# Optimal Estimator for Unlabeled Linear Regression

**Hang Zhang, Ping Li**
**Cognitive Computing Lab**
**Baidu Research**
**10900 NE 8th ST. Bellevue, WA 98004, USA**
**{zhanghanghitomi, pingli98}@gmail.com**

## Abstract

Unlabeled linear regression, or "linear regression with an unknown permutation", has attracted increasing attentions due to its applications in (e.g.,) linkage record and de-anonymization. However, the computation of unlabeled linear regression proves to be cumbersome and existing algorithms typically require considerable time, especially in the high dimensional regime. In this paper, we propose a one-step estimator which is optimal from both the computational and the statistical aspects. From the computational perspective, our estimator exhibits the same order of computational complexity as that of the oracle case (which means the regression coefficients are known in advance and only the permutation needs recovery). From the statistical perspective, when comparing with the necessary conditions for permutation recovery, our requirement on the *signal-to-noise ratio* (SNR) agrees up to merely $\Omega(\log\log n)$ difference when the stable rank of the regression coefficients $\mathbf{B}^\natural$ is much less than $\log n/\log\log n$. Numerical experiments are also provided to corroborate the theoretical claims.

## 1. Introduction

This paper studies the problem of unlabeled linear regression, with the sensing relation being written as

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

where $\mathbf{\Pi}^\natural \in \mathbb{R}^{n\times n}$ denotes the unknown permutation matrix, $\mathbf{X} \in \mathbb{R}^{n\times p}$ represents the design (sensing) matrix, $\mathbf{B}^\natural \in \mathbb{R}^{p\times m}$ presents the matrix of regression coefficients, $\mathbf{W} \in \mathbb{R}^{n\times m}$ is the additive noise, and $\mathbf{Y} \in \mathbb{R}^{n\times m}$ denotes the matrix of measurements. When the permutation matrix $\mathbf{\Pi}^\natural$ is known in advance, the model in Eq. (1) becomes the standard linear regression problem.

The study on unlabeled linear regression can be traced back to 1970s under the name "broken sample problem" (DeGroot et al., 1971; Goel, 1975; DeGroot and Goel, 1976; 1980; Chan and Loh, 2001; Bai and Hsing, 2005; Slawski et al., 2019a). To the best of our knowledge, the special term "unlabeled sensing" or "unlabeled linear regression" initially appeared in Unnikrishnan et al. (2015), which study the single observation model, i.e., $m = 1$, under the noiseless setting, namely, $\mathbf{W} = \mathbf{0} \in \mathbb{R}^n$. Assuming the entries of $\mathbf{X}$ come from a continuous distribution, Unnikrishnan et al. (2015) establish the necessary condition $n \geq 2p$ for the correct recovery. For this setting, similar results have also been discovered by Tsakiris (2018); Dokmanic (2019) but with different approaches.

Since the work of Unnikrishnan et al. (2015), a variety of estimation algorithms for solving Eq. (1) have been proposed and/or analyzed (Pananjady et al., 2017a; Abid et al., 2017; Hsu et al., 2017; Pananjady et al., 2017b; Slawski et al., 2019b; Slawski and Ben-David, 2019; Slawski et al., 2019a; Tsakiris and Peng, 2019; Zhang et al., 2019a;b). Those estimation methods, however, typically come with high computational complexity.

For example, in Pananjady et al. (2017a), the authors demonstrate that the *maximum likelihood* (ML) estimator of $\mathbf{\Pi}^\natural$ is NP-hard in general and no practical estimator is proposed. In the follow-up work (Pananjady et al., 2017b), instead of recovering $\mathbf{\Pi}^\natural$, the authors focus only on obtaining the product $\mathbf{\Pi}^\natural\mathbf{X}\mathbf{B}^\natural$. In Hsu et al. (2017); Abid et al. (2017), they both consider the single observation case ($m = 1$) but with the theoretical analysis focusing on $\|\mathbf{B}^\natural - \widehat{\mathbf{B}}\|_{\mathrm{F}}$. In Tsakiris and Peng (2019), an abstract view of unlabeled sensing is adopted and a branch-and-bound algorithm is proposed. In Zhang et al. (2019a;b), the authors consider the multiple observations setting (that is, $m > 1$). They first give the statistical lower bound for this scenario and prove that the requirement of SNR,

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020. 

which matches the order of the above bound, can drop drastically for correct permutation recovery. A heuristic estimator based on the *alternating direction method of multipliers* (ADMM) is proposed but without the performance guarantee. In Slawski and Ben-David (2019); Slawski et al. (2019a;b), the authors place a parsimonious constraint on the number of permuted rows and view the product $(\mathbf{I} - \mathbf{\Pi}^\natural)\mathbf{X}\mathbf{B}^\natural$ as the sparse outliers. Although their proposed estimators are computable with the performance guarantee, they require multiple rounds of iterations and typically only allow a small proportion of rows being permuted.

Note that, in the context of permutation recovery, no previous studies considered model sparsity, i.e., $\mathbf{B}^\natural$ is sparse, until very recently (Zhang and Li, 2020). The model sparsity problem is a challenging and interesting research topic.

### 1.1. Practical applications

In the past decade or so, one has witnessed a renaissance of the problem of unlabeled linear regression due to its wide applications in, for example, data integration, privacy protection, computer vision, sensor networks, robotics, etc. (Unnikrishnan et al., 2015; Pananjady et al., 2017a;b; Slawski and Ben-David, 2019; Slawski et al., 2019a). Here we would like to elaborate on three of the most important applications, namely *linkage record*, *de-anonymization*, and *header-free communication*.

In linkage record application (Winkler, 1995), one is interested in integrating multiple databases, where each database contains different pieces of information about the same identity, into one comprehensive database. In this process, the biggest challenge is how to find the matching across different databases. For de-anonymization (Nazarov et al., 2018), the task is to identify the hidden labels, which aims to preserve privacy, with public data. It can be seen as the inverse problem of privacy protection. One mathematical formulation is viewing the correspondence between the hidden labels and the public data as the unknown permutation matrix. For the application in header-free communication (Pananjady et al., 2017b), we have a sensor network where the sensor identity is omitted during communication to reduce the transmission cost and latency. In this scenario, reconstruction of the signal involves recovering the unknown correspondence. The above three applications are merely selected practical examples for using unlabelled linear regression.

### 1.2. Summary of contributions

Before describing the contributions of this paper, we first define the notation *signal-to-noise-ratio* (SNR) as

$$\mathsf{SNR} = \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2/(m\sigma^2). \tag{2}$$

Our contributions can be elaborated from two aspects:

Firstly, we propose a simple one-step estimator for the exact permutation recovery. In the previous works such as Pananjady et al. (2017a); Slawski and Ben-David (2019); Slawski et al. (2019a); Zhang et al. (2019a;b), the computation and analysis of estimators are largely parallel. In comparison, our estimator effectively exploits the sensing matrix's statistical properties in designing the estimator. By assuming $\mathbf{X}$ to be a Gaussian matrix, we propose to obtain an approximation of $c\mathbf{B}^\natural$ by the product $\mathbf{X}^\top\mathbf{Y}$, where $c$ is a non-negative scalar. Similar ideas have been used previously in picking the initialization points for the non-convex optimizations as in Candès et al. (2015); Balakrishnan et al. (2017); Chi et al. (2019). A detailed discussion on the similarities and differences between our estimator and their work can be found in Section 2.

As the second aspect of the contribution, we prove our proposed estimator achieves optimal performance in certain regime. First we show our estimator gets the ground truth $\mathbf{\Pi}^\natural$ provided $\log(\mathsf{SNR}) \gtrsim \log n$ under the special case $m = 1$, $p = 1$. This bound matches that of the statistical limit. Moreover, we consider the general setting when $m \gg 1$. Equipped with the *leave-one-out* trick (El Karoui, 2013; El Karoui et al., 2013; El Karoui, 2018; Chen et al., 2019; Sur et al., 2019), we are able to reduce the SNR requirement for correct $\mathbf{\Pi}^\natural$ to $\log(\mathsf{SNR}) \gtrsim \log\log n + \log n/\rho(\mathbf{B}^\natural)$, where $\rho(\cdot)$ denotes the stable rank and will be explained later. Meanwhile the SNR should be at least $\log(\mathsf{SNR}) \gtrsim \log n/\rho(\mathbf{B}^\natural)$ to avoid wrongly recovered $\mathbf{\Pi}$. When $\rho(\mathbf{B}^\natural) \ll \log n/\log\log n$, our estimator is optimal and gives the same order, namely $\log n/\rho(\mathbf{B}^\natural)$, with the difference only up to some multiplicative constants. Otherwise, our estimator may experience some performance loss, which is at most $\Omega(\log\log n)$. Numerical experiments are provided to corroborate our claim as well.

### 1.3. 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 $\mathbf{X}$, we denote $\mathbf{X}_{i,:}$ as its $i$-th row, $\mathbf{X}_{:,i}$ as its $i$-th column, and $X_{ij}$ as its $(i,j)$-th element. The Frobenius norm of $\mathbf{X}$ is defined as $\|\mathbf{X}\|_{\mathrm{F}}$ while the operator norm is denoted as $\|\mathbf{X}\|_{\mathrm{OP}}$, whose definitions can be found in Section 2.3 of Golub and Loan (2013) (P71). Its stable rank $\rho(\mathbf{X})$ is defined as the ratio $\|\mathbf{X}\|_{\mathrm{F}}^2/\|\mathbf{X}\|_{\mathrm{OP}}^2$ (see Section 2.1.15 in Tropp (2015)).

Consider a 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 permutation matrix $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ is defined as $\mathsf{d_H}(\mathbf{\Pi}_1, \mathbf{\Pi}_2) = \sum_{i=1}^{n} \mathbb{1}(\pi_1(i) \neq \pi_2(i))$. Again, the SNR is defined as $\mathsf{SNR} = \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2/(m\sigma^2)$.

### 1.4. Outline

In Section 2, we present our one-step estimator and its design insight. Then we separately investigate its statistical properties under the single observation model ($m = 1$) and multiple observations model ($m > 1$). The corresponding discussions are put in Section 3 and Section 4, respectively. Simulation results are presented in Section 5 and the conclusions are drawn in Section 6.

## 2. Estimator Description

We begin this section with a formal description of the sensing model, which reads

$$\mathbf{Y} = \mathbf{\Pi}^\natural \mathbf{X} \mathbf{B}^\natural + \mathbf{W}, \tag{3}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times m}$ denotes the observation, $\mathbf{\Pi}^\natural \in \mathbb{R}^{n\times n}$ is the unknown permutation matrix such that $\sum_i \Pi^\natural_{i,j} = \sum_j \Pi^\natural_{i,j} = 1$, $\Pi^\natural_{i,j} \in \{0,1\}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$ denotes the sensing matrix with $X_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$ follows the standard normal. $\mathbf{B}^\natural \in \mathbb{R}^{p\times m}$ is the matrix of regression coefficients, and $\mathbf{W} \in \mathbb{R}^{n\times m}$ represents the additive Gaussian noise with each entry $W_{ij}$ follows a Gaussian distribution with zero mean and $\sigma^2$ variance, $W_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.

In this paper, we propose a one-step estimator to estimate $(\mathbf{\Pi}^\natural, \mathbf{B}^\natural)$ from $\mathbf{Y}$, as summarized in Algorithm 1.

**Algorithm 1** The one-step estimator proposed in this paper.

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

$$\widehat{\mathbf{\Pi}} = \operatorname{argmax}_{\mathbf{\Pi}\in\mathcal{P}_n} \left\langle \mathbf{\Pi}, \mathbf{Y}\mathbf{Y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle, \tag{4}$$

$$\widehat{\mathbf{B}} = (\mathbf{X})^\dagger \widehat{\mathbf{\Pi}}^\top \mathbf{Y}, \tag{5}$$

where $\mathbf{X}^\dagger = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the pseudo-inverse of $\mathbf{X}$ and $\mathcal{P}_n$ is the set of all possible permutation matrices.

In Algorithm 1, the optimization task $\operatorname{argmax}_{\mathbf{\Pi}\in\mathcal{P}_n} \langle \mathbf{\Pi}, \cdot \rangle$ can be solved as a linear assignment problem (Kuhn, 1955; Bertsekas and Castañón, 1992), while Eq. (5) is simply the traditional least-square estimator for the linear regression (Golub and Loan, 2013). Compared with the previous estimators as in Pananjady et al. (2017a); Slawski et al. (2019a); Zhang et al. (2019a;b), our estimator in Eq. (4) and Eq. (5) demonstrates advantages from both the computational and statistical perspectives.

### 2.1. Computational aspects

Firstly, we can see that the computational complexity of our estimator in Eq. (4) and Eq. (5) would be $\Omega(n^3 + np^2m)$, where $\Omega(n^3)$ is for computing $\widehat{\mathbf{\Pi}}$ and $\Omega(np^2m)$ for $\widehat{\mathbf{B}}$.

We consider two types of oracle estimators as the baselines. The first oracle case assumes $\mathbf{B}^\natural$ is known in advance. The optimal estimator to recover $\mathbf{\Pi}^\natural$ would then become

$$\widehat{\mathbf{\Pi}} = \operatorname{argmax}_{\mathbf{\Pi}\in\mathcal{P}_n} \left\langle \mathbf{\Pi}, \mathbf{Y}\mathbf{B}^{\natural\top}\mathbf{X}^\top \right\rangle, \tag{6}$$

whose computational complexity is $\Omega(n^3)$. Comparing with our estimator in Eq. (4), we notice that we only sacrifice one matrix multiplication, i.e., replacing $\mathbf{B}^\natural$ by the product $\mathbf{X}^\top\mathbf{Y}$. Since the computational bottleneck lies in solving the linear assignment problem (Kuhn, 1955; Bertsekas and Castañón, 1992), one additional matrix multiplication does not change the computational complexity, which is also of order $\Omega(n^3)$.

For the second type of oracle estimator, we consider $\mathbf{\Pi}^\natural$ is known. In this case, the sensing relation in Eq. (3) reduces to the classical multivariate linear regression, where the least square estimator has $\Omega(np^2m)$ computational complexity in order to estimate $\mathbf{B}^\natural$.

With the relation $\Omega(n^3 + np^2m) \lesssim \Omega(n^3) \vee \Omega(np^2m)$, we conclude that our estimator is computationally optimal since it has the same order as the oracle estimators.

### 2.2. Statistical limits

In Zhang et al. (2019a;b), it is shown that no estimator can recover the permutation matrix $\mathbf{\Pi}^\natural$ with high probability if the SNR satisfies

$$\log(\mathsf{SNR}) \lesssim \frac{\log n}{\rho(\mathbf{B}^\natural)}. \tag{7}$$

In this paper, we show that our estimator in Eq. (4) and Eq. (5) will generate the correct permutation matrix $\mathbf{\Pi}^\natural$ under certain regimes, with the SNR order satisfying

$$\log(\mathsf{SNR}) \gtrsim \frac{\log n}{\rho(\mathbf{B}^\natural)} + \log\log n,$$

which matches the lower bound in Eq. (7) with the difference up to $\Omega(\log\log n)$. Provided that $\rho(\mathbf{B}^\natural) \ll \frac{\log n}{\log\log n}$, we conclude Eq. (4) coincides with the statistical limits up to some multiplicative constants. The formal statement of theoretical result is presented as Theorem 2.

Furthermore, our estimator does not involve the noise variance $\sigma^2$, hence it is immune to the inaccurate estimation of $\sigma^2$. For a more comprehensive understanding, we compare our estimator with the previous results in Table 1. The detailed discussions are presented in Section 3 and Section 4.

*Table 1.* Comparison of 4 estimators: the ML estimator (Pananjady et al., 2017a; Zhang et al., 2019a;b), the ADMM estimator (Zhang et al., 2019a;b), the two-stage estimator (Slawski et al., 2019a), and ours. A question mark indicates the corresponding entry is uncertain, and an asterisk means that it is only correct under certain regime. $h$ denotes the number of permuted rows ($h = \mathsf{d_H}(\mathbf{I}, \mathbf{\Pi}^\natural)$).

| | Statistical Optimal | | Computational Cost | | Permuted Rows | |
|---|---|---|---|---|---|---|
| | $m = 1$ | $m \gg 1$ | $m = 1$ | $m \gg 1$ | $m = 1$ | $m \gg 1$ |
| ML Estimator | YES | YES (*) | $\Omega(1)$ (*) | $\gtrsim n!$ | $h \lesssim n$ | $h \lesssim \frac{n}{\log n}$ |
| ADMM Estimator | ? | NO | $\gg \Omega(1)$ | $\gg \Omega(1)$ | ? | ? |
| Two-Stage Estimator | ? | YES (*) | ? | $\gg \Omega(1)$ | $h \lesssim \frac{n}{\log(n/h)}$ | $h \lesssim \frac{n}{\log(n/h)}$ |
| **Our Estimator** | **YES** (*) | **YES (*)** | $\Omega(1)$ | $\Omega(1)$ | $h \lesssim n$ | $h \lesssim n$ |

### 2.3. Insights in designing our estimator

Before delving into analyzing the statistical properties of our proposed estimator in Algorithm 1, we would like to elaborate on some of the insights. First, we consider the maximum likelihood (ML) estimator, which is written as

$$(\widehat{\mathbf{\Pi}},\ \widehat{\mathbf{B}}) = \mathrm{argmin}_{\mathbf{\Pi},\mathbf{B}} \|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}\|_{\mathrm{F}}. \tag{8}$$

Note that the major difficulty of solving the ML estimator lies in the intervention of $\mathbf{\Pi}$ and $\mathbf{B}$. Provided either values of $\mathbf{\Pi}^\natural, \mathbf{B}^\natural$ is known, the ML estimator then becomes convex and can be easily solved. As it is not possible to directly access the true values of $\mathbf{\Pi}$ and $\mathbf{B}$, we resort to using approximate values instead. A similar idea is also adopted in Slawski et al. (2019a). Our estimator differs from Slawski et al. (2019a) in that they obtain the approximation of $\mathbf{B}^\natural$ via a group-Lasso-alike estimator while we instead propose to use $\mathbf{X}^\top\mathbf{Y}$.

In retrospect, the underlying logic for using $\mathbf{X}^\top\mathbf{Y}$ is actually not too surprising. First we notice the ML estimator in Eq. (8) is insensitive to the length $\|\mathbf{B}\|_{\mathrm{F}}$ since the same $\mathbf{\Pi}$ will be returned once $\mathbf{B}/\|\mathbf{B}\|_{\mathrm{F}}$ is fixed. Then we follow the same procedure as the initialization methods in Candès et al. (2015); Balakrishnan et al. (2017); Chi et al. (2019) and assume that the product $\mathbf{X}^\top\mathbf{Y}$ is close to its expectation, which is a scaled value of $\mathbf{B}^\natural$. Combing the above reasonings together yields our approximation.

Finally, to conclude this section, we would also like to emphasize that our approximation scheme is different from what is used in Candès et al. (2015); Balakrishnan et al. (2017); Chi et al. (2019): (i) their approximation method is only used for initialization while ours is to obtain the final result; (ii) their goal is to minimize the distance $\|\mathbf{B} - \mathbf{B}^\natural\|_{\mathrm{F}}$, which requires to estimate the length of $\mathbf{B}^\natural$, while our estimator does not need this value as we have explained. Thus, our proposed estimator avoids estimating the noise variance and the procedure is inherently robust.

## 3. Single Observation Model ($m = 1$)

This section considers the single observation model, namely $m = 1$. We will separately discuss the estimator's performance under the cases where $p = 1$ and $p > 1$.

### 3.1. A warm-up example: $m = p = 1$

First we consider the warm-up example when $m = p = 1$. To distinguish this case with the multiple observations model, i.e., $m > 1$, we rewrite the sensing relation as

$$\mathbf{y} = \mathbf{\Pi}^\natural\mathbf{X}\beta + \mathbf{w}, \tag{9}$$

where $\mathbf{X} \in \mathbb{R}^n$ reduces to a vector while $\beta \in \mathbb{R}$ is a scalar. Then we show the estimator has the following property,

**Theorem 1** *Provided that the Hamming distance* $h = \mathsf{d_H}(\mathbf{I}, \mathbf{\Pi}^\natural) \le \frac{n}{4}$, $n \ge 2p$, *if* $\mathsf{SNR}$ *satisfies*

$$\log(\mathsf{SNR}) \ge c_0 \log(n),$$

*then the estimator in Eq. (4) recovers the correct permutation matrix, i.e.,* $\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^\natural$, *with probability exceeding* $1 - c_1 n^{-1}$ *when* $n$ *is sufficiently large, where* $c_0, c_1$ *are some positive constants.*

According to Theorem 2 in Pananjady et al. (2017a), which is restated as Theorem 5, correct recovery of permutation matrix $\mathbf{\Pi}^\natural$ requires $\log(\mathsf{SNR}) \gtrsim \log n$ at least. As our estimator matches the statistical limits with the difference up to some multiplicative constants, we conclude its tightness.

The proof of Theorem 1 is deferred to the supplementary material. Here we give an intuitive explanation. First we consider the noiseless case, where $\mathbf{w} = \mathbf{0}$ and $\mathsf{SNR}$ is infinite. We can expand the inner product $\langle \mathbf{\Pi}, \mathbf{Y}\mathbf{Y}^\top\mathbf{X}\mathbf{X}^\top \rangle$, after some algebra, as

$$\langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top\mathbf{X}\mathbf{X}^\top \rangle = \beta^2 \langle \mathbf{X}, \mathbf{\Pi}^\natural\mathbf{X} \rangle \langle \mathbf{\Pi}\mathbf{X}, \mathbf{\Pi}^\natural\mathbf{X} \rangle,$$

Given that the term $\beta^2\langle \mathbf{X}, \mathbf{\Pi}^\natural \mathbf{X}\rangle$ concentrates around its expectation $\beta^2 \mathbb{E}_{\mathbf{X}}\langle \mathbf{X}, \mathbf{\Pi}^\natural \mathbf{X}\rangle = (n-h)\beta^2 > 0$ with high probability, we conclude that the maximum is reached when $\mathbf{\Pi} = \text{argmax}\langle \mathbf{\Pi X}, \mathbf{\Pi}^\natural \mathbf{X}\rangle$, namely, $\mathbf{\Pi} = \mathbf{\Pi}^\natural$ since $\|\mathbf{\Pi}^\natural \mathbf{X}\|_{\mathrm{F}}^2 = \|\mathbf{\Pi X}\|_{\mathrm{F}} \|\mathbf{\Pi}^\natural \mathbf{X}\|_{\mathrm{F}} \geq \langle \mathbf{\Pi X}, \mathbf{\Pi}^\natural \mathbf{X}\rangle$ (Cauchy-Schwarz inequality).

For the noisy case, we can interpret the observation $\mathbf{y}$ as some perturbed version of the product $\mathbf{\Pi}^\natural \mathbf{X}\beta$. Specifically, we define $\mathcal{T}_i(\mathbf{\Pi})$, $(1 \leq i \leq 3)$ as

$$\mathcal{T}_1(\mathbf{\Pi}) = \langle \mathbf{w}, \mathbf{\Pi}^\top \mathbf{X}\rangle \langle \mathbf{X}, \mathbf{w}\rangle;$$
$$\mathcal{T}_2(\mathbf{\Pi}) = \langle \mathbf{w}, \mathbf{X}\rangle \langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{\Pi}^\top \mathbf{X}\rangle + \langle \mathbf{w}, \mathbf{\Pi}^\top \mathbf{X}\rangle \langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{X}\rangle;$$
$$\mathcal{T}_3(\mathbf{\Pi}) = \langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{\Pi X}\rangle \langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{X}\rangle,$$

where $\mathcal{T}_1(\mathbf{\Pi})$ and $\mathcal{T}_2(\mathbf{\Pi})$ correspond to the perturbation incurred by the noise $\mathbf{w}$. Then we have

$$\begin{aligned}
&\langle \mathbf{\Pi}^\natural, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top\rangle - \langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top\rangle \\
&= \mathcal{T}_1(\mathbf{\Pi}^\natural) - \mathcal{T}_1(\mathbf{\Pi}) + \beta\left[\mathcal{T}_2\left(\mathbf{\Pi}^\natural\right) - \mathcal{T}_2(\mathbf{\Pi})\right] \\
&+ \beta^2\left[\mathcal{T}_3\left(\mathbf{\Pi}^\natural\right) - \mathcal{T}_3(\mathbf{\Pi})\right].
\end{aligned}$$

The gist is to prove the perturbation is significantly small, namely, $\mathcal{T}_3(\mathbf{\Pi}^\natural) - \mathcal{T}_3(\mathbf{\Pi})$ is large while $\left|\mathcal{T}_i(\mathbf{\Pi}^\natural) - \mathcal{T}_i(\mathbf{\Pi})\right|$, $(1 \leq i \leq 2)$ is small. We can show

$$\begin{aligned}
&\langle \mathbf{\Pi}^\natural, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top\rangle - \langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top\rangle \\
&\gtrsim \frac{c_0\beta^2}{n^{19}} - c_1\beta\sigma n^2\sqrt{\log n} - c_2\sigma^2 n^2 \log n,
\end{aligned}$$

holds with high probability. Provided $\log(\mathsf{SNR}) \gtrsim \log n$, we can show that

$$\langle \mathbf{\Pi}^\natural, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top\rangle > \langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top\rangle,$$

which completes the proof.

### 3.2. General case: $m = 1,\ p > 1$

For this scenario, we first consider the case where the direction of $\mathbf{B}^\natural$ is known, i.e., $\mathbf{e} = \mathbf{B}^\natural / \|\mathbf{B}^\natural\|_{\mathrm{F}}$. Construct an orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{p\times p}$ with the Gram-Schmidt process (Golub and Loan, 2013) whose first column is $\mathbf{e}$. We can then rewrite Eq. (3) as

$$\mathbf{Y} = \mathbf{\Pi}^\natural (\mathbf{XQ})_{:,1} \|\mathbf{B}^\natural\|_{\mathrm{F}} + \mathbf{W}.$$

Due to the rotation invariance of Gaussian distribution, we have $(\mathbf{XQ})_{:,1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$ and restore it to the warm-up example in Eq. (9), where $\beta$ is replaced by the length $\|\mathbf{B}^\natural\|_{\mathrm{F}}$. Hence we can obtain the correct permutation matrix $\mathbf{\Pi}^\natural$ once $\log(\mathsf{SNR}) \gtrsim \log n$, as illustrated in Theorem 1.

Apart from the above case (i.e., when $\mathbf{e}$ is known), our estimator cannot ensure correct recovery of permutation even under the noiseless case. A numerical example is given in Figure 1 for an illustration.

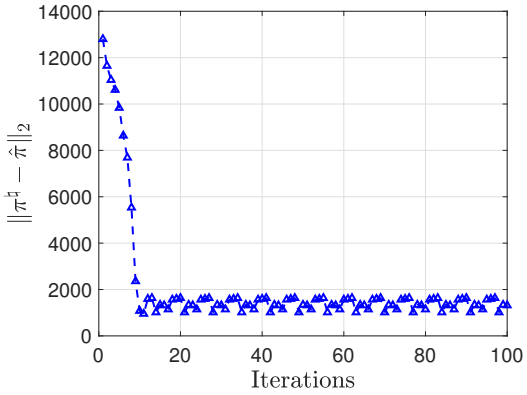


*Figure 1.* Error $\left\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^\natural\right\|_2$ across iterations when $n = 1000, p = 2, m = 1, \boldsymbol{\beta} = [1000\ 1000]^\top$, and $\sigma = 0$.

In this experiment, we use the results $(\widehat{\mathbf{\Pi}},\ \widehat{\mathbf{B}})$ returned by Eq. (4) and Eq. (5) as the initialization point. From Figure 1, we can see the error is approximately 13000. Then we try to refine the results with alternative minimization with the iterative equations being written as

$$\widehat{\mathbf{\Pi}}^{(t+1)} = \text{argmax}_{\mathbf{\Pi}} \left\langle \mathbf{\Pi}, \mathbf{Y}\widehat{\mathbf{B}}^{(t)\top}\mathbf{X}^\top \right\rangle;$$
$$\widehat{\mathbf{B}}^{(t+1)} = \mathbf{X}^\dagger \widehat{\mathbf{\Pi}}^{(t+1)\top}\mathbf{Y},$$

where $\widehat{\mathbf{B}}^{(t)}, \widehat{\mathbf{\Pi}}^{(t)}$ are the values in the $t$-th iteration. With iterative refinement, the error reduces to below 2000 within 20 iterations. This experiment suggests that our estimator returns a $\mathbf{\Pi}$ which is far from the ground truth $\mathbf{\Pi}^\natural$.

The underlying reason is due to the low stable rank $\rho(\mathbf{B}^\natural)$, which is 1 when $m = 1$. In the next section, we will show that the ground truth $\mathbf{\Pi}^\natural$ can be obtained much easier once $\rho(\mathbf{B}^\natural)$ exceeds certain threshold.

### 3.3. Prior research on the $m = 1$ case

In Pananjady et al. (2017a), the ML estimator is investigated, which is only computable for the special case $m = p = 1$ and NP-hard for the rest cases. Their estimator gets the same $\mathsf{SNR}$ requirement as ours, namely $\log(\mathsf{SNR}) \gtrsim \log n$. However, in theory their estimator can obtain the ground truth $\mathbf{\Pi}^\natural$ when $p > 1$. While our estimator will fail with high probability, as shown in Figure 1.

To handle the computational issue of the ML estimator, Hsu et al. (2017) propose an algorithm with polynomial-complexity to obtain an approximated solution. In addition, they focus on the recovery of $\mathbf{B}^\natural$ rather than $\mathbf{\Pi}^\natural$. Their $\mathsf{SNR}$ requirement is $\mathsf{SNR} \geq c\min(1, p/\log\log n)$, which has a gap compared with the bound in Pananjady et al. (2017a).

Slawski and Ben-David (2019) choose to put a sparse constraint on $h$, the number of permuted rows, to tackle the high computational complexity. Then a computable estimator is proposed, which works in both the $p = 1$ and $p > 1$ case. The similar idea has also been applied to the $m > 1$ case as in Slawski et al. (2019a;b).

We would also like to mention the work of Abid et al. (2017), where a consistent estimator is proposed based on the method-of-moments. Their analysis only considers the $m = p = 1$ case and focuses on the deviation $\|\widehat{\mathbf{B}} - \mathbf{B}^\natural\|_{\mathrm{F}}$.

## 4. Multiple Observations Model

Previous section has studied the single observation model, i.e., $m = 1$. This section focuses on our main contributions for the multiple observations model, namely $m \gg 1$. First, we state our main theoretical result as the next Theorem.

**Theorem 2** *Given that* $n \gtrsim p^4(\log n)^6(\log p)^4$, $\rho(\mathbf{B}^\natural) \gtrsim 18/c_0$, $h = \mathsf{d_H}(\mathbf{I}, \mathbf{\Pi}^\natural) \leq \frac{n}{4}$, *if* $\mathsf{SNR}$ *satisfies*

$$\log(\mathsf{SNR}) \gtrsim \frac{\log n}{\rho\left(\mathbf{B}^\natural\right)} + \log\log n, \tag{10}$$

*then we have* $\mathbb{P}(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural) \leq c_0 e^{-\left((\log n)^4 \wedge (\log n)^2 \rho(\mathbf{B}^\natural)\right)} + c_1 n e^{-c_2 m} + c_3 n e^{-c_4 n} + c_5 e^{-p} + c_6 p^{-2}$, *when* $n$ *is sufficiently large, where* $c_i$*'s are some positive constants.*

Once $\mathbf{\Pi}^\natural$ is estimated, this problem in Eq. (3) reduces to the traditional linear regression. We omit the discussion on the error $\|\widehat{\mathbf{B}} - \mathbf{B}^\natural\|_{\mathrm{F}}$, since it is well-studied in this scenario.

**Remark 3** *The requirement* $n \gg p^4$ *in Theorem 2 corresponds to the most stringent case, which can be relaxed with more specific constraints on* $h$ *and* $\rho(\mathbf{B}^\natural)$. *Perhaps the most interesting case is when* $h = O(1)$, *a fixed positive constant. In this case, when* $\rho(\mathbf{B}^\natural) \gtrsim 18/c_0$, *we only need* $n \gtrsim p^2(\log p)^{4/3}(\log n)^2$, *which can be further reduced to* $n \gtrsim p^{3/2}(\log p)(\log n)^{3/2}$ *when* $\rho(\mathbf{B}^\natural) \to \infty$. *On the other hand, even when* $h \asymp n$, *as long as* $\rho(\mathbf{B}^\natural) \to \infty$, *we can relax* $n$ *to be of the order* $n \gtrsim p^2(\log p)^2(\log n)^3$.

**Remark 4** *Numerical experiments suggest correct recovery can still be obtained when* $n \asymp p$, *we believe the requirement* $n \gg p^{3/2}$ *is an artifact of our analysis, which can be improved to* $n \asymp p$ *with more advanced analytical tools. Currently, there is still a gap of* $\Omega(\sqrt{p})$.

The rigorous proof of Theorem 2 is provided in the supplementary material, including supporting lemmas. Here, we would like to explain the main technical challenges in the proof, which lies in the proof that

$$\left\langle \mathbf{\Pi}^\natural, \mathbf{Y}\mathbf{Y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle \geq \left\langle \mathbf{\Pi}, \mathbf{Y}\mathbf{Y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle, \ \forall\, \mathbf{\Pi},$$

holds with high probability given Eq. (10). A direct analysis appears difficult, as it involves the fourth order of Gaussian random variables, especially for the product of random matrices. Meanwhile, bypassing the higher order by considering a relaxed event risks resulting in a loose bound for $\mathsf{SNR}$. How to balance these two issues constitutes the main challenge. In the following context, we give an outline of the proof and refer the interested readers to the supplementary material for the technical details.

We define $\widetilde{\mathbf{B}}$ and $\mathbf{B}^*$ respectively as

$$\widetilde{\mathbf{B}} = (n-h)^{-1}\mathbf{X}^\top \mathbf{\Pi}^\natural \mathbf{X}\mathbf{B}^\natural,$$
$$\mathbf{B}^* = \widetilde{\mathbf{B}} + (n-h)^{-1}\mathbf{X}^\top \mathbf{W}.$$

Firstly we relax the wrong recovery $\{\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural\}$ to event $\mathcal{E}$, i.e. $\{\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural\} \subseteq \mathcal{E}$, which reads as

$$\mathcal{E} \triangleq \left\{ \left\|\mathbf{Y}_{i,:} - \mathbf{X}_{\pi^\natural(i),:}\mathbf{B}^*\right\|_2^2 \geq \|\mathbf{Y}_{i,:} - \mathbf{X}_{j,:}\mathbf{B}^*\|_2^2, \ \exists\, i, j \right\}.$$

The physical meaning of $\mathcal{E}$ is that we may reduce the residual $\|\mathbf{Y} - \mathbf{\Pi}^\natural \mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}$ by changing $\pi^\natural(i)$ to $j$. With this relaxation, we reduce the computation of the fourth order to that of the third order. The same relaxation method has also been adopted in Collier and Dalalyan (2016); Slawski et al. (2019a); Zhang et al. (2019a;b).

Secondly, we upper-bound the probability $\mathbb{P}(\mathcal{E})$ under the SNR assumption in Eq. (10). We should emphasize that Theorem 2 is not proved by defining $\mathbf{B}^* = (n-h)^{-1}\mathbf{X}^\top \mathbf{Y}$ and invoking Theorem 2 as in Slawski et al. (2019a), which requires the $\mathsf{SNR}$ to satisfy

$$\log(\mathsf{SNR}) \gtrsim \log p + \frac{\log n}{\rho(\mathbf{B}^\natural)} + \log\log n,$$

for the correct permutation recovery. With the above bound, we will fail to prove the benefits brought by high $\rho(\mathbf{B}^\natural)$ as shown in Eq. (10), since we still need $\log(\mathsf{SNR}) \gtrsim \log n$ for the ground truth $\mathbf{\Pi}^\natural$ if $\log p \asymp \log n$. Instead, we transform the task to proving the following relations hold with high probability, namely,

$$\left\|\mathbf{X}_{i,:}\left(\widetilde{\mathbf{B}} - \mathbf{B}^\natural\right)\right\|_2 \lesssim \frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}}\|\mathbf{B}^\natural\|_{\mathrm{F}};$$
$$\left\|\mathbf{X}_{i,:}\mathbf{X}^\top \mathbf{W}\right\|_2 \lesssim \sqrt{m}(\log n)\sigma(n+p). \tag{11}$$

In particular, we would like to mention the technique used in bounding $\left\|\mathbf{X}_{i,:}\mathbf{X}^\top \mathbf{W}\right\|_2$. First we review the widely-used bounding procedure, which proceeds as

$$\left\|\mathbf{X}_{i,:}\mathbf{X}^\top \mathbf{W}\right\|_2 \leq \|\mathbf{X}_{i,:}\|_2 \|\mathbf{X}\|_{\mathrm{OP}}\|\mathbf{W}\|_{\mathrm{OP}}$$
$$\overset{①}{\lesssim} \sqrt{p\log n}\left(\sqrt{n}+\sqrt{p}\right)\sigma\left(\sqrt{n}+\sqrt{m}\right)$$
$$\overset{②}{\lesssim} \sqrt{\log n}(n^{3/2})\sigma + \sqrt{mn\log n}\sigma,$$

where in ① we use the fact $\|\mathbf{X}_{i,:}\|_2 \lesssim \sqrt{p \log n}$, $\|\mathbf{X}\|_{\text{OP}} \lesssim \sqrt{n} + \sqrt{p}$, $\|\mathbf{W}\|_{\text{OP}} \lesssim \sigma(\sqrt{n} + \sqrt{m})$ hold with high probability, and in ② we use $n \gg p$. Comparing with our results in Eq. (11), this bound experience inflations when $m \ll n$ and will lift the SNR requirement to $\log(\mathsf{SNR}) \gtrsim \log n$, which hides the role of $\rho(\mathbf{B}^\natural)$ compared with our current result in Theorem 2.

To handle such problem, we adopt the "leave-one-out trick" as in El Karoui (2013); El Karoui et al. (2013); El Karoui (2018); Chen et al. (2019); Sur et al. (2019) and we refer interested readers to the supplementary material for the technical details.

### 4.1. Tightness of the bound

We compare Eq. (10) for the correct permutation recovery with the statistical limit in Theorem 1 in Zhang et al. (2019a), which is also listed as Theorem 6 in this paper. For the convenience of comparison, we consider the special case where the stable rank $\rho(\mathbf{B}^\natural)$ equals to the rank of $\mathbf{B}^\natural$ (i.e., $\mathbf{B}^\natural$'s signal strength is uniformly spread over all eigenvalues). Theorem 6 suggests that wrong permutation matrix $\widehat{\mathbf{\Pi}}$ is expected with high probability if

$$\rho(\mathbf{B}^\natural) \log(\mathsf{SNR}) \lesssim \rho(\mathbf{B}^\natural) \log(1 + \mathsf{SNR}) \lesssim \log n,$$

which holds true regardless of the estimator form. Meanwhile, our estimator recovers $\mathbf{\Pi}^\natural$ correctly provided Eq. (10) holds. When $\rho(\mathbf{B}^\natural) \lesssim \frac{\log n}{\log\log n}$, to put more clear, $\frac{\log n}{\rho(\mathbf{B}^\natural)}$ is the dominant term, our bound is tight and matches the statistical limits with the difference up to some multiplicative constants. Provided that $\rho(\mathbf{B}^\natural) \gg \frac{\log n}{\log\log n}$, we have $\log\log n$ be the dominant term and our estimator experiences a loss of at most $\Omega(\log\log n)$.

### 4.2. Benefits from multiple observations

We investigate the benefits from the high $\rho(\mathbf{B}^\natural)$ by contrasting Theorem 2 with Theorem 1. First, correct permutation matrix $\mathbf{\Pi}^\natural$ can be obtained for all $p$. Hence we can enjoy its computational benefits without worrying the correctness problem. Second, the SNR requirement has been reduced significantly. Under the single observation model, accurate reconstruction of $\mathbf{\Pi}^\natural$ requires SNR to be the order of $\Omega(n^c)$; while this requirement has been decreased to $\Omega(\log n) \vee \Omega(n^{c/\rho(\mathbf{B}^\natural)})$ when more observations are drawn, where $c$ is some positive constant.

Our simulations in Section 5 confirm the benefits from multiple observations as shown in the left panels of Figure 2 and Figure 3. In summary, diversity, i.e., large $\rho(\mathbf{B}^\natural)$, can help both in the computational and statistical perspectives.

## 5. Simulations

This section presents the numerical results. Since our estimator cannot guarantee the correct permutation matrix $\mathbf{\Pi}^\natural$ under the single observation model, our simulations focus on the multiple observations model, i.e., $m > 1$.

### 5.1. Experiment setting

We closely follow the experiment setting in Zhang et al. (2019b). We set the $i$-th column $\mathbf{B}^\natural_{:,i}$ $(1 \le i \le \min(m,p))$ to be the $i$-th canonical basis, which has 1 on the $i$-th entry and 0 elsewhere. One benefit of this setting is that the stable rank $\rho(\mathbf{B}^\natural)$ is easy to compute, i.e., $\min(m,p)$. Then we can use $m$ $(m \le p)$ as a shortcut to denote the stable rank $\rho(\mathbf{B}^\natural)$. We report experiments for $n = 500$ and $n = 1000$, in Figure 2 and Figure 3, respectively.

For each $n$, we choose $p \in \{0.1n, 0.2n\}$ and $h \in \{n/10, n/4\}$. That is, when $n = 500$, we have $p \in \{50, 100\}$ and $h \in \{50, 125\}$; and when $n = 1000$, we have $p \in \{100, 200\}$ and $h \in \{100, 250\}$.

For each chosen set of parameters $(n, p, m, h)$ and SNR value, we simulate the data 1000 times and report the success rate of exact recovery of $\mathbf{\Pi}^\natural$ using our proposed estimator in Algorithm 1. For each $(n, p, m, h)$, we choose the grid of SNR values to ensure that we are able to report the full curve of success rate of recovery from 0% to 100%.

Note that, if $m$ is too small (i.e., the stable rank $\rho(\mathbf{B}^\natural)$ is too small), then the success rate may not reach 100%. In the plots, the smallest $m$ values (e.g., $m = 15$ or $m = 20$) are selected to ensure a 100% success rate can be reached.

In the left panels of Figure 2 and Figure 3, we plot the success rate with respect to SNR. However, in the right panels, we plot the success rate with respect to $\log\det\left(\mathbf{I} + \frac{\mathbf{B}^{\natural\top}\mathbf{B}^\natural}{\sigma^2}\right) / \log n$, for the convenience of comparing with the experiments in Zhang et al. (2019b).

According to the statistical lower bounds as in Theorem 1 in Zhang et al. (2019a) (also listed as Theorem 6), the correct recovery of $\mathbf{\Pi}^\natural$ at least requires

$$\log\det\left(\mathbf{I} + \frac{\mathbf{B}^{\natural\top}\mathbf{B}^\natural}{\sigma^2}\right) \gtrsim \log n. \tag{12}$$

Therefore, Zhang et al. (2019b) report the experimental results by using the ratio $\log\det\left(\mathbf{I} + \frac{\mathbf{B}^{\natural\top}\mathbf{B}^\natural}{\sigma^2}\right) / \log n$ as the x-axis in their plots. Of course, Zhang et al. (2019b) also report the success rate of recovery with respect to SNR.

In summary, readers can directly compare our experimental results with those in Zhang et al. (2019b).

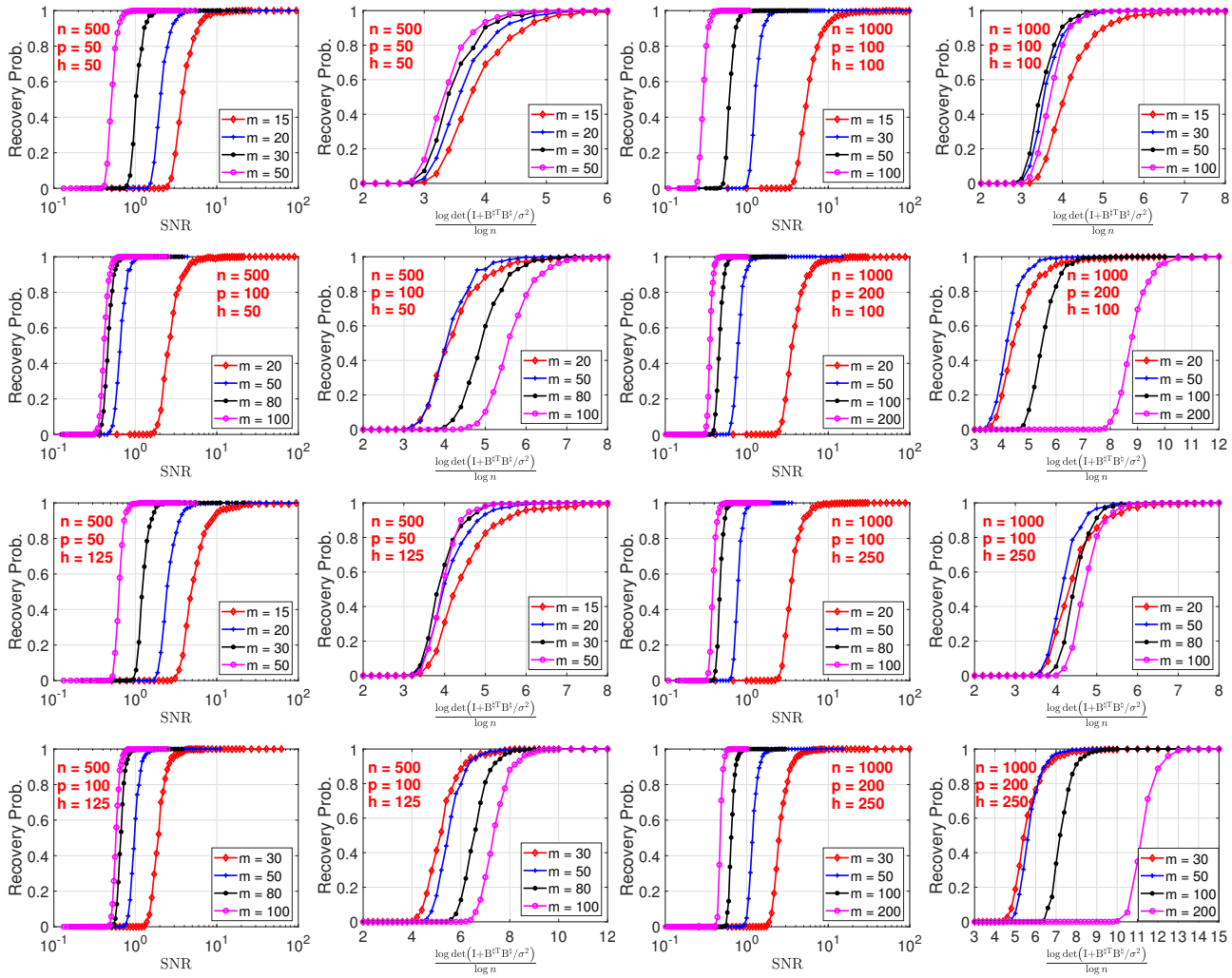


*Figure 2.* The simulated success rate of recovery $\mathbb{P}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^{\natural})$, with $n = 500$, $p \in \{50, 100\}$, $h \in \{50, 125\}$, with respect to SNR (**left panels**) and $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{\natural\top}\mathbf{B}^{\natural}/\sigma^2\right)}{\log n}$ (**right panels**).

*Figure 3.* The simulated success rate of recovery $\mathbb{P}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^{\natural})$, with $n = 1000$, $p \in \{100, 200\}$, $h \in \{100, 250\}$, with respect to SNR (**left panels**) and $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{\natural\top}\mathbf{B}^{\natural}/\sigma^2\right)}{\log n}$ (**right panels**).

### 5.2. Simulation results

The simulation results are reported in Figure 2 and Figure 3 for $n = 500$ and $n = 1000$, respectively. The simulations well match our theoretical results, in particular, Theorem 2.

Recall that, in our setting, the stable $\rho(\mathbf{B}^{\natural})$ is the same as $m$. As shown in the plots, the required SNR values increase with increasing $m$, in order to ensure success of recovery. When $m$ is too small (e.g., $m < 15$), Algorithm 1 cannot reach a $100\%$ success rate in the reasonable range of SNR values we have experimented with.

These plots also demonstrate that the permutation recovery problem becomes increasingly more challenging when the number of permuted rows $h$ becomes larger or when the dimension $p$ gets larger (for fixed $n$).

## 6. Conclusion

This paper has studied the well-known challenging problem of "unlabelled linear regression". Unlike classical linear regression, in this problem setting, a fraction of the rows of the observation matrix are permuted. The goal is to recover not only the regression coefficients but also the permutation matrix. In recent years, this problem has attracted increasingly more attentions because it arises in many important applications including data integration, privacy protection, computer vision, sensor networks, robotics, etc.

In this paper, we propose a (perhaps surprisingly) simple one-step estimator for unlabelled linear regression. Our theoretical analysis reveals that the proposed estimator is optimal from both the computational aspect and the statistical

perspective, by comparing our solution with the oracle estimators. Simulations confirm that the proposed estimator is efficient and accurate, as predicted by theoretical analysis.

One major limitation of the propose estimator is that it requires the matrix of regression coefficients to have a minimum "stable rank" $\rho(\mathbf{B}^\natural)$, which makes the estimator applicable mainly in the multiple observations settings (i.e., $m \gg 1$). In the single observation setting (i.e., $m = 1$), our estimator still works well for the special case when $m = 1$ and $p = 1$. We leave it for future research to modify the estimator so that it will work for more general settings.

## Acknowledgement

We thank the anonymous reviewers and area chair of ICML 2020 for their constructive comments which have helped us improve the quality of the paper.

## Appendix: Lower bounds in prior literature

For the convenience of comparison, we collect some previous results concerning the statistical lower bound for the correct recovery of $\mathbf{\Pi}^\natural$. Note that Theorem 6 gives almost the same order of Theorem 5 when we set $m = 1$. Hence, we can view Theorem 5 as a special case of Theorem 6.

**Theorem 5 (Theorem 2 in Pananjady et al. (2017a))** *For any estimator $\widehat{\mathbf{\Pi}}$, we have the error probability $\mathbb{P}(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural)$ exceed $1 - c_0 e^{-c_1 n\delta}$ provided that $2 + \log(1 + \mathsf{SNR}) \leq (2 - \delta)\log n$, $0 < \delta < 2$.*

**Theorem 6 (Theorem 1 in Zhang et al. (2019a))** *For any estimator $\widehat{\mathbf{\Pi}}$, we have the error probability $\mathbb{P}(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural)$ exceed $1/2$, provided that*

$$\frac{1}{2}\log\det\left(\mathbf{I} + \frac{\mathbf{B}^{\natural\top}\mathbf{B}^\natural}{\sigma^2}\right) + \frac{\log(|\mathcal{H}|)}{2n} < \frac{\mathsf{H}(\mathbf{\Pi}^\natural) - 1}{n}, \tag{13}$$

*where $\mathcal{H}$ and $\mathsf{H}(\mathbf{\Pi}^\natural)$ denote the support and the entropy of the random permutation matrix $\mathbf{\Pi}^\natural$, respectively.*

## 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. *Probab. Theory Relat. Fields*, 131(4):528–552, 2005.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.*, 45 (1):77–120, 02 2017. doi: 10.1214/16-AOS1435.

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

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Trans. Inf. Theory*, 61(4): 1985–2007, 2015.

Hock-Peng Chan and Wei-Liem Loh. A file linkage problem of degroot and goel revisited. *Statistica Sinica*, 11 (4):1031–1045, 2001.

Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *arXiv preprint arXiv:1902.07698*, 2019.

Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans. Signal Process.*, 67(20):5239–5269, 2019.

Olivier Collier and Arnak S. Dalalyan. Minimax rates in permutation estimation for feature matching. *J. Mach. Learn. Res.*, 17:6:1–6:31, 2016.

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. *Ann. Statist.*, 8(2):264–278, 03 1980.

Morris H. DeGroot, Paul I. Feder, and Prem K. Goel. Matchmaking. *Ann. Math. Statist.*, 42(2):578–593, 04 1971.

Ivan Dokmanic. Permutations unlabeled beyond sampling unknown. *IEEE Signal Process. Lett.*, 26(6):823–827, 2019.

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. *Probab. Theory Relat. Fields*, 170(1-2):95–175, 2018.

Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

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

Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins University Press Baltimore, 4th edition, 2013.

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.

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

Rafal Latala, Piotr Mankiewicz, Krzysztof Oleszkiewicz, and Nicole Tomczak-Jaegermann. Banach-mazur distances and projections on random subgaussian polytopes. *Discret. Comput. Geom.*, 38(1):29–50, 2007.

Ivan Nazarov, Boris Shirokikh, Maria Burkina, Gennady Fedonin, and Maxim Panov. Sparse group inductive matrix completion. *arXiv preprint arXiv:1804.10653*, 2018.

Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 64(5):3286–3300, 2017a.

Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Denoising linear models with permuted data. In *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT)*, pages 446–450, Aachen, Germany, 2017b.

Martin Slawski and Emanuel Ben-David. Linear regression with sparsely permuted data. *Electron. J. Statist.*, 13(1): 1–36, 2019.

Martin Slawski, Emanuel Ben-David, and Ping Li. A two-stage approach to multivariate linear regression with sparsely mismatched data. *arXiv preprint arXiv:1907.07148*, 2019a.

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)*, page 7, Tel Aviv, Israel, 2019b.

Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probab. Theory Relat. Fields*, 175(1-2):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. Eigenspace conditions for homomorphic sensing. arXiv:1812.07966, December 2018.

Manolis C. Tsakiris and Liangzu Peng. Homomorphic sensing. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6335–6344, Long Beach, CA, 2019.

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*. Cambridge Series in Statistical and Probabilistic Mathematics. 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.

William E Winkler. Matching and record linkage. *Business survey methods*, 1:355–384, 1995.

Hang Zhang and Ping Li. Sparse recovery with permuted labels: Statistical limits and practical estimators. Technical report, 2020.

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

Hang Zhang, Martin Slawski, and Ping Li. The benefits of diversity: Permutation recovery in unlabeled sensing from multiple measurement vectors. *arXiv preprint arXiv:1909.02496*, 2019b.

## A. Notations

We begin the appendix with a restatement of the notations. Denote $c$, $c^{'}$, $c_i$ as some universal positive constants. Notice that their values may not necessarily the same even for those with same notations. We denote $a \lesssim b$ if there exists some positive constant $c_0 > 0$ 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 $\mathbf{X}$, we denote $\mathbf{X}_{i,:}$ as the $i$-th row, $\mathbf{X}_{:,i}$ as its $i$-th column, and $X_{ij}$ as the $(i,j)$-th element. The Frobenius norm of $\mathbf{X}$ is defined as $\|\mathbf{X}\|_{\mathrm{F}}$ while the operator norm is denoted as $\|\mathbf{X}\|_{\mathrm{OP}}$, whose definition can be found in Section 2.3 of Golub and Loan (2013) (P71). Its stable rank $\rho(\mathbf{X})$ is defined as the ratio $\|\mathbf{X}\|_{\mathrm{F}}^2/\|\mathbf{X}\|_{\mathrm{OP}}^2$ (Section 2.1.15 in Tropp (2015)). The inner product $\langle \mathbf{A}, \mathbf{C} \rangle$ is defined as $\sum_{ij} A_{ij} C_{ij}$.

Associate with each permutation matrix $\mathbf{\Pi}$, we define the operator $\pi(\cdot)$ that transforms index $i$ to $\pi(i)$. The Hamming distance $\mathsf{d}_{\mathsf{H}}(\mathbf{\Pi}_1, \mathbf{\Pi}_2)$ between permutation matrix $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ is defined as $\mathsf{d}_{\mathsf{H}}(\mathbf{\Pi}_1, \mathbf{\Pi}_2) = \sum_{i=1}^{n} \mathbb{1}(\pi_1(i) \neq \pi_2(i))$. Additionally, we denote $\overline{\mathcal{E}}$ as the complement of the event $\mathcal{E}$ and the *signal-to-noise-ratio* (SNR) as $\mathsf{SNR} = \left\|\mathbf{B}^{\natural}\right\|_{\mathrm{F}}^2/(m\sigma^2)$.

## B. Problem Restatement

To begin with, we recall the problem formulation, which reads as

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

where $\mathbf{Y} \in \mathbb{R}^{n\times m}$ represents the observation, $\mathbf{\Pi} \in \mathbb{R}^{n\times n}$ denotes the unknown permutation matrix, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the sensing matrix (design matrix) with $X_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$ being a standard normal random variable (RV), $\mathbf{B}^{\natural} \in \mathbb{R}^{p\times m}$ is the matrix of regression coefficients, and $\mathbf{W} \in \mathbb{R}^{n\times m}$ is the additive Gaussian noise matrix such that $W_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.

Our goal is to reconstruct the pair $(\widehat{\mathbf{\Pi}}, \widehat{\mathbf{B}})$ from the observation $\mathbf{Y}$ and sensing matrix (design matrix) $\mathbf{X}$. The proposed one-step estimator can be written as

$$\begin{aligned}
\widehat{\mathbf{\Pi}} &= \mathrm{argmax}_{\mathbf{\Pi}\in\mathcal{P}_n} \left\langle \mathbf{\Pi}, \mathbf{Y}\mathbf{Y}^{\top}\mathbf{X}\mathbf{X}^{\top} \right\rangle, \\
\widehat{\mathbf{B}} &= (\mathbf{X})^{\dagger}\, \widehat{\mathbf{\Pi}}^{\top}\mathbf{Y},
\end{aligned}$$

where $\mathbf{X}^{\dagger} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}$ denotes the pseudo-inverse of $\mathbf{X}$. In the following, we will separately investigate its properties under the single observation model $(m = 1)$ and multiple observations model $(m > 1)$. The formal statement is packaged in Theorem 1 and Theorem 2.

## C. Appendix for Section 3

This section focuses on the special case where $p = 1, m = 1$. Consider $\mathbf{X} \in \mathbb{R}^n$ to be a Gaussian distributed RV such that $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$, and permutation matrix $\mathbf{\Pi}^{\natural}$ which satisfies $\mathsf{d}_{\mathsf{H}}(\mathbf{I}, \mathbf{\Pi}^{\natural}) = h \leq n/4$.

### C.1. Notations

First we define the following events $\mathcal{E}_i$, $(1 \leq i \leq 5)$, which reads

$$\begin{aligned}
\mathcal{E}_1 &\triangleq \left\{ \left\langle \mathbf{X}, \mathbf{\Pi}^{\natural}\mathbf{X} \right\rangle \geq c_0 n \right\}, \\
\mathcal{E}_2 &\triangleq \left\{ \|\mathbf{X}\|_2 \leq 2\sqrt{n} \right\} \\
\mathcal{E}_3(\mathbf{\Pi}) &\triangleq \left\{ \mathbf{W}^{\top}\mathbf{X}\mathbf{X}^{\top}\left(\mathbf{\Pi}^{\natural} - \mathbf{\Pi}\right)\mathbf{W} \lesssim \sigma^2 n^2 \log n \right\}, \\
\mathcal{E}_4(\mathbf{\Pi}) &\triangleq \left\{ \left| \langle \mathbf{W}, \mathbf{X} \rangle \left\langle \mathbf{\Pi}^{\natural}\mathbf{X}, \left(\mathbf{\Pi}^{\natural} - \mathbf{\Pi}\right)^{\top}\mathbf{X} \right\rangle + \left\langle \mathbf{W}, \left(\mathbf{\Pi}^{\natural} - \mathbf{\Pi}\right)^{\top}\mathbf{X} \right\rangle \left\langle \mathbf{\Pi}^{\natural}\mathbf{X}, \mathbf{X} \right\rangle \right| \lesssim \sigma n^2 \sqrt{\log n} \right\} \\
\mathcal{E}_5(\mathbf{\Pi}; \ell) &\triangleq \left\{ \|\mathbf{X} - \mathbf{\Pi}\mathbf{X}\|_2^2 \geq \frac{12\ell}{5en^{20}},\ \mathsf{d}_{\mathsf{H}}(\mathbf{I}, \mathbf{\Pi}) = \ell \right\},
\end{aligned}$$

where $\mathbf{\Pi}$ is an arbitrary permutation matrix, and $c_0 > 0$ is some positive constant.

### C.2. Outline of proof

We will prove that ground truth permutation matrix $\mathbf{\Pi}^\natural$ will be returned with high probability under the assumptions in Theorem 1. The formal statement is shown in Theorem 1. Before we delve into the proof details, we give a roadmap of the proof, which is

- **Step I:** Under the events $\mathcal{E}_1 \bigcap_{\mathbf{\Pi}} (\mathcal{E}_3(\mathbf{\Pi}) \bigcap \mathcal{E}_4(\mathbf{\Pi}) \bigcap \mathcal{E}_5(\mathbf{\Pi}; \ell))$, we have

$$\left\langle \mathbf{\Pi}^\natural, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle - \left\langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle \gtrsim \frac{c_0 \beta^2}{n^{19}} - c_1 \beta \sigma n^2 \sqrt{\log n} - c_2 \sigma^2 n^2 \log n.$$

  Notice that under assumptions in Theorem 1, we conclude that $\left\langle \mathbf{\Pi}^\natural, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle > \left\langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle$, $\forall\ \mathbf{\Pi}$, which suggests that $\mathbf{\Pi}^\natural$ will always be returned by our estimator in Eq. (3).

- **Step II:** We upper-bound the probability $\mathbb{P}(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural)$ by $\mathbb{P}\left(\overline{\mathcal{E}}_1 \bigcup_{\mathbf{\Pi}} \left(\overline{\mathcal{E}}_3(\mathbf{\Pi}) \bigcup \overline{\mathcal{E}}_4(\mathbf{\Pi}) \bigcup \overline{\mathcal{E}}_5(\mathbf{\Pi}; \ell)\right)\right)$ and complete the proof by showing it is at most $cn^{-1}$.

Having illustrated the proof strategy, we turn to the proof details. The main proof is attached in Section C.3 while the supporting lemmas bounding $\mathbb{P}(\mathcal{E}_i)$, $(1 \leq i \leq 5)$, are put in Section C.4.

### C.3. Proof of Theorem 1

**Proof 1** *For an arbitrary permutation matrix $\mathbf{\Pi}$, we can expand the term $\left\langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle$ as*

$$\left\langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle = \mathcal{T}_1(\mathbf{\Pi}) + \beta \mathcal{T}_2(\mathbf{\Pi}) + \beta^2 \mathcal{T}_3(\mathbf{\Pi}),$$

*where $\mathcal{T}_i(\mathbf{\Pi})$, $(1 \leq i \leq 3)$, are defined as*

$$\begin{aligned}
\mathcal{T}_1(\mathbf{\Pi}) &= \left\langle \mathbf{W}, \mathbf{\Pi}^\top \mathbf{X} \right\rangle \langle \mathbf{X}, \mathbf{W} \rangle; \\
\mathcal{T}_2(\mathbf{\Pi}) &= \langle \mathbf{W}, \mathbf{X} \rangle \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{\Pi}^\top \mathbf{X} \right\rangle + \left\langle \mathbf{W}, \mathbf{\Pi}^\top \mathbf{X} \right\rangle \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{X} \right\rangle; \\
\mathcal{T}_3(\mathbf{\Pi}) &= \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{\Pi} \mathbf{X} \right\rangle \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{X} \right\rangle.
\end{aligned}$$

***Step I:*** *We rewrite the difference $\left\langle \mathbf{\Pi}^\natural, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle - \left\langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle$ as*

$$\begin{aligned}
&\left\langle \mathbf{\Pi}^\natural, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle - \left\langle \mathbf{\Pi}, \mathbf{y}\mathbf{y}^\top \mathbf{X}\mathbf{X}^\top \right\rangle \\
=\ & \mathcal{T}_1(\mathbf{\Pi}^\natural) - \mathcal{T}_1(\mathbf{\Pi}) + \beta \left(\mathcal{T}_2\left(\mathbf{\Pi}^\natural\right) - \mathcal{T}_2(\mathbf{\Pi})\right) + \beta^2 \left(\mathcal{T}_3\left(\mathbf{\Pi}^\natural\right) - \mathcal{T}_3(\mathbf{\Pi})\right) \\
\overset{①}{=}\ & \frac{\beta^2}{2} \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{X} \right\rangle \left\| \mathbf{X} - \mathbf{\Pi}^{\natural\top} \mathbf{\Pi} \mathbf{X} \right\|_2^2 + \beta \left(\mathcal{T}_2\left(\mathbf{\Pi}^\natural\right) - \mathcal{T}_2(\mathbf{\Pi})\right) + \mathcal{T}_1(\mathbf{\Pi}^\natural) - \mathcal{T}_1(\mathbf{\Pi}) \\
\overset{②}{\geq}\ & \frac{\beta^2}{2} c_0 n \frac{24}{5en^{20}} - \beta \left|\mathcal{T}_2\left(\mathbf{\Pi}^\natural\right) - \mathcal{T}_2(\mathbf{\Pi})\right| - \left|\mathcal{T}_1(\mathbf{\Pi}^\natural) - \mathcal{T}_1(\mathbf{\Pi})\right| \\
\overset{③}{\gtrsim}\ & \frac{c_0 \beta^2}{n^{19}} - c_1 \beta \sigma n^2 \sqrt{\log n} - c_2 \sigma^2 n^2 \log n \overset{④}{>} 0,
\end{aligned}$$

*where in ① we rewrite $\|\mathbf{X}\|_2^2 - \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{\Pi} \mathbf{X} \right\rangle$ as*

$$\|\mathbf{X}\|_2^2 - \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{\Pi} \mathbf{X} \right\rangle = \frac{1}{2} \left( \|\mathbf{X}\|_2^2 + \left\| \mathbf{\Pi}^{\natural\top} \mathbf{\Pi} \mathbf{X} \right\|_2^2 - 2 \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{\Pi} \mathbf{X} \right\rangle \right) = \frac{1}{2} \left\| \mathbf{X} - \mathbf{\Pi}^{\natural\top} \mathbf{\Pi} \mathbf{X} \right\|_2^2,$$

*in ② we condition on event $\mathcal{E}_1, \mathcal{E}_5(\mathbf{\Pi}; \ell)$ and have $\|\mathbf{X} - \mathbf{\Pi}\mathbf{X}\|_2^2 \geq \frac{12\ell}{5en^{20}} \geq \frac{24}{5en^{20}}$, in ③ we condition on $\mathcal{E}_3(\mathbf{\Pi}), \mathcal{E}_4(\mathbf{\Pi})$, and in ④ we use the assumption $\log(\mathsf{SNR}) \gtrsim \log n$ in Theorem 1.*

***Step II:*** *The error probability* $\mathbb{P}\left(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural\right)$ *is hence be upper-bounded as*

$$\begin{aligned}
\mathbb{P}\left(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural\right) \leq\ & \mathbb{P}\left(\overline{\mathcal{E}}_1 \bigcup_{\mathbf{\Pi}} \left(\overline{\mathcal{E}}_3(\mathbf{\Pi}) \bigcup \overline{\mathcal{E}}_4(\mathbf{\Pi}) \bigcup \overline{\mathcal{E}}_5(\mathbf{\Pi}; \ell)\right)\right) \\
\overset{⑤}{\leq}\ & \mathbb{P}\left(\bigcup_{\mathbf{\Pi}} \left(\overline{\mathcal{E}}_3(\mathbf{\Pi}) \bigcup \overline{\mathcal{E}}_4(\mathbf{\Pi}) \bigcup \overline{\mathcal{E}}_5(\mathbf{\Pi})\right) \bigcap \mathcal{E}_1 \bigcap \mathcal{E}_2\right) + \mathbb{P}\left(\overline{\mathcal{E}}_1\right) + \mathbb{P}\left(\overline{\mathcal{E}}_2\right) \\
\overset{⑥}{\leq}\ & \sum_{\mathbf{\Pi}^\natural \neq \mathbf{\Pi}} \mathbb{P}\left(\overline{\mathcal{E}}_3(\mathbf{\Pi}) \bigcap \mathcal{E}_1 \bigcap \mathcal{E}_2\right) + \sum_{\mathbf{\Pi}^\natural \neq \mathbf{\Pi}} \mathbb{P}\left(\overline{\mathcal{E}}_4(\mathbf{\Pi}) \bigcap \mathcal{E}_1 \bigcap \mathcal{E}_2\right) \\
& + \sum_{\ell \geq 2} \mathbb{P}\left(\overline{\mathcal{E}}_5(\mathbf{\Pi}; \ell) \bigcap \mathcal{E}_1 \bigcap \mathcal{E}_2\right) + 8n^{-1} + 2e^{-c_0 n} \\
\overset{⑦}{\leq}\ & 2n^{-n} + 3\sum_{\ell \geq 2} \binom{n}{\ell} \ell! n^{-2\ell} + 8n^{-1} + 2e^{-c_0 n} \\
\overset{⑧}{\lesssim}\ & c_0 n^{-n} + n^{-1} + 3\sum_{\ell \geq 2} n^\ell n^{-2\ell} \lesssim c_0 n^{-1} + \frac{3}{n(n-1)} \lesssim n^{-1},
\end{aligned}$$

*where in ⑤ we use the union bound, in ⑥ we complete the proof with Lemma 1 and the fact* $\mathbb{P}\left(\overline{\mathcal{E}}_2\right) \leq e^{-0.8n}$*, in ⑦ we invoke Lemma 2, Lemma 3, Lemma 4, and in ⑧ we use* $n!/(n-\ell)! \leq n^\ell$ *and complete the proof.*

### C.4. Supporting Lemmas for Theorem 1

This subsection collects the supporting lemmas for the proof of Theorem 1.

**Lemma 1** *We have* $\mathbb{P}\left(\overline{\mathcal{E}}_1\right) \leq 8n^{-1} + e^{-0.238n}$ *when* $n$ *is sufficiently large.*

**Proof 2** *Different from the proof in Lemma 9, we consider the case where* $\mathbf{X} \in \mathbb{R}^n$ *is a vector and would lower-bound* $\left\langle \mathbf{X}, \mathbf{\Pi}^\natural \mathbf{X} \right\rangle$*. W.l.o.g, we assume the first* $h$ *entries are permuted and expand the inner product* $\left\langle \mathbf{X}, \mathbf{\Pi}^\natural \mathbf{X} \right\rangle$ *as*

$$\left\langle \mathbf{X}, \mathbf{\Pi}^\natural \mathbf{X} \right\rangle = \sum_{i=1}^{h} X_i X_{\pi(i)} + \sum_{i=h+1}^{n} X_i^2.$$

*With union bound, we can upper bound* $\mathbb{P}\left(\left\langle \mathbf{X}, \mathbf{\Pi}^\natural \mathbf{X} \right\rangle \leq c_0 n\right)$ *as*

$$\mathbb{P}\left(\left\langle \mathbf{X}, \mathbf{\Pi}^\natural \mathbf{X} \right\rangle \leq c_0 n\right) \overset{①}{\leq} \underbrace{\mathbb{P}\left(\sum_{i=h+1}^{n} X_i^2 \leq \frac{1}{4}(n-h)\right)}_{\zeta_1} + \underbrace{\mathbb{P}\left(\sum_{i=1}^{h} X_i X_{\pi(i)} \leq -\frac{4\sqrt{2}+\sqrt{35}}{\sqrt{2}}\sqrt{n \log n}\right)}_{\zeta_2},$$

*where* $c_0 > 0$ *is some positive constant, in ① we use the fact*

$$\frac{n-h}{4} - \frac{4\sqrt{2}+\sqrt{35}}{\sqrt{2}}\sqrt{n \log n} \overset{(h \leq \frac{n}{4})}{\geq} \frac{3n}{16} - \frac{4\sqrt{2}+\sqrt{35}}{\sqrt{2}}\sqrt{n \log n} \geq c_0 n,$$

*when* $n$ *is large. We finish the proof by separately upper-bounding* $\zeta_1 \leq e^{-0.2386n}$ *and* $\zeta_2 \leq 8n^{-1}$*. The detailed computation comes as follows.*

***Phase I:*** *For* $\zeta_1$*, we can view* $\sum_{i=h+1}^{n} X_i^2$ *as a* $\chi^2$*-RV with* $(n-h)$ *freedom and have*

$$\zeta_1 \overset{②}{\leq} \exp\left(\frac{n-h}{2}\left(\log\frac{1}{4} - \frac{1}{4} + 1\right)\right) \overset{③}{\leq} e^{-0.2386n},$$

*where in ② we use Lemma 11, and ③ is because $h \le n/4$.*

***Phase II:*** *To bound $\zeta_2$, we divide the index set $\{j : j \neq \pi(j)\}$ into 3 disjoint sets $\mathcal{I}_i$, $1 \le i \le 3$, as in Lemma 8 in Pananjady et al. (2017a) (restated as Lemma 13). This division has two properties: (i) indices $j$ and $\pi(j)$ lies in different sets; (ii) the cardinality $h_i$ of each $\mathcal{I}_i$ satisfies $\lfloor h/5 \rfloor \le h_i \le h/3$. Then we obtain*

$$\begin{aligned}\zeta_2 &\le \mathbb{P}\left(\sum_{i=1}^{h} X_i X_{\pi(i)} \le -\frac{4\sqrt{2}+\sqrt{35}}{\sqrt{2}}\sqrt{n\log n},\ |X_i| \le 2\sqrt{\log n},\ \forall\, i\right) + \mathbb{P}\left(|X_i| \ge 2\sqrt{\log n},\ \exists\, i\right)\\ &\overset{④}{\le} \sum_{i=1}^{3} \underbrace{\mathbb{P}\left(\sum_{j\in\mathcal{I}_i} X_j X_{\pi(j)} \le -\frac{4\sqrt{2}+\sqrt{35}}{3\sqrt{2}}\sqrt{n\log n},\ |X_i| \le 2\sqrt{\log n},\ \forall\, i\right)}_{\zeta_{2,i}} + n\underbrace{\mathbb{P}\left(|X_i| \ge 2\sqrt{\log n}\right)}_{\le 2n^{-2}},\end{aligned}$$

*where in ④ we use the union bound for $\sum_{i=1}^{h} X_i X_{\pi(i)}$ and the tail bounds for Gaussian distributed $X_i$.*

*Then we define $Z_i = \sum_{j\in\mathcal{I}_i} X_j X_{\pi(j)}$ and bound $\zeta_{2,i}$ via the Bernstein inequality (Theorem 2.8.4 in Vershynin (2018)). First, we verify that $\mathbb{E}\left(X_j X_{\pi(j)}\right) = \left(\mathbb{E}X_j\right)\left(\mathbb{E}X_{\pi(j)}\right) = 0$. Meanwhile we compute $\sigma^2 = \sum_{j\in\mathcal{I}_i}\mathbb{E}\left(X_j X_{\pi(j)}\right)^2 = h_i$. According to the Bernstein inequality, we have*

$$\left|\sum_{j\in\mathcal{I}_i} X_j X_{\pi(j)}\right| \ge \frac{4}{3}(\log n)^2 + \sqrt{\frac{16}{9}(\log n)^4 + 2(\log n)h_i},$$

*holds with probability $2n^{-1}$. Meanwhile, we can upper bound as*

$$\frac{4}{3}(\log n)^2 + \sqrt{\frac{16}{9}(\log n)^4 + 2(\log n)h_i} \le \frac{4}{3}(\log n)^2 + \sqrt{\frac{16}{9}(\log n)^4 + \frac{n\log n}{6}} \overset{⑤}{\le} \frac{4\sqrt{2}+\sqrt{35}}{3\sqrt{2}}\sqrt{n\log n},$$

*where ⑤ is because $n \ge \log^3(n)$ for $n \ge 95$. Hence, we conclude that $\zeta_{2,i} \le 2n^{-1}$ and complete the proof by combining the bound for $\zeta_1$ and $\zeta_2$.*

**Lemma 2** *We have $\mathbb{P}\left(\overline{\mathcal{E}}_3(\mathbf{\Pi})\bigcap\mathcal{E}_2\right) \le n^{-2n}$.*

**Proof 3** *For the conciseness of notation, we define $\mathbf{\Xi}$ as $\mathbf{\Xi} \triangleq \mathbf{X}\mathbf{X}^\top\left(\mathbf{\Pi}^\natural - \mathbf{\Pi}\right)$. Due to the independence of the $\mathbf{X}$ and $\mathbf{W}$, we can condition on $\mathbf{X}$ and bound $\mathbb{P}(\overline{\mathcal{E}}_3(\mathbf{\Pi})\bigcap\mathcal{E}_2)$ as*

$$\begin{aligned}\mathbb{P}\left(\overline{\mathcal{E}}_3(\mathbf{\Pi})\bigcap\mathcal{E}_2\right) &\overset{①}{\le} \mathbb{P}\left(\mathbf{W}^\top\mathbf{\Xi}\mathbf{W} \ge \mathbb{E}\mathbf{W}^\top\mathbf{\Xi}\mathbf{W} + c\sigma^2 n^2\log n\right)\\ &\overset{②}{\le} \exp\left(-\left(\frac{c_0 n^4\log^2 n}{\|\mathbf{\Xi}\|_{\mathrm{F}}^2} \wedge \frac{c_1 n^2\log n}{\|\mathbf{\Xi}\|_2}\right)\right) \overset{③}{\le} n^{-2n},\end{aligned}$$

*where in ① we condition on $\mathcal{E}_2$ and use the fact*

$$\mathbb{E}\mathbf{W}^\top\mathbf{\Xi}\mathbf{W} + c\sigma^2 n^2\log n \lesssim \sigma^2\|\mathbf{X}\|_2^2 + c\sigma^2 n^2\log n \lesssim \sigma^2 n^2\log n,$$

*in ② we use Hanson-Wright inequality (Theorem 6.2.1 in Vershynin (2018)), and in ③ we condition on $\mathcal{E}_2$ and use $\|\mathbf{\Xi}\|_2 \lesssim \|\mathbf{X}\|_2^2 \lesssim n$.*

**Lemma 3** *We have $\mathbb{P}\left(\overline{\mathcal{E}}_4(\mathbf{\Pi})\bigcap\mathcal{E}_2\right) \le n^{-2n}$.*

**Proof 4** *Due to the independence between $\mathbf{W}$ and $\mathbf{X}$, we would like to condition on $\mathbf{X}$ and bound $\mathbb{P}(\overline{\mathcal{E}}_4(\mathbf{\Pi})\bigcap\mathcal{E}_2)$ as*

$$\mathbb{P}\left(\overline{\mathcal{E}}_4(\mathbf{\Pi})\bigcap\mathcal{E}_2\right) \le \exp\left(-\frac{4c\sigma^2 n^4\log n}{2\sigma_{\mathbf{\Pi}}^2}\right),$$

*where* $\sigma_{\mathbf{\Pi}}^2$ *is defined as*

$$\sigma_{\mathbf{\Pi}}^2 = \sigma^2 \left\| \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \left(\mathbf{\Pi}^\natural - \mathbf{\Pi}\right)^\top \mathbf{X} \right\rangle \mathbf{X} + \left\langle \mathbf{\Pi}^\natural \mathbf{X}, \mathbf{X} \right\rangle \left(\mathbf{\Pi}^\natural - \mathbf{\Pi}\right) \mathbf{X} \right\|_{\mathrm{F}}^2,$$

*Notice under* $\mathcal{E}_2$*, we have* $\sigma_{\mathbf{\Pi}}^2 \lesssim \sigma^2 \left(4\|\mathbf{X}\|_2^3\right)^2 = c\sigma^2 n^3$*, and complete the proof by showing*

$$\exp\left(-\frac{4c\sigma^2 n^4 \log n}{2\sigma_{\mathbf{\Pi}}^2}\right) \le \exp\left(-\frac{4c\sigma^2 n^4 \log n}{2c\sigma^2 n^3}\right) = n^{-2n}.$$

**Lemma 4** *We have* $\mathbb{P}\left(\overline{\mathcal{E}}_5(\mathbf{\Pi}); \ell\right) \le 3n^{-2\ell}$.

**Proof 5** *Adopting a similar approach as in proving Lemma 1, we can decompose the index sets* $\{j : j \neq \pi(j)\}$ *into* 3 *disjoint sets* $\mathcal{I}_i$ $(1 \le i \le 3)$ *such that:* (1) $j$ *and* $\pi(j)$ *do not lie within the same index set* $\mathcal{I}_i$*; and* (2) *the cardinality* $\ell_i$ *of* $\mathcal{I}_i$ *satisfies* $\lfloor \ell/5 \rfloor \le \ell_i \le \ell/3$*. Then we can bound* $\mathbb{P}\left(\mathcal{E}_5(\mathbf{\Pi}; \ell)\right)$ *as*

$$\mathbb{P}\left(\left\|\mathbf{X} - \mathbf{\Pi}^\natural \mathbf{X}\right\|_2^2 \le \frac{12\ell}{5en^{20}}\right) \overset{①}{=} \sum_{i=1}^{3} \mathbb{P}\left(\sum_{j \in \mathcal{I}_i} \left(X_j - X_{\pi(j)}\right)^2 \le \frac{4\ell}{5en^{20}}\right)$$
$$\overset{②}{\le} \sum_{i=1}^{3} \exp\left(\frac{\ell_i}{2}\left(\log \frac{2l}{5en^{20}\ell_i} - \frac{2l}{5en^{20}\ell_i} + 1\right)\right) \overset{③}{\le} 3n^{-2\ell}.$$

*where* ① *is due to the decomposition* $\mathcal{I}_i$*,* $1 \le i \le 3$*,* ② *is because* $\sum \left(X_j - X_{\pi(j)}\right)^2 / 2$ *is a* $\chi^2$ *RV with freedom* $\ell_i$ *and Lemma 11, and* ③ *is due to* $\lfloor \ell/5 \rfloor \le \ell_i \le \ell/3$ *and hence*

$$\frac{\ell_i}{2}\left(\log \frac{2l}{5en^{20}\ell_i} - \frac{2l}{5en^{20}\ell_i} + 1\right) \le \frac{\ell_i}{2}\left(\log \frac{2l}{5\ell_i} - 20\log n\right) \le -10\ell_i \log n \le -2\ell \log n.$$

## D. Appendix for Section 4

This section provides theoretical analysis for the multiple observations model, i.e., $m > 1$. We will show that our estimator in Eq. (3) gives correct permutation matrix $\mathbf{\Pi}^\natural$ once

$$\log(\mathsf{SNR}) \gtrsim \frac{\log n}{\rho(\mathbf{B}^\natural)} + \log\log n.$$

The formal statement is packaged in Theorem 2.

### D.1. Notations

Before our discussion, first we define $\widetilde{\mathbf{B}}$ and $\mathbf{B}^*$ respectively as

$$\widetilde{\mathbf{B}} = (n-h)^{-1} \mathbf{X}^\top \mathbf{\Pi}^\natural \mathbf{X} \mathbf{B}^\natural,$$
$$\mathbf{B}^* = (n-h)^{-1} \mathbf{X}^\top \mathbf{Y} = \widetilde{\mathbf{B}} + (n-h)^{-1} \mathbf{X}^\top \mathbf{W},$$

where $h$ is denoted as the Hamming distance between identity matrix $\mathbf{I}$ and the ground truth permutation matrix $\mathbf{\Pi}^\natural$, i.e., $h = \mathsf{d}_{\mathsf{H}}(\mathbf{I}, \mathbf{\Pi}^\natural)$. Similar as in Section C, we define events $\mathcal{E}_i$, $(6 \le i \le 9)$ as

$$\mathcal{E}_6 \triangleq \left\{\|\mathbf{X}_{i,:}\|_2 \le 2\sqrt{p \log n},\ \forall i\right\};$$
$$\mathcal{E}_7 \triangleq \left\{\left\|\mathbf{X}_{i,:}\left(\mathbf{B}^* - \mathbf{B}^\natural\right)\right\|_2 \lesssim c_0 \frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}} \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} + c_1\sqrt{m}(\log n)\sigma\left(1 + \frac{p}{n}\right),\ \forall\, i\right\};$$
$$\mathcal{E}_8 \triangleq \left\{\left\langle \mathbf{W}_{i,:}, \left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^\natural(i),:}\right)\mathbf{B}^*\right\rangle \ge \Delta,\ \exists\, i, j\right\};$$
$$\mathcal{E}_9 \triangleq \left\{\left\|\left(\mathbf{X}_{\pi^\natural(i),:} - \mathbf{X}_{j,:}\right)\mathbf{B}^\natural\right\|_2^2 + 2\left\langle\left(\mathbf{X}_{\pi^\natural(i),:} - \mathbf{X}_{j,:}\right)\mathbf{B}^\natural, \mathbf{X}_{j,:}\left(\mathbf{B}^\natural - \mathbf{B}^*\right)\right\rangle - \left\|\mathbf{X}_{\pi^\natural(i),:}\left(\mathbf{B}^\natural - \mathbf{B}^*\right)\right\|_2^2 \le \Delta,\ \exists\, i, j\right\},$$

where $\Delta$ is defined as

$$\Delta = 16\sqrt{2}c_0\sigma\frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} + 16c_1\sqrt{2m}(\log n)\sigma^2\left(1 + \frac{p}{n}\right) + 4\sqrt{2}c_2(\log n)\sigma\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}.$$

### D.2. Outline of proof

In front of the rigorous proof in Section D.3, we first illustrate our proof strategy as

- **Step I:** We relax the wrong recovery $\left\{\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural\right\}$ to event $\mathcal{E}$, i.e. $\left\{\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^\natural\right\} \subseteq \mathcal{E}$, which reads as

$$\mathcal{E} \triangleq \left\{\left\|\mathbf{Y}_{i,:} - \mathbf{X}_{\pi^\natural(i),:}\mathbf{B}^*\right\|_2^2 \geq \left\|\mathbf{Y}_{i,:} - \mathbf{X}_{j,:}\mathbf{B}^*\right\|_2^2,\ \exists\, i, j\right\}. \tag{14}$$

  The physical meaning of $\mathcal{E}$ is that we may reduce the residual $\left\|\mathbf{Y} - \mathbf{\Pi}^\natural\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}$ by changing $\pi^\natural(i)$ to $j$. Same relaxation has been previously used in Collier and Dalalyan (2016); Slawski et al. (2019a); Zhang et al. (2019a;b).

- **Step II:** The core in this step lies in how to lower bound $\mathbb{P}(\mathcal{E}_7)$. First we decompose $\mathcal{E}$ into $\mathcal{E}_8 \bigcup \mathcal{E}_9$ with some simple algebraic manipulations. Under the SNR assumption in Eq. (7), we show both $\mathbb{P}(\mathcal{E}_8)$ and $\mathbb{P}(\mathcal{E}_9)$ are approximately $\mathbb{P}(\overline{\mathcal{E}}_7)$, as in Lemma 5 and Lemma 6, respectively.

  To show $\mathbb{P}(\overline{\mathcal{E}}_7)$ is with low probability, in another words, $\mathbb{P}(\mathcal{E}_7)$ is highly likely, we prove the following relations hold with high probability under $\mathcal{E}_6$,

$$\left\|\mathbf{X}_{i,:}\left(\widetilde{\mathbf{B}} - \mathbf{B}^\natural\right)\right\|_2 \lesssim \frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}};$$
$$\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2 \lesssim \sqrt{m}(\log n)\sigma(n+p),$$

  whose proof are in Lemma 9 and Lemma 10, respectively, and hence finish the proof by

$$\left\|\mathbf{X}_{i,:}\left(\mathbf{B}^* - \mathbf{B}^\natural\right)\right\|_2 \leq \left\|\mathbf{X}_{i,:}\left(\widetilde{\mathbf{B}} - \mathbf{B}^\natural\right)\right\|_2 + \frac{1}{n-h}\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2.$$

  In particular, we would like to mention the technique used in bounding $\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2$. First we review the widely-used bounding procedure, which proceeds as

$$\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2 \leq \|\mathbf{X}_{i,:}\|_2\|\mathbf{X}\|_2\|\mathbf{W}\|_2 \overset{①}{\lesssim} \sqrt{p\log n}\left(\sqrt{n}+\sqrt{p}\right)\sigma\left(\sqrt{n}+\sqrt{m}\right) \overset{②}{\asymp} \sqrt{\log n}(n^{3/2})\sigma + \sqrt{mn\log n}\sigma,$$

  where in ① we use the fact $\|\mathbf{X}_{i,:}\|_2 \lesssim \sqrt{p\log n}$, $\|\mathbf{X}\|_2 \lesssim \sqrt{n}+\sqrt{p}$, $\|\mathbf{W}\|_2 \lesssim \sigma(\sqrt{n}+\sqrt{m})$ hold with high probability, and in ② we use $p \asymp n$. Comparing with our results in Lemma 10, this bound experience inflations when $m \ll n$ and will lift the $\mathsf{SNR}$ requirement to $\log(\mathsf{SNR}) \gtrsim \log n$, which hides the role of $\rho(\mathbf{B}^\natural)$ compared with our current result in Theorem 2. To handle such problem, we adopt the leave-one-out trick as in El Karoui (2013; 2018); Chen et al. (2019); Sur et al. (2019) and refer to Lemma 10 for the technical details.

Having illustrated our proof strategies, we leave the detailed calculation to Section D.3.

### D.3. Proof of Theorem 2

**Proof 6** *We restate the definition of event $\mathcal{E}$ as*

$$\mathcal{E} \triangleq \left\{\left\|\mathbf{Y}_{i,:} - \mathbf{X}_{\pi^\natural(i),:}\mathbf{B}^*\right\|_2^2 \geq \left\|\mathbf{Y}_{i,:} - \mathbf{X}_{j,:}\mathbf{B}^*\right\|_2^2,\ \exists\, i, j\right\}.$$

***Step I:*** *First we verify that*

$$\widehat{\mathbf{\Pi}} = \mathrm{argmin}_{\mathbf{\Pi}}\left\|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}$$

*returns the same permutation matrix $\widehat{\mathbf{\Pi}}$ as that by Eq. (3). Hence, correct recovery of the ground truth permutation matrix $\mathbf{\Pi}^\natural$ suggests that*

$$\left\|\mathbf{Y} - \mathbf{\Pi}^\natural\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}} < \left\|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}},\ \forall\, \mathbf{\Pi} \neq \mathbf{\Pi}^\natural.$$

*Then we finish the proof by showing that* $\overline{\mathcal{E}} \subseteq \left\{\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^\natural\right\}$. *Assuming the claim is not true, which means we have matrix* $\mathbf{\Pi}$ *such that*

$$\left\|\mathbf{Y} - \mathbf{\Pi}^\natural \mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}^2 \geq \left\|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}^2,$$

*conditional on event* $\overline{\mathcal{E}}$. *Meanwhile we have*

$$\left\|\mathbf{Y} - \mathbf{\Pi}^\natural \mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}^2 = \sum_{i=1}^n \left\|\mathbf{Y}_{i,:} - \mathbf{X}_{\pi^\natural(i),:}\mathbf{B}^*\right\|_2^2 \overset{①}{<} \sum_{i=1}^n \left\|\mathbf{Y}_{i,:} - \mathbf{X}_{\pi(i),:}\mathbf{B}^*\right\|_2^2 = \left\|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}^2,$$

*which leads to contradiction, where in* ① *we use the definition of* $\overline{\mathcal{E}}$.

***Step II:*** *We verify that* $\left\|\mathbf{Y}_{i,:} - \mathbf{X}_{\pi^\natural(i),:}\mathbf{B}^*\right\|_2^2 \geq \left\|\mathbf{Y}_{i,:} - \mathbf{X}_{j,:}\mathbf{B}^*\right\|_2^2$ *is equivalent to*

$$\begin{aligned} 2\left\langle \mathbf{W}_{i,:},\ \left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^\natural(i),:}\right)\mathbf{B}^*\right\rangle \ \geq\ & \left\|\left(\mathbf{X}_{\pi^\natural(i),:} - \mathbf{X}_{j,:}\right)\mathbf{B}^\natural\right\|_2^2 + \left\|\mathbf{X}_{j,:}\left(\mathbf{B}^\natural - \mathbf{B}^*\right)\right\|_2^2 \\ & + 2\left\langle \left(\mathbf{X}_{\pi^\natural(i),:} - \mathbf{X}_{j,:}\right)\mathbf{B}^\natural, \mathbf{X}_{j,:}\left(\mathbf{B}^\natural - \mathbf{B}^*\right)\right\rangle - \left\|\mathbf{X}_{\pi^\natural(i),:}\left(\mathbf{B}^\natural - \mathbf{B}^*\right)\right\|_2^2, \end{aligned}$$

*which suggests that* $\mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{E}_8) + \mathbb{P}(\mathcal{E}_9)$ *and completes the proof with Lemma 5 and Lemma 6.*

**Lemma 5** *We have* $\mathbb{P}(\mathcal{E}_8) \leq c_0 e^{-\left((\log n)^4 \wedge (\log n)^2 \rho(\mathbf{B}^\natural)\right)} + c_1 n^{-1} + c_2 n e^{-c_3 n} + c_4 n e^{-c_0 m} + 2e^{-p} + 6p^{-2}$.

**Proof 7** *For the conciseness of notation, we define* $\Delta_1$ *and* $\Delta_2$ *as*

$$\begin{aligned} \Delta_1 =\ & 4c_0 \frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}} \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} + 4c_1\sqrt{m}(\log n)\sigma\left(1 + \frac{p}{n}\right); \\ \Delta_2 =\ & c_2(\log n)\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}. \end{aligned}$$

*Then we can bound* $\mathbb{P}(\mathcal{E}_8)$ *as*

$$\begin{aligned} \mathbb{P}(\mathcal{E}_8) \ &\overset{①}{\leq} \mathbb{P}\left(\left\|\left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^\natural(i),:}\right)\mathbf{B}^*\right\|_2 \geq \Delta_1 + \Delta_2,\ \exists\, i, j\right) + \exp\left(-\frac{\Delta^2}{2\sigma^2\left(\Delta_1 + \Delta_2\right)^2}\right) \\ &\overset{②}{\leq} \underbrace{\mathbb{P}\left(\left\|\left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^\natural(i),:}\right)\left(\mathbf{B}^* - \mathbf{B}^\natural\right)\right\|_2 \geq \Delta_1,\ \exists\, i, j\right)}_{\zeta_1} + \underbrace{\mathbb{P}\left(\left\|\left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^\natural(i),:}\right)\mathbf{B}^\natural\right\|_2 \geq \Delta_2,\ \exists\, i, j\right)}_{\zeta_2} + n^{-8}, \end{aligned} \tag{15}$$

*where in* ① *we use the independence between* $\mathbf{W}$ *and* $\mathbf{X}$ *and condition on* $\mathbf{X}$, *in* ② *we use the relation* $\Delta = 4\sqrt{2}\sigma\left(\Delta_1 + \Delta_2\right)$. *Then we will prove that* $\zeta_1 \leq \mathbb{P}(\overline{\mathcal{E}}_7)$ *and* $\zeta_2 \asymp e^{-\left((\log n)^4 \wedge (\log n)^2 \rho(\mathbf{B}^\natural)\right)}$.

***Phase I: bounding*** $\zeta_1$ *Conditional on* $\mathcal{E}_7$, *we have*

$$\begin{aligned} \left\|\left(\mathbf{X}_{j,:} - \mathbf{X}_{\pi^\natural(i),:}\right)\left(\mathbf{B}^* - \mathbf{B}^\natural\right)\right\|_2 \leq\ & \left\|\mathbf{X}_{j,:}\left(\mathbf{B}^* - \mathbf{B}^\natural\right)\right\|_2 + \left\|\mathbf{X}_{\pi^\natural(i),:}\left(\mathbf{B}^* - \mathbf{B}^\natural\right)\right\|_2 \\ \overset{③}{\leq}\ & 2c_0 \frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}} \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} + 2c_1\sqrt{m}(\log n)\sigma\left(1 + \frac{p}{n}\right) < \frac{\Delta_1}{2}, \end{aligned}$$

*and obtain* $\zeta_1 = 0$, *where* ③ *is due to the definition of* $\mathcal{E}_7$. *Then we conclude that* $\zeta_1 \leq \mathbb{P}(\overline{\mathcal{E}}_7)$.

***Phase II: bounding*** $\zeta_2$ *For* $\zeta_2$, *we upper-bound it as*

$$\begin{aligned} \zeta_2 &\overset{④}{\leq} \sum_{\pi^\natural(i), j} \mathbb{P}\left(Z \geq c_2(\log n)^2 \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2\right) \overset{⑤}{\leq} n^2 \mathbb{P}\left(|Z - \mathbb{E}Z| \geq c_3(\log n)^2 \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2\right) \\ &\overset{⑥}{\leq} n^2 \exp\left(-\left(\frac{(\log n)^4 \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^4}{\left\|\mathbf{B}^\natural \mathbf{B}^{\natural\top}\right\|_{\mathrm{F}}^2} \wedge \frac{(\log n)^2 \left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2}{\left\|\mathbf{B}^\natural \mathbf{B}^{\natural\top}\right\|_{\mathrm{OP}}}\right)\right) = n^2 e^{-\left((\log n)^4 \wedge (\log n)^2 \rho(\mathbf{B}^\natural)\right)} \\ &\asymp e^{-\left((\log n)^4 \wedge (\log n)^2 \rho(\mathbf{B}^\natural)\right)}, \end{aligned} \tag{16}$$

*where in ④ we define* $Z \triangleq \left\|\left(\mathbf{X}_{j,:}-\mathbf{X}_{\pi^\natural(i),:}\right)\mathbf{B}^\natural\right\|_2^2$*, in ⑤ we have* $\mathbb{E}Z = 4\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2$ *and use* $c_2(\log n)^2\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2 \geq \left(4+c_3(\log n)^2\right)\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2$ *when* $n$ *is sufficiently large, and in ⑥ we use the Hanson-Wright inequality (Theorem* 6.2.1 *in Vershynin (2018)). Combining Eq. (15), Eq. (16) and Lemma 8 together, we complete the proof.*

**Lemma 6** *Consider the same setting of Theorem* 2*. Provided the* SNR *satisfies*

$$\log(\mathsf{SNR}) \gtrsim \frac{6\log n}{\rho\left(\mathbf{B}^\natural\right)} + \log\log n,$$

*we have* $\mathbb{P}\left(\mathcal{E}_9\right) \leq 2e^{-p} + ne^{-c_1 m} + c_2 p^{-2} + c_3 n e^{-c_4 n}$*, when* $n$ *is sufficiently large, where* $c_i > 0$*,* $0 \leq i \leq 4$ *are some positive constants.*

**Proof 8** *We upper bound* $\mathbb{P}\left(\mathcal{E}_9\right)$ *as*

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_9\right) &\leq \mathbb{P}\left(\left\|\left(\mathbf{X}_{\pi^\natural(i),:}-\mathbf{X}_{j,:}\right)\mathbf{B}^\natural\right\|_2^2 - 2\left\|\left(\mathbf{X}_{\pi^\natural(i),:}-\mathbf{X}_{j,:}\right)\mathbf{B}^\natural\right\|_2\left\|\mathbf{X}_{j,:}\left(\mathbf{B}^\natural-\mathbf{B}^*\right)\right\|_2 - \left\|\mathbf{X}_{\pi^\natural(i),:}\left(\mathbf{B}^\natural-\mathbf{B}^*\right)\right\|_2^2 \leq \Delta,\ \exists\, i,j\right) \\
&\leq \underbrace{\mathbb{P}\left(\left\|\left(\mathbf{X}_{\pi^\natural(i),:}-\mathbf{X}_{j,:}\right)\mathbf{B}^\natural\right\|_2 \leq \delta,\ \exists\, i,j\right)}_{\triangleq\ \zeta_1} + \underbrace{\mathbb{P}\left(\frac{\left\|\mathbf{X}_{\pi^\natural(i),:}\left(\mathbf{B}^\natural-\mathbf{B}^*\right)\right\|_2^2}{\delta^2} + \frac{2\left\|\mathbf{X}_{\pi^\natural(i),:}\left(\mathbf{B}^\natural-\mathbf{B}^*\right)\right\|_2}{\delta} + \frac{\Delta}{\delta^2} \geq 1,\ \exists\, i,j\right)}_{\triangleq\ \zeta_2}.
\end{aligned}$$

*Setting* $\delta$ *as* $\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} n^{-\frac{3}{c\rho\left(\mathbf{B}^\natural\right)}}$*, we would like to show* $\zeta_1 \lesssim n^{-1}$ *and* $\zeta_2 \leq \mathbb{P}(\overline{\mathcal{E}}_7)$ *under the assumptions in Lemma 6.*

***Phase I: bounding*** $\zeta_1$ *We set* $\delta$ *as* $\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} n^{-\frac{3}{c\rho\left(\mathbf{B}^\natural\right)}}$*, and can upper bound* $\zeta_1$ *as*

$$\zeta_1 \leq \sum_{i=1}^{n}\sum_{j\neq\pi^\natural(i)} \mathbb{P}\left(\left\|\left(\mathbf{X}_{\pi^\natural(i),:}-\mathbf{X}_{j,:}\right)\mathbf{B}^\natural\right\|_2 \leq \delta\right) \overset{①}{\leq} \sum_{i=1}^{n}\sum_{j\neq\pi^\natural(i)} n^{-3} \lesssim n^{-1}, \tag{17}$$

*where ① comes from the* small ball *probability as in Lemma 2.6 in Latala et al. (2007), which is also stated as Lemma 12.*

***Phase II: bounding*** $\zeta_2$ *Then we prove that* $\zeta_2$ *can be arbitrarily small under the* SNR *requirement in Eq. (7). Conditional on event* $\mathcal{E}_7$*, we have*

$$\begin{aligned}
\frac{\left\|\mathbf{X}_{\pi^\natural(i),:}\left(\mathbf{B}^\natural-\mathbf{B}^*\right)\right\|_2^2}{\delta^2} &\leq \frac{2c_0^2\frac{p^2(\log n)^3(\log p)^2}{n}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2 + 2c_1^2 m(\log n)^2\sigma^2\left(1+p/n\right)^2}{\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2 n^{-\frac{6}{c\rho\left(\mathbf{B}^\natural\right)}}} \\
&\overset{②}{\leq} \underbrace{\frac{2c_0^2 p^2(\log n)^3(\log p)^2}{n^{1-6/(c\rho(\mathbf{B}^\natural))}}}_{\eta_1} + \underbrace{8c_1^2\frac{(\log n)^2 n^{\frac{6}{c\rho\left(\mathbf{B}^\natural\right)}}}{\mathsf{SNR}}}_{\eta_2},
\end{aligned} \tag{18}$$

*in ② we use the fact* $p \leq n$*. Since we have* $n \geq p^4(\log n)^6(\log p)^4$ *and* $\rho(\mathbf{B}^\natural) \geq 18/c$*, we conclude* $\eta_1 \to 0$ *as* $n$ *goes to infinity. Meanwhile, because of the assumptions in Eq. (7), we have* $\eta_2$ *to be a small positive constants.*

*Additionally, we can expand* $\Delta/\delta^2$ *as*

$$\begin{aligned}
\frac{\Delta}{\delta^2} &\lesssim \frac{n^{\frac{6}{c\rho\left(\mathbf{B}^\natural\right)}}\sigma}{\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}^2}\left(c_0\frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} + c_1\sqrt{m}(\log n)\sigma\left(1+\frac{p}{n}\right) + c_2(\log n)\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}\right) \\
&\lesssim c_0\frac{p(\log n)^{3/2}(\log p)}{\sqrt{mn}} \times \frac{n^{\frac{6}{c\rho\left(\mathbf{B}^\natural\right)}}}{\sqrt{\mathsf{SNR}}} + c_1\frac{\log n}{\sqrt{m}} \times \frac{n^{\frac{6}{c\rho\left(\mathbf{B}^\natural\right)}}}{\sqrt{\mathsf{SNR}}} + c_2\frac{\log n}{\sqrt{m}} \times \frac{n^{\frac{6}{c\rho\left(\mathbf{B}^\natural\right)}}}{\mathsf{SNR}}.
\end{aligned} \tag{19}$$

*Following similar procedures as above, we can prove* $\Delta/\delta^2$ *to be a small positive constant given Eq. (7). Combing Eq. (18) and Eq. (19) together, we conclude*

$$\eta_1 + \eta_2 + 2\sqrt{\eta_1+\eta_2} + \frac{\Delta}{\delta^2} < 1,$$

*which suggests that* $\zeta_2$ *equals zero conditional on events* $\mathcal{E}_7$*. Therefore, we obtain*

$$\zeta_2 \leq \mathbb{P}\left(\overline{\mathcal{E}}_7\right) \overset{③}{\leq} 2e^{-p} + 6p^{-2} + ne^{-c_0 m} + c_0 n^{-1} + c_1 n e^{-c_2 n} \overset{④}{\lesssim} 2e^{-p} + ne^{-c_0 m} + c_0 p^{-2} + c_1 n e^{-c_2 n}$$

*and completes the proof together with Eq. (17), where ③ is due to Lemma 8, and ④ is because of* $n \gtrsim p^2$*.*

### D.4. Supporting Lemmas for Theorem 2

**Lemma 7** *For arbitrary row $\mathbf{X}_{i,:}$, we have*

$$\|\mathbf{X}_{i,:}\|_2 \le 2\sqrt{p\log n},$$

*with probability exceeding $1-n^{-p}$.*

**Proof 9** *Notice that $\|\mathbf{X}_{i,:}\|_2^2$ is a $\chi^2$-RV with freedom $p$, we have*

$$\mathbb{P}\left(\|\mathbf{X}_{i,:}\|_2^2 \ge 4p\log n\right) \le \exp\left(\frac{p}{2}\left(\log(4p\log n) - 4\log n + 1\right)\right) \overset{①}{\le} \exp\left(-p\log n\right) = n^{-p},$$

*where in ① we use $2\log n \ge \log(4\log n)+1$, when $n \ge 4$.*

**Lemma 8** *We have $\mathbb{P}(\mathcal{E}_7) \ge 1 - 2e^{-p} - 6p^{-2} - ne^{-c_0 m} - c_0 n^{-1} - c_1 n e^{-c_2 n}$.*

**Proof 10** *Invoking Lemma 10, we have*

$$\begin{aligned}
&\mathbb{P}\left(\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2 \le\ c_0\sqrt{m}(\log n)\sigma\,(n+p)\,, \forall\, i\right)\\
&= 1 - \mathbb{P}\left(\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2 >\ c_0\sqrt{m}(\log n)\sigma\,(n+p)\,, \exists\, i\right)\\
&\ge 1 - \sum_i \mathbb{P}\left(\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2 >\ c_0\sqrt{m}(\log n)\sigma\,(n+p)\right)\\
&\ge 1 - n^{1-p} - ne^{-c_0 m} - n^{-1} - c_1 n e^{-c_2 n}. \qquad (20)
\end{aligned}$$

*Then we conclude*

$$\begin{aligned}
\left\|\mathbf{X}_{i,:}\left(\mathbf{B}^* - \mathbf{B}^\natural\right)\right\|_2 &\le \left\|\mathbf{X}_{i,:}\left(\widetilde{\mathbf{B}} - \mathbf{B}^\natural\right)\right\|_2 + \frac{1}{n-h}\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2\\
&\le \|\mathbf{X}_{i,:}\|_2\left\|\widetilde{\mathbf{B}} - \mathbf{B}^\natural\right\|_{\mathrm{F}} + \frac{1}{n-h}\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2\\
&\overset{①}{\le} c_0\frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} + \frac{c_1\sqrt{m}(\log n)\sigma\,(n+p)}{n-h}\\
&\overset{②}{\le} c_0\frac{p(\log n)^{3/2}(\log p)}{\sqrt{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} + \frac{4}{3}c_1\sqrt{m}(\log n)\sigma\left(1+\frac{p}{n}\right),
\end{aligned}$$

*where in ① we condition on Lemma 9 and Eq. (20), and in ② we use the fact $h \le n/4$.*

**Lemma 9** *Provided that $n \gtrsim p^2$, $h \le n/4$, we have*

$$\left\|\widetilde{\mathbf{B}} - \mathbf{B}^\natural\right\|_{\mathrm{F}} \le \sqrt{\frac{p}{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}\left(4\sqrt{6} + (\log n)(\log p)\right),$$

*with probability at least $1-2e^{-p}-6p^{-2}$ when $n, p$ are sufficiently large.*

**Proof 11** *We assume that the first $h$ rows of $\mathbf{X}$ are permuted w.l.o.g. First, we expand $\mathbf{X}^\top\mathbf{\Pi}^\natural\mathbf{X}$ as*

$$\mathbf{X}^\top\mathbf{\Pi}^\natural\mathbf{X} = \sum_{i=1}^{h}\mathbf{X}_{\pi(i),:}^\top\mathbf{X}_{i,:} + \sum_{i=h+1}^{n}\mathbf{X}_{i,:}^\top\mathbf{X}_{i,:},$$

*and obtain*

$$\begin{aligned}
&\mathbb{P}\left(\left\|\mathbf{B}^\natural-\widetilde{\mathbf{B}}\right\|_2 \geq \sqrt{\frac{p}{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}\left(4\sqrt{6}+(\log n)(\log p)\right)\right)\\
\leq\ &\mathbb{P}\left(\frac{1}{n-h}\left\|\sum_{i=1}^{h}\mathbf{X}_{\pi(i),:}^\top\mathbf{X}_{i,:}\mathbf{B}^\natural\right\|_{\mathrm{F}}+\frac{1}{n-h}\left\|\sum_{i=h+1}^{n}\left(\mathbf{X}_{i,:}^\top\mathbf{X}_{i,:}-\mathbf{I}\right)\mathbf{B}^\natural\right\|_{\mathrm{F}}\geq\sqrt{\frac{p}{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}\left(4\sqrt{6}+(\log n)(\log p)\right)\right)\\
\overset{①}{\leq}\ &\underbrace{\mathbb{P}\left(\frac{1}{n-h}\left\|\sum_{i=1}^{h}\mathbf{X}_{\pi(i),:}^\top\mathbf{X}_{i,:}\mathbf{B}^\natural\right\|_{\mathrm{F}}\geq\frac{(\log n)(\log p)\sqrt{p}}{\sqrt{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}\right)}_{\zeta_1}\\
&+\underbrace{\mathbb{P}\left(\frac{1}{n-h}\left\|\sum_{i=h+1}^{n}\left(\mathbf{X}_{i,:}^\top\mathbf{X}_{i,:}-\mathbf{I}\right)\mathbf{B}^\natural\right\|_{\mathrm{F}}\geq 4\sqrt{\frac{6p}{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}\right)}_{\zeta_2},
\end{aligned}$$

*where ① is because of the union bound. Then we separately bound $\zeta_1$ and $\zeta_2$.*

***Phase I: Bounding*** *$\zeta_1$ According to Lemma 8 in Pananjady et al. (2017a) (restated as Lemma 13), we can decompose the set $\{j:\pi(j)\neq j\}$ into three disjoint sets $\mathcal{I}_i$, $1\leq i\leq 3$, such that $j$ and $\pi(j)$ does not lie in the same set. And the cardinality of set $\mathcal{I}_i$ is $h_i$ satisfies $\lfloor h/5\rfloor\leq h_i\leq h/3$. Adopting the union bound, we can upper-bound $\zeta_1$ as*

$$\begin{aligned}
\zeta_1&\leq\sum_{i=1}^{3}\mathbb{P}\left(\frac{1}{n-h}\left\|\sum_{j\in\mathcal{I}_i}\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}\mathbf{B}^\natural\right\|_{\mathrm{F}}\geq\frac{(\log n)(\log p)\sqrt{p}}{3\sqrt{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}\right)\\
&\leq\sum_{i=1}^{3}\mathbb{P}\left(\frac{1}{n-h}\left\|\sum_{j\in\mathcal{I}_i}\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}\right\|_{\mathrm{F}}\geq\frac{(\log n)(\log p)\sqrt{p}}{3\sqrt{n}}\right).
\end{aligned}$$

*Defining $\mathbf{Z}_i$ as $\mathbf{Z}_i=\sum_{j\in\mathcal{I}_i}\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}$, we would bound the above probability by invoking the matrix Bernstein inequality (cf. Thm 7.3.1 in Tropp (2015)). First, we have*

$$\mathbb{E}\left(\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}\right)=\left(\mathbb{E}\mathbf{X}_{\pi(j),:}\right)^\top\left(\mathbb{E}\mathbf{X}_{j,:}\right)=0,$$

*due to the independence between $\mathbf{X}_{\pi(j),:}$ and $\mathbf{X}_{j,:}$. Then we upper bound $\left\|\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}\right\|_2$ as*

$$\left\|\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}\right\|_2\overset{②}{=}\left\|\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}\right\|_{\mathrm{F}}\overset{③}{=}\left\|\mathbf{X}_{\pi(j),:}\right\|_2\left\|\mathbf{X}_{j,:}\right\|_2\overset{④}{\leq}4p\log n,$$

*where ② is because $\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}$ is rank-1, ③ is due to the fact $\left\|\mathbf{u}\mathbf{v}^\top\right\|_{\mathrm{F}}^2=\mathrm{Tr}\left(\mathbf{u}\mathbf{v}^\top\mathbf{v}\mathbf{u}^\top\right)=\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2$ for arbitrary vector $\mathbf{u},\mathbf{v}\in\mathbb{R}^p$, and ④ is because of Lemma 7.*

*In the end, we compute $\mathbb{E}\left(\mathbf{Z}_i\mathbf{Z}_i^\top\right)$ and $\mathbb{E}\left(\mathbf{Z}_i^\top\mathbf{Z}_i\right)$ as*

$$\begin{aligned}
\mathbb{E}\left(\mathbf{Z}_i^\top\mathbf{Z}_i\right)&=\mathbb{E}\left(\sum_{j_1,j_2\in\mathcal{I}_i}\mathbf{X}_{\pi(j_1),:}^\top\mathbf{X}_{j_1,:}\mathbf{X}_{j_2,:}^\top\mathbf{X}_{\pi(j_2),:}\right)\overset{⑤}{=}\mathbb{E}\left(\sum_{j\in\mathcal{I}_i}\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}\mathbf{X}_{j,:}^\top\mathbf{X}_{\pi(j),:}\right)\\
&\overset{⑥}{=}\mathbb{E}\left(\sum_{j\in\mathcal{I}_i}\mathbf{X}_{\pi(j),:}^\top\mathbb{E}\left(\mathbf{X}_{j,:}\mathbf{X}_{j,:}^\top\right)\mathbf{X}_{\pi(j),:}\right)=p\left(\sum_{j\in\mathcal{I}_i}\mathbb{E}\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{\pi(j),:}\right)=ph_i\mathbf{I}_{p\times p}=\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^\top\right),
\end{aligned}$$

*where ⑤ and ⑥ is because of the fact such that $j$ and $\pi(j)$ are not within the set $\mathcal{I}_i$ simultaneously. To sum up, we invoke the matrix Bernstein inequality (cf. Thm 7.3.1 in Tropp (2015)) and have*

$$\frac{1}{n-h}\left\|\sum_{j\in\mathcal{I}}\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}\right\|_2\leq\frac{1}{3}\left(\frac{4p(\log n)(\log p)}{n-h}+\frac{p\sqrt{16(\log n)^2(\log p)^2+6h_i\log p/p}}{n-h}\right)$$

*holds with probability* $1-2p^{-2}$.

*Exploiting the fact such that* $h \le n/4$, $h_i \le h/3$, *and* $p \lesssim \sqrt{n}$, *we obtain*

$$\frac{p\sqrt{16(\log n)^2(\log p)^2+6h_i\log p/p}}{n-h} \le \frac{4p}{3n}\sqrt{16(\log n)^2(\log p)^2+\frac{n}{2p}(\log n)(\log p)} \overset{⑦}{\le} \frac{4\sqrt{p}}{3\sqrt{n}}\times(\log n)(\log p),$$

*in* ⑦ *we* $n \gtrsim p^2 \ge 32p$ *and hence*

$$\frac{1}{n-h}\left\|\sum_{j\in\mathcal{I}}\mathbf{X}_{\pi(j),:}^\top\mathbf{X}_{j,:}\right\|_2 \le (\log n)(\log p)\left(\frac{16p}{9n}+\frac{4\sqrt{p}}{9\sqrt{n}}\right) \overset{⑧}{\le} \sqrt{\frac{p}{n}}(\log n)(\log p),$$

*holds with probability exceeding* $1-6p^{-2}$, *where in* ⑧ *we use* $n \ge 256p/25$.

***Phase II: Bounding*** $\zeta_2$ *We upper bound* $\zeta_2$ *as*

$$\begin{aligned}\zeta_2 &\le \mathbb{P}\left(\frac{1}{n-h}\left\|\sum_{i=h+1}^{n}\left(\mathbf{X}_{i,:}^\top\mathbf{X}_{i,:}-\mathbf{I}\right)\mathbf{B}^\natural\right\|_{\mathrm{F}} \ge 4\sqrt{\frac{6p}{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}\right)\\ &\le \mathbb{P}\left(\frac{1}{n-h}\left\|\sum_{i=h+1}^{n}\left(\mathbf{X}_{i,:}^\top\mathbf{X}_{i,:}-\mathbf{I}\right)\right\|_{\mathrm{OP}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}} \ge 4\sqrt{\frac{6p}{n}}\left\|\mathbf{B}^\natural\right\|_{\mathrm{F}}\right) \overset{⑨}{\le} 2e^{-p}.\end{aligned}$$

*where* ⑨ *is because of* $(n-h)^{-1}\left\|\sum_{i=h+1}^{n}\left(\mathbf{X}_{i,:}\mathbf{X}_{i,:}^\top-\mathbf{I}\right)\right\|_2 \le 6\sqrt{2p/(n-h)}$ *with probability* $2e^{-p}$ *in Example 6.1 in Wainwright (2019) (also listed as Lemma 14) and* $h \le n/4$.

*The proof is completed via combing the results in Phase I and Phase II.*

**Lemma 10** *For an arbitrary index* $i$, *we have*

$$\mathbb{P}\left(\left\|\mathbf{X}_{i,:}\mathbf{X}^\top\mathbf{W}\right\|_2 \ge c_0\sqrt{m}(\log n)\sigma\left(n+p\right)\right) \le n^{-p}+e^{-c_0m}+n^{-2}+c_1e^{-c_2n}.$$

**Proof 12** *For the conciseness of notation, we define* $\delta$ *as* $c_0\sqrt{m}(\log n)\sigma\left(n+p\right)$. *In addition, we assume that* $i=1$ *w.l.o.g and prove this lemma with the* leave-one-out *trick, which is previously used in El Karoui (2013); El Karoui et al. (2013); El Karoui (2018); Chen et al. (2019); Sur et al. (2019). First we define a perturbed matrix* $\widetilde{\mathbf{X}}$ *such that* $\widetilde{\mathbf{X}}_{j,:}=\mathbf{X}_{j,:}$, $2\le j\le n$, *while* $\widetilde{\mathbf{X}}_{1,:}\in\mathbb{R}^{1\times p}$ *is a independent identically distributed Gaussian vector as* $\mathbf{X}_{1,:}$, *namely,* $\mathcal{N}\left(\mathbf{0},\mathbf{I}\right)$.

*Then we can upper-bound the probability as*

$$\begin{aligned}&\mathbb{P}\left(\left\|\mathbf{X}_{1,:}\mathbf{X}^\top\mathbf{W}\right\|_2 \ge \delta\right) \le \mathbb{P}\left(\left\|\mathbf{X}_{1,:}\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_2+\left\|\mathbf{X}_{1,:}\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^\top\mathbf{W}\right\|_2 \ge \delta\right)\\ \le\ &\underbrace{\mathbb{P}\left(\left\|\mathbf{X}_{1,:}\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^\top\mathbf{W}\right\|_2 \ge 4p\left(\log n\right)\sqrt{m}\sigma\right)}_{\zeta_1}+\underbrace{\mathbb{P}\left(\left\|\mathbf{X}_{i,:}\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_2 \ge \delta-4p\left(\log n\right)\sqrt{m}\sigma\right)}_{\zeta_2}.\end{aligned}$$

***Phase I: bounding*** $\zeta_1$ *To bound* $\zeta_1$, *easily we can verify the following relation*

$$\left\|\mathbf{X}_{1,:}\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^\top\mathbf{W}\right\|_2 \le \left\|\mathbf{X}_{1,:}\right\|_2\left\|\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^\top\mathbf{W}\right\|_{\mathrm{F}} \overset{①}{=} \left\|\mathbf{X}_{1,:}\right\|_2\left\|\mathbf{X}_{1,:}-\widetilde{\mathbf{X}}_{1,:}\right\|_2\left\|\mathbf{W}_{1,:}\right\|_2 \overset{②}{\le} 4p\left(\log n\right)\sqrt{m}\sigma.$$

*with probability exceeding* $1-n^{-p}-e^{-c_0m}$, *where* ① *is because only the first row of* $\mathbf{X}-\widetilde{\mathbf{X}}$ *is nonzero, and* ② *conditions on* $\mathcal{E}_6$ *and* $\left\|\mathbf{W}_{1,:}\right\|_2 \le 2\sqrt{m}\sigma$ *holds with probability at least* $1-e^{-c_0m}$.

***Phase II: bounding*** $\zeta_2$ *Since* $\delta-4p(\log n)\sqrt{m}\sigma \gtrsim n(\log n)\sqrt{m}\sigma$, *we can upper-bound* $\zeta_2$ *as*

$$\zeta_2 \le \mathbb{P}\left(\left\|\mathbf{X}_{i,:}\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_2 \ge c_1n(\log n)\sqrt{m}\sigma\right).$$

*Due to the construction of* $\widetilde{\mathbf{X}}$*, we have* $\mathbf{X}_{1,:}$ *to be independent of* $\widetilde{\mathbf{X}}$*. Hence, we condition on* $\widetilde{\mathbf{X}}^\top\mathbf{W}$ *and obtain*

$$\begin{aligned}\zeta_2 \le\ & \mathbb{P}\left(\left\|\mathbf{X}_{i,:}\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_2 \ge c_1 n(\log n)\sqrt{m}\sigma,\ \left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_{\mathrm{F}} < 8n\sqrt{m}\sigma\right) + \mathbb{P}\left(\left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_{\mathrm{F}} \ge 8n\sqrt{m}\sigma\right)\\ \le\ & \underbrace{\mathbb{E}_{\widetilde{\mathbf{X}}^\top\mathbf{W}}\mathbb{1}\left(\left\|\mathbf{X}_{i,:}\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_2 \ge c_2(\log n)\left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_{\mathrm{F}}\right)}_{\zeta_{2,1}} + \underbrace{\mathbb{P}\left(\left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_{\mathrm{F}} \ge 8n\sqrt{m}\sigma\right)}_{\zeta_{2,2}}.\end{aligned}$$

*For* $\zeta_{2,1}$*, we define* $Z = \left\|\mathbf{X}_{i,:}\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_2^2$ *and have*

$$\begin{aligned}\zeta_{2,1} &\le \mathbb{E}_{\widetilde{\mathbf{X}}^\top\mathbf{W}}\mathbb{1}\left(|Z-\mathbb{E}Z| \ge c_3(\log n)^2\left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_{\mathrm{F}}^2\right)\\ &\overset{③}{\le} \mathbb{E}_{\widetilde{\mathbf{X}}^\top\mathbf{W}}\exp\left(-\left(\frac{(\log n)^4\left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_{\mathrm{F}}^4}{\left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\mathbf{W}^\top\widetilde{\mathbf{X}}\right\|_{\mathrm{F}}^2} \wedge \frac{(\log n)^2\left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_{\mathrm{F}}^2}{\left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\mathbf{W}^\top\widetilde{\mathbf{X}}\right\|_{\mathrm{OP}}}\right)\right) \overset{④}{\le} n^{-2},\end{aligned}$$

*where* ③ *is because of the Hanson-Wright inequality (Theorem 6.2.1 in Vershynin (2018)), and* ④ *is due to the stable rank* $\rho(\widetilde{\mathbf{X}}^\top\mathbf{W}) \ge 1$*. Meanwhile we upper-bound* $\zeta_{2,2}$ *as*

$$\begin{aligned}\mathbb{P}\left(\left\|\widetilde{\mathbf{X}}^\top\mathbf{W}\right\|_2 \ge 8n\sqrt{m}\sigma\right) \le\ & \mathbb{P}\left(\left\|\widetilde{\mathbf{X}}\right\|_{\mathrm{OP}}\|\mathbf{W}\|_{\mathrm{F}} \ge 8n\sqrt{m}\sigma\right)\\ \overset{⑤}{\le}\ & \mathbb{P}\left(\left\|\widetilde{\mathbf{X}}\right\|_{\mathrm{OP}} \ge 2\left(\sqrt{n}+\sqrt{p}\right)\right) + \mathbb{P}\left(\|\mathbf{W}\|_{\mathrm{F}} \ge \frac{8n\sqrt{m}\sigma}{2\left(\sqrt{n}+\sqrt{p}\right)},\ \left\|\widetilde{\mathbf{X}}\right\|_{\mathrm{OP}} \le 2\left(\sqrt{n}+\sqrt{p}\right)\right)\\ \overset{⑥}{\le}\ & \mathbb{P}\left(\left\|\widetilde{\mathbf{X}}\right\|_{\mathrm{OP}} \ge 2\left(\sqrt{n}+\sqrt{p}\right)\right) + \mathbb{P}\left(\|\mathbf{W}\|_{\mathrm{F}} \ge \sqrt{2nm}\sigma\right) \overset{⑦}{\le} e^{-c_0 n} + e^{-0.8nm},\end{aligned}$$

*where* ⑤ *is because of the union bound, in* ⑥ *we use* $p \le n$*, and in* ⑦ *we use* $\|\mathbf{X}\|_{\mathrm{OP}} \ge 2\left(\sqrt{n}+\sqrt{p}\right)$ *with probability less than* $e^{-c_0 n}$ *(Chandrasekaran et al., 2012) and the fact* $\|\mathbf{W}\|_{\mathrm{F}}^2/\sigma^2$ *is a* $\chi^2$*-RV with* $nm$ *freedom, and Lemma 11.*

## E. Useful Facts

This section lists some useful facts for the sake of self-containing.

**Lemma 11** *For a* $\chi^2$*-RV* $Z$ *with* $\ell$ *freedom, we have*

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

**Lemma 12 (Small ball probability, Lemma 2.6 in Latala et al. (2007))** *Given an arbitrary fixed vector* $\mathbf{y} \in \mathbb{R}^n$*, we have*

$$\mathbb{P}\left(\|\mathbf{y}-\mathbf{A}\mathbf{g}\|_2 \le \alpha\|\mathbf{A}\|_{\mathrm{F}}\right) \le \exp\left(\kappa\log(\alpha)\varrho(\mathbf{A})\right), \quad \forall\ \alpha \in (0,\alpha_0),$$

*where* $\mathbf{g}$ *is a Gaussian RV following* $\mathcal{N}(\mathbf{0},\mathbf{I}_{n\times n})$*,* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is a non-zero matrix, and* $\alpha_0 \in (0,1)$ *and* $\kappa > 0$ *are some universal constants.*

**Lemma 13 (Lemma 8 in Pananjady et al. (2017a))** *Consider an arbitrary permutation map* $\boldsymbol{\pi}$ *with Hamming distance* $k$ *from the identity map, i.e.,* $\mathsf{d}_{\mathsf{H}}(\boldsymbol{\pi},\mathbf{I}) = k$*. We define the index set* $\{i:\ i \ne \pi(i)\}$ *and can decompose it into* 3 *independent sets* $\mathcal{I}_j$ $(1 \le j \le 3)$*, i.e.,* $i$ *and* $\pi(i)$ *are in different sets* $\mathcal{I}_j$ *for arbitrary* $i \in \{i:\ i \ne \pi(i)\}$*, such that the cardinality of each set satisfies* $|\mathcal{I}_j| \ge \lfloor k/3\rfloor \ge k/5$*.*

**Lemma 14 (Example 6.1 in Wainwright (2019))** *Let* $\mathbf{G} \in \mathbb{R}^{n_1\times n_2}$ *be generated with iid standard normal random variables, we have* $\|\mathbf{G}\|_{\mathrm{OP}} \le 4\sqrt{n_2/n_1}$*, hold with probability exceeding* $1-2e^{-n_2/2}$*.*