbol{x}, \boldsymbol{\beta}^{\natural}\rangle| \lesssim \sqrt{\log np}\|\boldsymbol{\beta}^{\natural}\|_2$, and $\|\boldsymbol{w}\|_{\infty} \lesssim \sigma\sqrt{\log nm}$, in other words, ϑ is zero. □

C. Useful Facts

This section collects some useful facts about probability for the sake of self-containing.

Lemma 15 (Dasgupta and Gupta (2003))**.** *For a χ^2-random variable Z, which is with freedom ℓ, we conclude*

$$
\begin{aligned}
\mathbb{P}(Z \le t) &\le \exp\left[\frac{\ell}{2}\left(\frac{t}{\ell} - \frac{t}{\ell} + 1\right)\right],\ \ t < \ell; \\
\mathbb{P}(Z \ge t) &\le \exp\left[\frac{\ell}{2}\left(\frac{t}{\ell} - \frac{t}{\ell} + 1\right)\right],\ \ t > \ell.
\end{aligned}
$$

Lemma 16 (Pananjady et al. (2018))**.** *Suppose the permutation matrix $\mathbf{\Pi}$ with Hamming distance h from the identity matrix $\mathbf{I}$, namely, $\mathsf{d}_{\mathsf{H}}(\mathbf{I}, \mathbf{\Pi}) = h$. We can decompose the index set $\{i :\ i \ne \pi(i)\}$ into 3 independent sets $\mathcal{I}_i$ $(1 \le i \le 3)$ such that the cardinality of each set satisfies $|\mathcal{I}_i| \ge \lfloor h/3 \rfloor \ge h/5$.*