arXiv:2303.11233v1 [cs.IT] 20 Mar 2023

Sparse Recovery with Shuffled Labels: Statistical Limits and Practical Estimators

Hang Zhang, Ping Li

Cognitive Computing Lab
Baidu Research
10900 NE 8th ST. Bellevue, WA 98004, USA
{zhanghanghitomi, pinli98}@gmail.com

Abstract

[1]This paper considers the sparse recovery with shuffled labels, i.e., $\boldsymbol{y} = \mathbf{\Pi}^{\natural}\mathbf{X}\boldsymbol{\beta}^{\natural} + \boldsymbol{w}$, where $\boldsymbol{y} \in \mathbb{R}^n$, $\mathbf{\Pi} \in \mathbb{R}^{n\times n}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta}^{\natural} \in \mathbb{R}^p$, $\boldsymbol{w} \in \mathbb{R}^n$ denote the sensing result, the unknown permutation matrix, the design matrix, the sparse signal, and the additive noise, respectively. Our goal is to reconstruct both the permutation matrix $\mathbf{\Pi}^{\natural}$ and the sparse signal $\boldsymbol{\beta}^{\natural}$. We investigate this problem from both the statistical and computational aspects. From the statistical aspect, we first establish the minimax lower bounds on the sample number n and the *signal-to-noise ratio* (SNR) for the correct recovery of permutation matrix $\mathbf{\Pi}^{\natural}$ and the support set $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$, to be more specific, $n \gtrsim k\log p$ and $\log \mathsf{SNR} \gtrsim \log n + \frac{k\log p}{n}$. Then, we confirm the tightness of these minimax lower bounds by presenting an exhaustive-search based estimator whose performance matches the lower bounds thereof up to some multiplicative constants. From the computational aspect, we impose a parsimonious assumption on the number of permuted rows and propose a computationally-efficient estimator accordingly. Moreover, we show that our proposed estimator can obtain the ground-truth $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$ under mild conditions. Furthermore, we provide numerical experiments to corroborate our claims.

In this study, we focus on the "single measurement" problem, i.e., $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{\beta}^{\natural} \in \mathbb{R}^p$, and require SNR being at least of the order $\Omega(n^c \cdot (k \cdot \log p/n)^c)$. A recent work (Zhang and Li, 2023) studies the permuted sparse recovery problem with multiple measurements, i.e., $\boldsymbol{y} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{\beta}^{\natural} \in \mathbb{R}^{p\times m}$, with $m > 1$. They exploit the strategy of "borrowing strength" across different sets of measurements and reduce the SNR requirement for the permutation recovery. They propose a new estimator and develop a series of new techniques (including a novel modification of the "leave-one-out" method), which however do not apply to our problem (with $m = 1$) in this paper.

[1]Preliminary work appeared in Proceedings of the IEEE International Symposium on Information Theory (ISIT'21).

1 Introduction

This paper considers the linear sensing relation with shuffled labels, which can be formulated as

$$\boldsymbol{y} = \boldsymbol{\Pi}^{\natural}\mathbf{X}\boldsymbol{\beta}^{\natural} + \boldsymbol{w}.$$

Here $\boldsymbol{y} \in \mathbb{R}^n$ denotes the sensing result, $\boldsymbol{\Pi}^{\natural} \in \mathbb{R}^{n\times n}$ is the unknown permutation matrix, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the design (sensing) matrix, $\boldsymbol{\beta}^{\natural} \in \mathbb{R}^p$ represents the signals of interests, and $\boldsymbol{w} \in \mathbb{R}^n$ denotes the additive noise. In real life, we have witnessed a broad spectrum of its applications, which spans from communication to data privacy to computer vision to curve registration to natural language processing (Pananjady et al., 2018; Slawski and Ben-David, 2019; Unnikrishnan et al., 2015; Zhang et al., 2019). Two prominent examples include *linkage record*, which is to merge two datasets containing different pieces of information about the same objects into one comprehensive dataset; and *data de-anonymization*, which infers the hidden labels and can be viewed as the inverse problem of data anonymization used for privacy protection. Apart from the above applications, other applications include correspondence estimation between pose and estimation in graphics to time-domain sampling in the presence of clock jitter to multi-target tracking in radar. For a detailed discussion, we refer the interested readers to Pananjady et al. (2018); Slawski and Ben-David (2019); Unnikrishnan et al. (2015); Zhang et al. (2019); Slawski et al. (2020).

In the majority of existing works (Pananjady et al., 2018; Slawski and Ben-David, 2019; Unnikrishnan et al., 2015; Zhang et al., 2019; Slawski et al., 2020; Hsu et al., 2017; Zhang et al., 2022; Zhang and Li, 2020), their focus is usually on the regime of sufficient samples, in other words, the sample number n is larger than the signal length p (i.e., $n \geq p$). For a general case where the signal $\beta^{\natural} \in \mathbb{R}^p$ an arbitrary vector residing within the linear space $\mathbb{R}^p$, the requirement $n \geq$ seems to be inevitable, even with the perfect correspondence information, namely, the permutation matrix $\boldsymbol{\Pi}^{\natural}$. Meanwhile, the sample number can be reduced given some prior knowledge of the signal $\boldsymbol{\beta}^{\natural}$, e.g., we know that $\boldsymbol{\beta}^{\natural}$ lies within a small subspace, or equivalently, $\boldsymbol{\beta}^{\natural}$ is with a low inherent dimension. One typical example is the literature on the "compressed sensing" (or *sparse recovery*) (Donoho, 2006; Candes et al., 2006; Candès et al., 2006). Assuming the signal $\boldsymbol{\beta}^{\natural}$ is k-sparse (with k-nonzero entries), it is proved that the required sample number n can be reduced to $\Omega(k \log p)$, which is far less than p provided $k \ll p$. For other low-dimensional structures, similar results have been obtained under the names *M-estimator with regularizer*s (Negahban et al., 2012), *atomic norm*s (Chandrasekaran et al., 2012), *random convex optimization*s (Amelunxen et al., 2014), etc.

Inspired by these works, we investigate the shuffled linear sensing problem with insufficient samples, namely, $n \ll p$, by placing a parsimonious assumption on the signal $\boldsymbol{\beta}^{\natural}$. Assuming $\boldsymbol{\beta}^{\natural}$ to be k-sparse, we show that the correspondence information can be restored with $n = \Omega(k \log p)$. Notice that this order is the same as the classical works on compressed sensing/sparse recovery (Donoho, 2006; Candes et al., 2006; Candès et al., 2006) and is far less than the previously required sample number such that $n = \Omega(p)$ (Pananjady et al., 2018; Slawski and Ben-David, 2019; Hsu et al., 2017; Unnikrishnan et al., 2015).

Related work. The research on unlabeled linear regression has a long history and can be at least traced back to 70s under the name "broken sample problems" (DeGroot and Goel, 1976; Goel, 1975; Bai and Hsing, 2005; DeGroot and Goel, 1980). In recent years, we have witnessed a renaissance of the study in this area. (Unnikrishnan et al., 2015) investigate the permutation recovery under the noiseless setting, i.e., $\boldsymbol{w} = \mathbf{0}$; and establish the necessary condition $n \geq 2p$ for the general signal recovery. In Pananjady et al. (2018), the noisy observation is considered and a thorough analysis of the *maximum likelihood* (ML) estimator is presented. From the statistical perspective, it is shown that the ML estimator can reach the statistical optimality with respect to the *signal-to-noise ratio* ($\mathsf{SNR} \triangleq \|\boldsymbol{\beta}^\natural\|_2^2/\sigma^2$) requirement for correct permutation recovery, to be more specific, $\mathsf{SNR} \asymp \Omega(n^c)$. From the computational perspective, Pananjady et al. (2018) show the ML estimator is NP-hard except for the special case $p = 1$. This computational issue is later tackled by Hsu et al. (2017), where an approximation algorithm for the permutation recovery is proposed with polynomial complexity. In Slawski and Ben-David (2019), the authors take an alternative path and impose parsimonious constraints on the number of permuted rows. By viewing $(\mathbf{I} - \mathbf{\Pi}^\natural)\mathbf{X}\boldsymbol{\beta}^\natural$ as sparse outliers, (Slawski and Ben-David, 2019) reconstruct the correspondence information from the viewpoint of de-noising. In this work, we adopt a similar viewpoint in designing the estimator for practical usage. However, some modifications are required to handle the insufficient sample problem (i.e., $n \ll p$). A detailed explanation of our proposed estimator can be found in Section 5.

In Emiya et al. (2014), they consider a similar setting as ours, namely, a sparse signal $\boldsymbol{\beta}^\natural$. A branch-and-bound scheme is proposed to reconstruct the permutation matrix $\mathbf{\Pi}^\natural$. Potential drawbacks of this work include their high computational cost and the missing performance guarantee. (Zhang and Li, 2021) is the conference version of this work and proposes different practical estimators. While the estimators in Zhang and Li (2021) are rooted in the literature about the sign consistency in Lasso, Dantzig estimator, etc (Zhao and Yu, 2006; Wainwright, 2009; Meinshausen et al., 2009; Donoho et al., 2005; Rosenbaum et al., 2010; Zhang et al., 2017; Lounici, 2008; Zhang et al., 2018), the estimator in this work is more related to the study of robustness (Nguyen and Tran, 2013; Dalalyan and Thompson, 2019). Despite the above differences and their distinct looks, we should mention that they actually share the same spirit, i.e., the viewpoint of de-noising: Zhang and Li (2021) performs de-noising in an implicit way while this work takes an explicit approach. Together with the change brings a noticeable performance improvement, which is discussed in Remark 3.

Apart from the above-mentioned articles, there are other works that are worth mentioning, e.g., (Tsakiris and Peng, 2019; Haghighatshoar and Caire, 2018; Emiya et al., 2014; Zhang et al., 2019; Slawski et al., 2020; Fang and Li, 2022; Slawski and Sen, 2022). Since their connection to our work are rather loose, we only mention their name without giving detailed discussion.

Contributions. Our contributions are summarized as follows:

- We establish the statistical lower bounds for the correct recovery of $(\mathbf{\Pi}^\natural, \mathrm{supp}(\boldsymbol{\beta}^\natural))$. Different from the previous works, our work focuses on the situation with insufficient samples, i.e., $n \ll p$. Exploiting the signal sparsity, we manage to reduce the sample number n from $\Omega(p)$ to $\Omega(k \log p)$, where k denotes the sparsity number of the signal $\boldsymbol{\beta}^\natural$. As compensation, our required SNR inflates from $\log \mathsf{SNR} \gtrsim \log n$ to $\log \mathsf{SNR} \gtrsim \log n + \frac{k \log p}{n}$, which turns out to be marginal since $n \gtrsim k \log p$. Moreover, we show an exhaustive-search-based estimator can match the above lower bounds up to some multiplicative constants and thus conclude the tightness of the minimax lower bounds thereof.

- We propose a computational-friendly estimator for the recovery of $(\mathbf{\Pi}^\natural, \mathrm{supp}(\boldsymbol{\beta}^\natural))$. By imposing a parsimonious assumption on the number of permuted rows, we view $(\mathbf{I} - \mathbf{\Pi}^\natural)\mathbf{X}\boldsymbol{\beta}^\natural$ as a sparse outlier and obtain a rough estimation $\widetilde{\boldsymbol{\beta}}$ of the signal $\boldsymbol{\beta}^\natural$. Then, we reconstruct the missing correspondence based on the estimated value $\widetilde{\boldsymbol{\beta}}$. We prove that the ground-truth permutation matrix $\mathbf{\Pi}^\natural$ can be obtained under mild conditions. More importantly, we show these conditions almost match the minimax lower bounds thereof. Once the permutation matrix $\mathbf{\Pi}^\natural$ is given, we restore our problem to the classical setting of sparse recovery/compressed sensing and detect the support set of $\boldsymbol{\beta}^\natural$ accordingly.

Notations. We denote c, c_0, c' as some positive real constants. For two arbitrary real numbers, we denote $a \vee b$ as the maximum of a and b while $a \wedge b$ as the minimum. We denote $a \lesssim b$ if there exists a positive constant $c_0 > 0$ such that $a \leq c_0 b$. Similarly, we define $a \gtrsim b$ provided the inequality $a \geq c_0 b$ holds for some positive constants c_0. We write $a \asymp b$ when $a \lesssim b$ and $a \gtrsim b$ hold simultaneously.

We denote the set of possible values for the permutation matrix $\mathbf{\Pi}^\natural$ as $\mathcal{P}_n$. For an arbitrary permutation matrix $\mathbf{\Pi}$, we associate it with the operator $\pi(\cdot)$ which transforms index i to $\pi(i)$. We define the Hamming distance $\mathsf{d_H}(\cdot,\cdot)$ between two permutation matrices $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ as $\mathsf{d_H}(\mathbf{\Pi}_1, \mathbf{\Pi}_2) \triangleq \sum_{i=1}^n \mathbb{1}(\pi_1(i) \neq \pi_2(i))$. The support set $\mathrm{supp}(\boldsymbol{\beta}^\natural)$ is defined as the set of indices of non-zero entries (i.e., $\mathrm{supp}(\boldsymbol{\beta}^\natural) \triangleq \{i : \beta_i^\natural \neq 0\}$). The *signal-to-noise-ratio* (SNR) is defined as $\|\boldsymbol{\beta}^\natural\|_2^2/\sigma^2$.

2 Problem Statement

We start by giving a formal restatement of our problem. Consider the sensing relation

$$\boldsymbol{y} = \mathbf{\Pi}^\natural \mathbf{X} \boldsymbol{\beta}^\natural + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the sensing result, $\mathbf{\Pi}^\natural \in \mathbb{R}^{n\times n}$ denotes the permutation matrix, i.e., $\sum_i \mathbf{\Pi}_{ij}^\natural = \sum_j \mathbf{\Pi}_{ij}^\natural = 1$, $\mathbf{\Pi}_{ij}^\natural \in \{0,1\}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the sensing matrix, with its entries to be i.i.d. standard normal random variable, namely, $\mathbf{X}_{ij} \in \mathsf{N}(0,1)$, $\boldsymbol{\beta}^\natural \in \mathbb{R}^p$ represents the k-sparse signals, i.e., $\|\boldsymbol{\beta}^\natural\|_0 \leq k$, and $\boldsymbol{w} \in \mathbb{R}^n$ denotes the Gaussian noise following $\mathsf{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

Compared with the previous work as in Zhang et al. (2019); Slawski et al. (2020); Pananjady et al. (2018) that requires $n \geq 2p$, our work focuses on the regime where $n \leq p$. By exploiting the sparsity of the signals $\boldsymbol{\beta}^\natural$, we will show that $n = \Omega(k \log p) \ll p$ samples will be sufficient for the recovery of the permutation matrix. A graphical illustration of this paper's organization is presented in Figure 1.

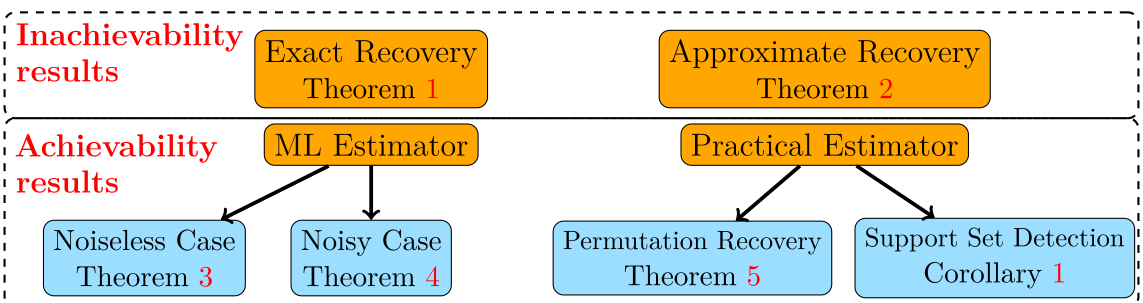

Figure 1: A diagram illustration for the roadmap of the main results to be presented in this paper. **Upper panel:** inachievability results; **Lower panel:** achievability results.

3 Statistical Lower Bounds

This paper focuses on recovering both the permutation matrix $\mathbf{\Pi}^\natural$ and the support set $\mathrm{supp}(\boldsymbol{\beta}^\natural)$, which are affected by both the sample number n and SNR. In this section, we will separately discuss their roles and establish the corresponding statistical lower bounds.

3.1 Lower bound on sample number n

To free us from the impact of SNR, we consider the oracle scenario with $\mathbf{\Pi}^{\natural}$ being known and limit ourselves to the small noise case, i.e., $\boldsymbol{w} \approxeq \mathbf{0}$. Then, our problem reduces to the classical setting of CS (Donoho, 2006; Candes et al., 2006; Candès et al., 2006), where $n \gtrsim k \log p$ is required for the Gaussian $\mathbf{X}$ (cf. P. 507, Example 15.18 in Wainwright (2019)). This order applies to our case as well since it is hopeless to recover $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$ provided we fail even in the oracle scenario with known $\mathbf{\Pi}^{\natural}$.

3.2 Lower bound on SNR

This subsection studies the mini-max lower bound for the SNR. The main result is the following.

Theorem 1. *We have*

$$\inf_{\widehat{\mathbf{\Pi}}, \widehat{\boldsymbol{\beta}}} \sup_{\mathbf{\Pi}^{\natural}, \boldsymbol{\beta}^{\natural}} \mathbb{E}_{\mathbf{X}, \boldsymbol{w}} \mathbb{1}\left[(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) \neq (\widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}))\right] \geq \frac{1}{2}, \tag{2}$$

if $n \log(1 + \mathsf{SNR}) + 2 \leq \log\left(|\mathcal{P}_n| \binom{p}{k}\right)$, *where* $\mathbb{E}_{\mathbf{X}, \boldsymbol{w}}(\cdot)$ *is taken w.r.t* $\mathbf{X}$ *and* $\boldsymbol{w}$, *and the infimum is over all possible estimators* $\widehat{\mathbf{\Pi}}$ *and* $\widehat{\boldsymbol{\beta}}$.

To better understand Theorem 1, we spell out the constants and only focus on the orders. Without any prior information about $\mathbf{\Pi}^{\natural}$, we can assume it to distribute uniformly among the set $\mathcal{P}_n$. Then, we have $|\mathcal{P}_n| = \log n!$ and can rewrite the SNR requirement in (2) as

$$\log(1 + \mathsf{SNR}) \lesssim \log n + \frac{k \log(p/k)}{n}. \tag{3}$$

Compared with Pananjady et al. (2018) which requires $\log(\mathsf{SNR}) \asymp \log n$ and $n = \Omega(p)$ for correct permutation recovery, our bound only has a slight increase of SNR requirement in (3) since $n \gtrsim k \log p$. Such an increase of required signal length is outweighed by the significant reduction of sample number, which is from $\Omega(p)$ to $\Omega(k \log p)$. In addition, we believe that this theorem can be safely relaxed to the scenario where $\mathbf{X}_{ij}$ is an i.i.d. sub-gaussian random variable with zero mean and unit variance.

The rigorous proof of Theorem 1 is in Section A.1. Here we only present an intuitive explanation, which comes from *coding theory* (Cover and Thomas, 2012). The basic idea is to recast the problem of recovering $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$ as a decoding problem (Pananjady et al., 2018; Zhang et al., 2019). First, we encode $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$ into the codeword $\mathbf{\Pi}^{\natural}\mathbf{X}\boldsymbol{\beta}^{\natural}$. Then, we pass it through the additive Gaussian channel (Cover and Thomas, 2012) and observe $\boldsymbol{y} = \mathbf{\Pi}^{\natural}\mathbf{X}\boldsymbol{\beta}^{\natural} + \boldsymbol{w}$. Our goal is to decode $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$ from the received signal $\boldsymbol{y}$. An illustration is available in Figure 2.

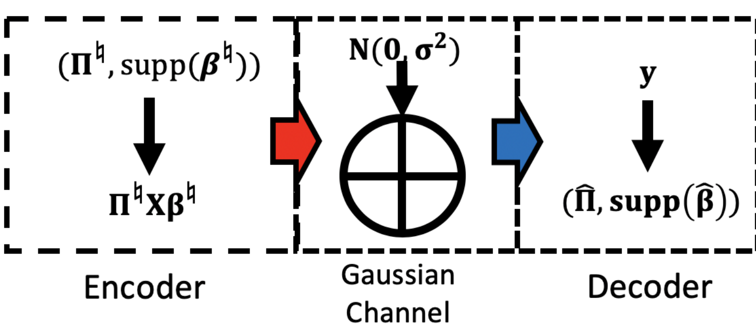

Figure 2: Interpretation of Theorem 1 from the viewpoint of coding theory.

Different from Pananjady et al. (2018), we cannot assume $\beta^\natural$ to be given a prior as this will lead to absence of sparsity number in the SNR requirement. Instead, we assume $\beta^\natural$ to be a binary vector, namely, $\beta^\natural \in \{0,1\}^n$. A bonus of this assumption is that $\mathrm{supp}(\beta^\natural)$ contains the same amount of information as β, in other words, there is no extra information (e.g., the specific values of the non-zero entries) required for encoding $(\mathbf{\Pi}^\natural, \mathrm{supp}(\beta^\natural))$ into $\mathbf{\Pi}^\natural \mathbf{X} \beta^\natural$. On one hand, the code rate Rate is computed as

$$\mathsf{Rate} \triangleq \frac{\log\left(\binom{p}{k} n!\right)}{n} \gtrsim \ (1 + {}^1\!/_{2n}) \log n + \ \frac{k\log(p/k)}{n}.$$

Meanwhile, the channel capacity is approximately

$$\mathsf{Capacity} \triangleq \frac{1}{2}\log\left(1 + \frac{\|\mathbf{\Pi}^\natural \mathbf{X}\beta^\natural\|_2^2}{n\sigma^2}\right) \approx \frac{1}{2}\log\left(1 + \frac{\|\beta^\natural\|_2^2}{\sigma^2}\right).$$

According to Cover and Thomas (2012), we require the code rate is no greater than the channel capacity, i.e., $\mathsf{Rate} < \mathsf{Capacity}$, to ensure correct recovery, which naturally yields the SNR requirement in (2).

In addition, we notice that the exact recovery may be unnecessary in certain applications. Following a similar approach, we obtain an analogous lower bound for the approximate recovery, namely, $\mathsf{d_H}(\widehat{\mathbf{\Pi}}, \mathbf{\Pi}^\natural) + \mathsf{d_H}(\mathrm{supp}(\beta^\natural), \mathrm{supp}(\widehat{\beta})) \geq \mathsf{D}$, where $\mathsf{D} \geq 0$ is some positive integer. A formal statement is given as follows.

Theorem 2. *Provided that $n\log(1+\mathsf{SNR}) + \log 4 \leq \log\zeta$, we conclude*

$$\inf_{\widehat{\mathbf{\Pi}}, \widehat{\beta}} \sup_{\mathbf{\Pi}^\natural, \beta^\natural} \mathbb{E}_{\mathbf{X}, \boldsymbol{w}} \mathbb{1}\left[\mathsf{d_H}(\widehat{\mathbf{\Pi}}, \mathbf{\Pi}^\natural) + \mathsf{d_H}(\mathrm{supp}(\beta^\natural), \mathrm{supp}(\widehat{\beta})) \geq \mathsf{D}\right] \geq \frac{1}{2},$$

where ζ is defined as

$$\zeta \triangleq \frac{p!}{(k!)^2[(p-k)!]^2} \cdot \left[\sum_{i=1}^{\mathsf{D}} \sum_{j=1}^{(\mathsf{D}-i)\wedge k} \frac{1}{(n-i)!(k-j)!(p-k-j)!(j!)^2}\right]^{-1}. \tag{4}$$

Remark 1. *Due to the complicated form of ζ in (4), we only calculate one special case, i.e., $\mathsf{D} = 0$, to illustrate its behavior. Notice that $\mathsf{D} = 0$ corresponds to the exact recovery, which restore the setting to Theorem 1. Parameter ζ under this case is written as $n!\binom{p}{k}$, which exhibits the same order as in Theorem 1.*

In this section, we have established the lower bounds, which remain valid regardless of the estimators being used. In the next section, we will confirm their tightness.

4 The Maximum Likelihood Estimator

We will show that the lower bounds in Section 3 can be matched with the differences only up to some multiplicative constants, to be more specific, (i) sample number n can be picked as $n \asymp k\log p$; and (ii) SNR can be set as $\log \mathsf{SNR} \asymp \log n + \frac{k\log(ep/k)}{n}$.

4.1 A warm-up example: the noiseless case

First, we study the role of sample number n. To free it from the impact of SNR, we consider the noiseless case, where SNR is infinite. Assuming the sparsity number k is given in advance, we will recover $(\mathbf{\Pi}^{\natural}, \boldsymbol{\beta}^{\natural})$ via the *maximal likelihood* (ML) estimator reading as

$$(\widehat{\mathbf{\Pi}}_{\mathsf{ML}}, \widehat{\boldsymbol{\beta}}_{\mathsf{ML}}) = \operatorname*{argmin}_{\substack{\mathbf{\Pi} \in \mathcal{P}_n \\ \|\boldsymbol{\beta}\|_0 \leq k}} \|\boldsymbol{y} - \mathbf{\Pi}\mathbf{X}\boldsymbol{\beta}\|_2. \tag{5}$$

Then we have the following theorem:

Theorem 3. *Consider the noiseless case where $\boldsymbol{w} = \mathbf{0}$. Given that $n \gtrsim k \log p$, we have $\mathbb{P}((\widehat{\mathbf{\Pi}}_{\mathsf{ML}}, \widehat{\boldsymbol{\beta}}_{\mathsf{ML}}) \neq (\mathbf{\Pi}^{\natural},\ \boldsymbol{\beta}^{\natural})) \lesssim n^{-2}$.*

Proof outline. We divide the analysis of the reconstruction error $(\widehat{\mathbf{\Pi}}_{\mathsf{ML}}, \widehat{\boldsymbol{\beta}}_{\mathsf{ML}}) \neq (\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$ into three categories: (i) $\{\mathbf{\Pi}^{\natural} = \widehat{\mathbf{\Pi}}_{\mathsf{ML}},\ \boldsymbol{\beta}^{\natural} \neq \widehat{\boldsymbol{\beta}}_{\mathsf{ML}}\}$; (ii) $\{\mathbf{\Pi}^{\natural} \neq \widehat{\mathbf{\Pi}}_{\mathsf{ML}},\ \boldsymbol{\beta}^{\natural} = \widehat{\boldsymbol{\beta}}_{\mathsf{ML}}\}$; and (iii) $\{\mathbf{\Pi}^{\natural} \neq \widehat{\mathbf{\Pi}}_{\mathsf{ML}},\ \boldsymbol{\beta}^{\natural} \neq \widehat{\boldsymbol{\beta}}_{\mathsf{ML}}\}$. Iterating over all possible pairs $(\mathbf{\Pi},\ \mathrm{supp}(\boldsymbol{\beta}))$, we will show the above 3 types of errors rarely happen. The detailed proof is deferred to Subsection B.2.

We notice that Theorem 3 directly recovers a sparse signal $\boldsymbol{\beta}^{\natural}$, which contains more information than its support set $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$. Moreover, we observe that the sample number requirement in Theorem 3 is well aligned with the lower bound presented in Subsection 3.1 and can hence confirm its tightness.

4.2 The noisy case

This subsection investigates the noisy case. To correctly recover the support set $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$, we need an additional assumption on non-zero entries' smallest magnitudes. Otherwise, even in the classical setting without any permutation, small sensing noise can lead to incorrect support set detection. The formal statement reads as follows.

Theorem 4. *Provided that*

(i) $n \gtrsim k \log p$,

(ii) $\log \mathsf{SNR} \gtrsim \log n + \frac{k}{n} \log(\frac{ep}{k})$

(iii) $\min_{i \in \mathrm{supp}(\boldsymbol{\beta}^{\natural})} {|\beta^{\natural}_i|^2}/{\sigma^2} \gtrsim 1$,

we have $\mathbb{P}((\widehat{\mathbf{\Pi}}_{\mathsf{ML}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{ML}})) \neq (\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) \lesssim n^{-2} + e^{-ck\log p}$.

Compared with the result pertaining to the noiseless case (i.e., Theorem 3), Theorem 4 requires non-zero entries' magnitudes to be at least some positive constants. Apart from this constraint, our bound matches the minimax lower bound in Theorem 1 up to some multiplicative constants, specifically, $n \gtrsim k \log p$ and $\log \mathsf{SNR} \gtrsim \log n + {k \log p}/{n}$.

Remark 2. *With a slight modification of the ML estimator, we can significantly relax the third assumption in the above theorem, i.e., $\min_{i \in \mathrm{supp}(\boldsymbol{\beta}^{\natural})} {|\beta^{\natural}_i|^2}/{\sigma^2} \gtrsim 1$. Notice that the correct permutation recovery only requires the first two assumptions in Theorem 4. Hence, we can first reconstruct the permutation matrix $\mathbf{\Pi}^{\natural}$ with the ML estimator. Afterwards, we restore the shuffled sparse recovery problem to its classical setting and invoke the previous works to detect the support set $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$. With the above modifications, we can improve the assumption $\min_{i \in \mathrm{supp}(\boldsymbol{\beta}^{\natural})} {|\beta^{\natural}_i|^2}/{\sigma^2} \gtrsim 1$ to $\min_{i \in \mathrm{supp}(\boldsymbol{\beta}^{\natural})} {|\beta^{\natural}_i|^2}/{\sigma^2} \gtrsim {\log p}/{n}$.*

We would like to emphasize that the ML estimator only serves to confirm the tightness of Theorem 1. It is unpractical due to its high computational cost: it needs to iterate (i) all possible k-sparse subsets, which consists of $\binom{p}{k}$ cases; and (ii) all possible permutation matrices $\mathbf{\Pi}$, which consists $n!$ cases. The next section will present a computational-friendly estimator.

5 Practical Estimator

This section proposes a practical estimator to combat the high computational cost associated with the ML estimator, which consists of two stages: permutation recovery and support set detection. A formal statement is in Algorithm 1.

5.1 Permutation Recovery

We note that the major difficulty in the permutation recovery stems from the missing value of signal $\boldsymbol{\beta}^{\natural}$. One natural solution is to restore the permutation with an approximate value of $\boldsymbol{\beta}^{\natural}$. To begin with, we impose a parsimonious assumption on the number of permuted rows, or equivalently, only a small proportion of rows are permuted. Then, we adopt a denoising viewpoint and treat $\left(\mathbf{I} - \mathbf{\Pi}^{\natural}\right)\mathbf{X}\boldsymbol{\beta}^{\natural}$ as a sparse outlier to be removed. Inspired by Nguyen and Tran (2013) and Slawski and Ben-David (2019), we can estimate the signal $\boldsymbol{\beta}^{\natural}$ as

$$(\widetilde{\mathbf{\Xi}}, \widetilde{\boldsymbol{\beta}}) = \operatorname{argmin}_{\mathbf{\Xi},\boldsymbol{\beta}} \ \frac{1}{2n}\left\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \sqrt{n}\cdot\mathbf{\Xi}\right\|_2^2 + \lambda_{\mathbf{\Xi}}\|\mathbf{\Xi}\|_1 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1.$$

Afterwards, we reconstruct the permutation matrix $\mathbf{\Pi}^{\natural}$ via the following *linear assignment problem* (LAP), which reads as

$$\widehat{\mathbf{\Pi}} = \operatorname{argmax}_{\mathbf{\Pi}}\langle \boldsymbol{y},\ \mathbf{\Pi}\mathbf{X}\widetilde{\boldsymbol{\beta}}\rangle,$$

where $\widehat{\boldsymbol{\beta}}$ is the solution of (6). Then, we conclude

Theorem 5. *We set $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\mathbf{\Xi}}$ in (6) as $c_0\sigma\sqrt{\log p/n}$ and $c_1\sigma\sqrt{\log n/n}$, respectively. Assuming that (i) $n \gtrsim k\log p$, (ii) $h \lesssim \frac{n}{\log n}$, and (iii)*

$$\mathsf{SNR} \gtrsim \frac{n^{2(1+\varepsilon)}(n-1)^2}{4\pi}\left[\sqrt{\log np}\left(k\sqrt{\frac{\log p}{n}} \vee h\sqrt{\frac{\log n}{n}}\right) + 2\log(n^{1+\varepsilon}(n-1))\right]^2,$$

we conclude that (7) can yield the ground truth with probability exceeding $1-2n^{-\varepsilon}$, i.e., $\mathbb{P}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^{\natural}) \geq 1-2n^{-\varepsilon}$.

First, we discuss the SNR requirement. From the above theorem, we conclude that the correct permutation matrix can be obtained provided that $\log\mathsf{SNR} \gtrsim \log n$, which matches the mini-max lower bound in Theorem 1 up to some multiplicative constant. Then, we consider the maximum allowed number of permuted rows, i.e., $h \lesssim \frac{n}{\log n}$. Compared with the optimal order $O(n)$, we experience a loss of logarithmic term. This is consistent with our parsimonious assumption on the number of permuted rows, i.e., $h \ll n$. Moreover, we discuss the minimum required sample number n, which is of the order $\Omega(k\log p)$. Notice that this is the same as the mini-max bound discussed in Subsection 3.1.

5.2 Support Set Detection

Once we have the correct permutation matrix $\mathbf{\Pi}^{\natural}$, we can restore (1) to the classical model and detect the support set $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$ with the Lasso estimator, which is written as

$$\widehat{\boldsymbol{\beta}}_{\mathsf{Lasso}} = \operatorname{argmin}_{\boldsymbol{\beta}} \frac{1}{2n}\|\widehat{\mathbf{\Pi}}^{\top}\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_{\mathsf{Lasso}(n)}\|\boldsymbol{\beta}\|_1,$$

where $\widehat{\mathbf{\Pi}}$ denotes the solution of (7). Then, we detect the support set $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$ by selecting the entries with the first k-largest magnitude. With the standard results concerning the sign consistency of Lasso estimator, e.g., Lounici (2008), we can show the support set can be detected with high probability under the settings of Theorem 5.

Corollary 1. *Under the settings of Theorem 5, we pick $\lambda_{\mathsf{Lasso}(n)}$ in (8) as $c\sigma\sqrt{\log p/n}$. Provided that $\min_{\boldsymbol{\beta}_i^{\natural}\neq 0}(|\boldsymbol{\beta}_i^{\natural}|^2/\sigma^2) \gtrsim \frac{\log p}{n}$, we have $\mathrm{sign}(\mathsf{thres}(\widehat{\boldsymbol{\beta}}; k)) = \mathrm{sign}(\boldsymbol{\beta}^{\natural})$ hold with probability $1 - o(1)$, where $\mathsf{thres}(\cdot; k)$ selects the entries with the first k-largest magnitude and is defined in (9).*

This corollary suggests that the support set $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$ can be detected with high probability. Compared with Theorem 5, Corollary 1 has one additional assumption on the smallest magnitude of the non-zero entries in $\boldsymbol{\beta}^{\natural}$, e.g., $\min_{\boldsymbol{\beta}_i^{\natural}\neq 0}(|\boldsymbol{\beta}_i^{\natural}|^2/\sigma^2) \gtrsim {}^{\log p}/_{n}$. Notice that this assumption is quite standard (Lounici, 2008; Zhao and Yu, 2006; Wainwright, 2009) in studying the property of sign consistency.

Remark 3. *In Zhang and Li (2021), we need*

$$\min_{\boldsymbol{\beta}_i^{\natural}\neq 0} |\boldsymbol{\beta}_i^{\natural}| \gtrsim (1 + k\sqrt{{}^{\log p}/_{n}})\sqrt{{}^{\log p}/_{n}} \cdot \left(\|\boldsymbol{\beta}^{\natural}\|_2\ \sqrt{h\log n} \vee \sigma\right).$$

Meanwhile, our estimator improves this requirement to

$$\min_{\boldsymbol{\beta}_i^{\natural}\neq 0} |\boldsymbol{\beta}_i^{\natural}| \gtrsim \sigma\sqrt{\frac{\log p}{n}}.$$

Compared with Zhang and Li (2021), our estimator has a significant improvement. First, our assumption on $\min_{\boldsymbol{\beta}_i^{\natural}\neq 0} |\boldsymbol{\beta}_i^{\natural}|$ is free from the total energy $\|\boldsymbol{\beta}^{\natural}\|_2$. Even after we factor out the impact of $\|\boldsymbol{\beta}^{\natural}\|_2$, (Zhang and Li, 2021) still requires $\min_{\boldsymbol{\beta}_i^{\natural}\neq 0} |\boldsymbol{\beta}_i^{\natural}| \gtrsim \frac{\sigma(k\log p)}{n}$ while our estimator reduces the requirement to $\min_{\boldsymbol{\beta}_i^{\natural}\neq 0} |\boldsymbol{\beta}_i^{\natural}| \gtrsim \sigma\sqrt{{}^{\log p}/_{n}}$.

In the end, we will briefly discuss the potential methods of recovering $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$. Notice that Algorithm 1 only consists of one step of permutation recovery and support set detection. One natural way for the performance improvement is to iteratively perform the permutation recovery and the support set detection. In addition, we find that $\widetilde{\mathbf{\Xi}}$ in (6) is largely ignored. Since it contains information about $(\mathbf{I} - \mathbf{\Pi}^{\natural})\mathbf{X}\boldsymbol{\beta}^{\natural}$, in other words, it has information about the permutation matrix, we can use it to refine the reconstructed permutation.

Algorithm 1 Permuted-Lasso Estimator.

- **Input:** observation $\boldsymbol{y}$, sensing matrix $\mathbf{X}$, and sparsity number k.
- **Stage I: Permutation Recovery.** We pick λ_{β} and λ_{Ξ} as $c_0\sigma\sqrt{\log p/n}$ and $c_1\sigma\sqrt{\log n/n}$. We restore the correspondence information as

$$(\widetilde{\boldsymbol{\Xi}}, \widetilde{\boldsymbol{\beta}}) = \operatorname{argmin}_{\boldsymbol{\Xi},\boldsymbol{\beta}} \frac{1}{2n}\left\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \sqrt{n}\cdot\boldsymbol{\Xi}\right\|_2^2 + \lambda_{\Xi}\|\boldsymbol{\Xi}\|_1 + \lambda_{\beta}\|\boldsymbol{\beta}\|_1; \tag{6}$$

$$\widehat{\boldsymbol{\Pi}} = \operatorname{argmax}_{\boldsymbol{\Pi}}\langle \boldsymbol{y}, \boldsymbol{\Pi}\mathbf{X}\widetilde{\boldsymbol{\beta}}\rangle. \tag{7}$$

- **Stage II: Support Set Detection.** With the permutation matrix $\widehat{\boldsymbol{\Pi}}$ in (7), we pick $\lambda_{\mathsf{Lasso}(n)}$ in (8) as $c_3\sigma\sqrt{\log p/n}$ and detect the support set by first (i) computing $\widehat{\boldsymbol{\beta}}_{\mathsf{Lasso}}$ as

$$\widehat{\boldsymbol{\beta}}_{\mathsf{Lasso}} = \operatorname{argmin}_{\boldsymbol{\beta}} \frac{1}{2n}\left\|\widehat{\boldsymbol{\Pi}}^{\top}\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \lambda_{\mathsf{Lasso}(n)}\|\boldsymbol{\beta}\|_1, \tag{8}$$

and then (ii) performing hard-thresholding to $\widehat{\boldsymbol{\beta}}_{\mathsf{Lasso}}$, which is

$$(\mathsf{thres}(\widehat{\boldsymbol{\beta}}_{\mathsf{Lasso}}; k))_i \triangleq \begin{cases} (\widehat{\boldsymbol{\beta}}_{\mathsf{Lasso}})_i, \text{ if } |(\widehat{\boldsymbol{\beta}}_{\mathsf{Lasso}})_i| \text{ is among the } k\text{-largest absolute entries;} \\ 0, \text{ otherwise.} \end{cases} \tag{9}$$

- **Output:** we return $(\widehat{\boldsymbol{\Pi}}, \mathsf{thres}(\widehat{\boldsymbol{\beta}}_{\mathsf{Lasso}}; k))$.

6 Simulations

This section presents the numerical results, where the permutation matrix $\boldsymbol{\Pi}^{\natural}$ and the support set $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$ are reconstructed via Algorithm 1. The regularizer coefficients, i.e., $\lambda_{\beta}, \lambda_{\Xi}$, and $\lambda_{\mathsf{Lasso}(n)}$, are all picked as 2.0. First, we consider the Gaussian setting, where each entry $\mathbf{X}_{ij}$ are i.i.d. standard normal random variables, i.e., $\mathsf{N}(0,1)$. Moreover, we extend it to the setting of sub-gaussian distributions, where $\mathbf{X}_{ij}$ are i.i.d. sub-gaussian random variables, to be more specific, $\mathbf{X}_{ij}$ are uniformly distributed within the region $[-1,1]$, namely, $\mathbf{X}_{ij} \overset{\text{i.i.d}}{\sim} \text{Unif}[-1,1]$.

We evaluate the performance in terms of the ratio ${}^{\log \mathsf{SNR}}/_{\log n}$, which is widely used in the study of permuted linear regression. We only plot the correct rate for the permutation recovery, since the support set detection in (8) and (9) seldom makes any mistake, even with a wrong permutation matrix $\widehat{\boldsymbol{\Pi}}$ returned in (7).

6.1 Impact of sparsity number

This subsection studies the impact of sparsity number k. We fix the signal length p and the permuted row number h as 500 and 20, respectively. We let the sample number $n \in \{180, 200, 220\}$ and vary the sparsity number k within the set $\{5, 10, 20\}$. The numerical results are put in Figure 3.

Discussion. First, we discuss the Gaussian setting. From the curves in Figure 3, we confirm the correctness of Theorem 5, which suggests that the correct permutation matrix can be obtained once $\log \mathsf{SNR} \gtrsim \log n$. In addition, we notice that the correct permutation reconstruction requires a larger SNR with an increasing sparsity number k. For example, we can obtain the ground-truth permutation matrix with $\log \mathsf{SNR} = 5.5\log n$ when $(n, p, h, k) = (180, 500, 20, 5)$. When the sparsity number k increases to 20, the requirement for the correct permutation recovery increases to $\log \mathsf{SNR} > 6\log n$. Similar phenomena can be observed for other settings as well. Second, we discuss the uniform

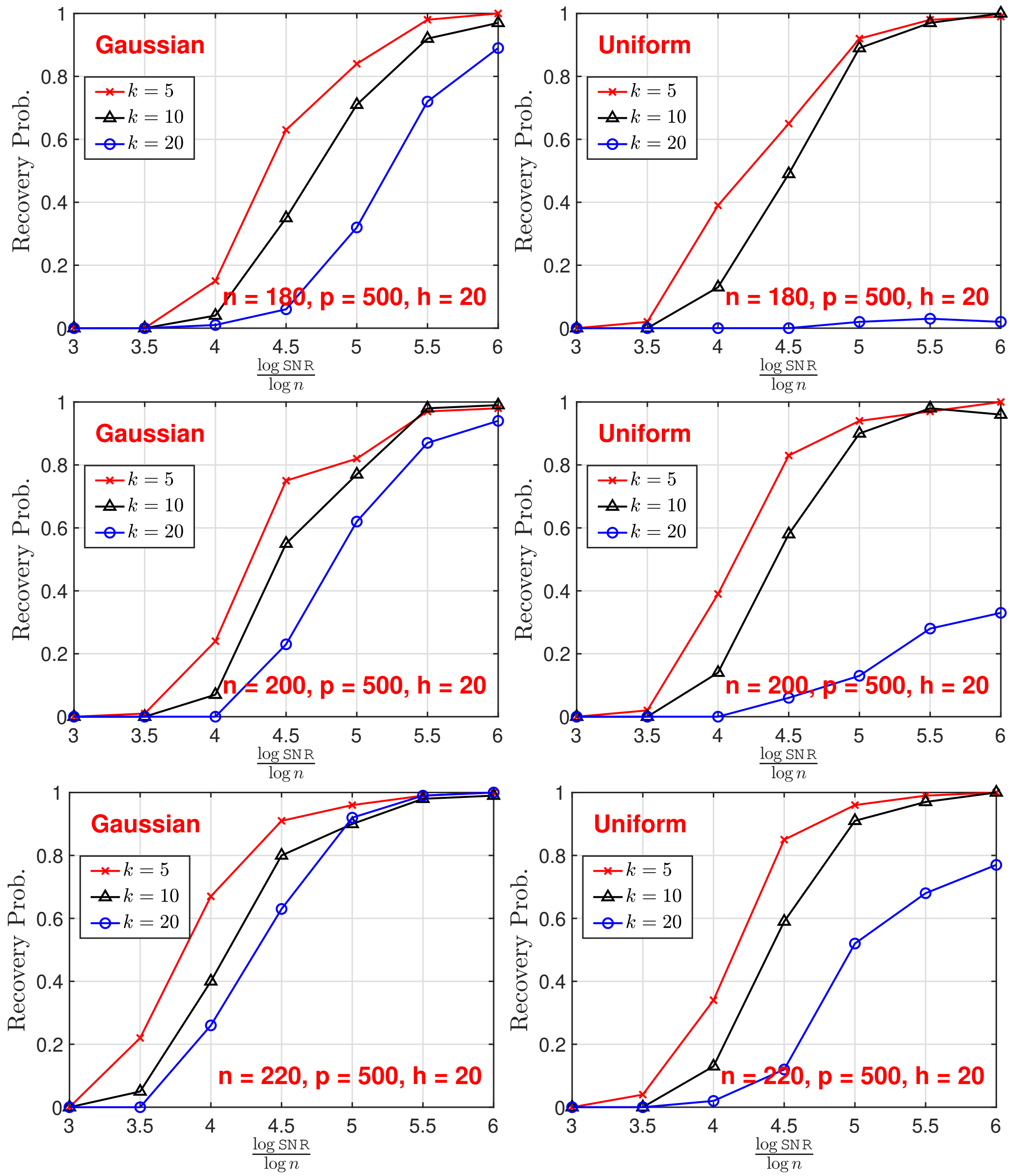

Figure 3: Simulated permutation recovery rate $\mathbb{P}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^{\natural})$ with $n = \{180, 200, 220\}$, $p = 500$, $h = 20$, and $k = \{5, 10, 20\}$, w.r.t. ${}^{\log \mathsf{SNR}}/{}_{\log n}$. (**Left Panel**) We have $\mathbf{X}_{ij}$ be i.i.d. normal random variables, i.e., $\mathbf{X}_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{N}(0, 1)$; (**Right Panel**) We have $\mathbf{X}_{ij}$ be i.i.d. sub-gaussian random variables, to be more specific, $\mathbf{X}_{ij} \overset{\text{i.i.d}}{\sim} \text{Unif}[-1, 1]$.

distribution setting. Numerical results show a similar behavior as that of the Gaussian setting and suggest that our estimator in Algorithm 1 can work beyond the setting in Theorem 5.

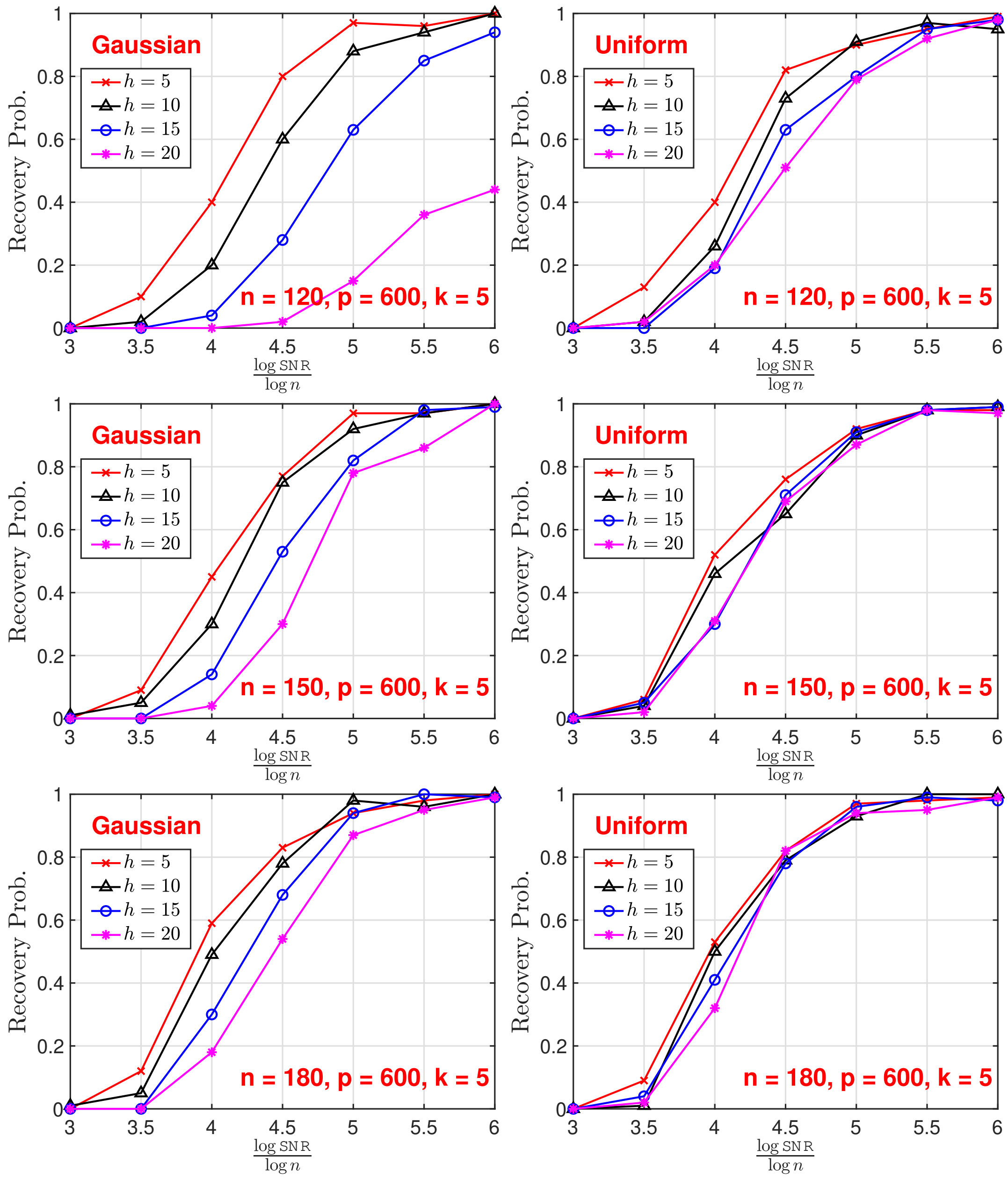

Figure 4: Simulated permutation recovery rate $\mathbb{P}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^{\natural})$ with $n = \{120, 150, 180\}$, $p = 600$, $k = 5$, and $h = \{5, 10, 15, 20\}$, w.r.t. ${}^{\log \mathsf{SNR}}/_{\log n}$. (**Left Panel**) We have $\mathbf{X}_{ij}$ be i.i.d. normal random variables, i.e., $\mathbf{X}_{ij} \overset{\text{i.i.d}}{\sim} \mathsf{N}(0,1)$; (**Right Panel**) We have $\mathbf{X}_{ij}$ be i.i.d. sub-gaussian random variables, to be more specific, $\mathbf{X}_{ij} \overset{\text{i.i.d}}{\sim} \text{Unif}[-1,1]$.

6.2 Impact of permuted row number

We investigate the impact of the permuted row number h on the simulated permutation recovery rate $\mathbb{P}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^{\natural})$. We fix the signal length p and sparsity number k as 600 and 5, respectively. We let the sample number $n = \{120, 150, 180\}$ and set the permuted row number $h \in \{5, 10, 15, 20\}$. The experiment results are presented in Figure 4.

Discussion. We notice that the permutation recovery becomes more difficult, in other words, requires a larger SNR, with an increasing number of permuted rows. Under the Gaussian setting $(n, p, h, k) = (120, 600, 5, 5)$, we can obtain the ground-truth $\mathbf{\Pi}^{\natural}$ when $\log \mathsf{SNR} \approx 5 \log n$. When h increases to 20, the requirement on SNR is strengthened to $\log \mathsf{SNR} > 6 \log n$. We believe that this conclusion should hold universally. However, numerical experiments do suggest that the performance difference becomes less distinguishable with a higher n/p ratio.

7 Conclusion

We have studied sparse recovery with shuffled labels. First, we establish the statistical lower bounds for both the sample number n and SNR. For the sample number n, by exploiting the sparsity of signals, we manage to reduce the required sample number from $n \geq 2p$ to the order of $\Omega(k \log p)$. For SNR, we have a marginal increase from $\log \mathsf{SNR} \gtrsim \log n$ to $\log \mathsf{SNR} \gtrsim \log n + {}^{k}\!/_{n} \log(\frac{ep}{k})$. Then, we present an exhaustive-search based estimator to confirm the tightness of the above bounds. Afterwards, we propose a practical estimator and show they can yield the correct $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$ under mild conditions. Simulations confirm our theorems and suggest that large sparsity number and Hamming distance require more samples and stronger signal energy for correct reconstruction of $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$.

References

Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

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

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

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

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

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

Arnak S. Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13188–13198, Vancouver, Canada, 2019.

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

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

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

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

David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.

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

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

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

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

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

Karim Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.

Nicolai Meinshausen, Bin Yu, et al. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Nam H. Nguyen and Trac D. Tran. Robust lasso with missing and grossly corrupted observations. *IEEE Trans. Inf. Theory*, 59(4):2036–2058, 2013.

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

Mathieu Rosenbaum, Alexandre B Tsybakov, et al. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

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

Martin Slawski and Bodhisattva Sen. Permuted and unlinked monotone regression in $\mathbb{R}^d$: an approach based on mixture modeling and optimal transport. *arXiv preprint arXiv:2201.03528*, 2022.

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

Manolis C. Tsakiris and Liangzu Peng. Homomorphic sensing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *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*, volume 47. Cambridge university press, 2018.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theory*, 55(5):2183–2202, 2009.

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

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

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

Hang Zhang and Ping Li. One-step estimator for permuted sparse recovery. *preprint*, 2023.

Hang Zhang, Afshin Abdi, and Faramarz Fekri. Recovery of sign vectors in quadratic compressed sensing. In *Proceedings of the IEEE Information Theory Workshop (ITW)*, pages 379–383, Kaohsiung, Taiwan, 2017.

Hang Zhang, Afshin Abdi, and Faramarz Fekri. Sparse recovery of sign vectors under uncertain sensing matrices. In *Proceedings of the IEEE Information Theory Workshop (ITW)*, pages 1–5, Guangzhou, China, 2018.

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

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

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

Appendices

A Proof of Statistical Lower Bound

A.1 Proof of Theorem 1

Proof. To begin with, we assume $\boldsymbol{\beta}^{\natural} \in \{0,1\}^p$ and place a uniform distribution prior on $\boldsymbol{\Pi}^{\natural}$ and $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$. Then, we notice the relation

$$
\begin{aligned}
&\sup_{\boldsymbol{\Pi}^{\natural},\boldsymbol{\beta}^{\natural}} \mathbb{E}_{\mathbf{X},\boldsymbol{w}} \mathbb{1}\left[(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) \neq (\widehat{\boldsymbol{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}))\right] \\
\geq\ & \mathbb{P}_{\mathbf{X},\boldsymbol{w},\boldsymbol{\Pi}^{\natural},\mathrm{supp}(\boldsymbol{\beta}^{\natural})}\left[(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) \neq (\widehat{\boldsymbol{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}))\right] \triangleq \vartheta,
\end{aligned}
\tag{10}
$$

where $\mathbb{E}_{\mathbf{X},\boldsymbol{w}}(\cdot)$ denotes the expectation w.r.t. $\mathbf{X}$ and $\boldsymbol{w}$, and $\mathbb{P}_{\mathbf{X},\boldsymbol{w},\boldsymbol{\Pi}^{\natural},\mathrm{supp}(\boldsymbol{\beta}^{\natural})}$ puts uniform prior on $\boldsymbol{\Pi}^{\natural}$ and $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$ as well. Since (10) holds universally, we can safely add $\inf_{\widehat{\boldsymbol{\Pi}},\widehat{\boldsymbol{\beta}}}$ to the left-hand side in (10) and complete the proof. In the following context, we lower bound the error probability by adapting the techniques used in proving the Fano's inequality in Theorem 2.10.1 in Cover and Thomas (2012).

Denote $\mathsf{H}(\cdot)$ as the entropy, while $\mathrm{I}(\cdot;\cdot)$ as the mutual information. With the Fano's method as illustrated in Cover and Thomas (2012), we would like to lower bound the error probability $\mathbb{P}((\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) \neq (\widehat{\boldsymbol{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}})))$ as

$$
\begin{aligned}
&\mathsf{H}(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) = \mathsf{H}\left(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \mathbf{X}\right) \\
=\ & \mathsf{H}(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \mathbf{X}, \widehat{\boldsymbol{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}})) + \mathrm{I}(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}); \widehat{\boldsymbol{\Pi}}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \mathbf{X}) \\
\overset{①}{\leq}\ & \mathsf{H}(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) | \widehat{\boldsymbol{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}})) + \mathrm{I}(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}); \widehat{\boldsymbol{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}) | \mathbf{X}) \\
\overset{②}{\leq}\ & \mathsf{H}(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \widehat{\boldsymbol{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}})) + \ \mathrm{I}\left(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}); \boldsymbol{y} \mid \mathbf{X}\right) \\
\overset{③}{\leq}\ & 1 + \log\left(|\mathcal{P}_n| \times \binom{p}{k}\right)\vartheta + \mathrm{I}\left(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}); \boldsymbol{y} | \mathbf{X}\right),
\end{aligned}
\tag{11}
$$

where ① is because of the property such that conditioning reduces entropy (Cover and Thomas, 2012, Eq. (2.157)), ② is due to the fact $(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) \to \boldsymbol{y} \to (\widehat{\boldsymbol{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}))$ forms a Markov chain and the data-processing inequality (Cover and Thomas, 2012, Theorem 2.8.1); and ③ is a direct consequence of Fano's inequality (Cover and Thomas, 2012, Theorem 2.10.1). Exploiting the independence between $\boldsymbol{\Pi}^{\natural}$ and $\boldsymbol{\beta}^{\natural}$, we have

$$
\mathsf{H}(\boldsymbol{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) = \mathsf{H}(\boldsymbol{\Pi}^{\natural}) + \mathsf{H}(\mathrm{supp}(\boldsymbol{\beta}^{\natural})) = \log|\mathcal{P}_n| + \log\binom{p}{k}.
$$

Combing (11) with Lemma 1 then complete the proof. □

A.2 Proof of Theorem 2

Proof. We assume $\mathbf{\Pi}^{\natural}$ is uniformly distributed over the set $\mathcal{P}_n$, which corresponds to the case where no prior knowledge about $\mathbf{\Pi}^{\natural}$ is unavailable. First, we define $\mathcal{E} \triangleq \mathbb{1}\{\mathsf{d_H}(\widehat{\mathbf{\Pi}}, \mathbf{\Pi}^{\natural}) + \mathsf{d_H}\left(\mathrm{supp}(\boldsymbol{\beta}^{\natural}), \mathrm{supp}(\widehat{\boldsymbol{\beta}})\right) \geq \mathsf{D}\}$, which indicates the failure of approximate recovery of $\mathbf{\Pi}^{\natural}$. We give a roadmap before going into the details

- **Step I:** We consider the conditional entropy $\mathsf{H}(\mathcal{E}, \mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) | \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X})$ and prove

$$\mathsf{H}(\mathcal{E}, \mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X}) = \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \boldsymbol{y}, \mathbf{X}).$$

- **Step II:** We show that

$$\mathsf{H}(\mathcal{E}, \mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X}) \leq \log 2 + \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) - \mathbb{P}(\mathcal{E} = 0) \log \zeta,$$

  where ζ is defined in (4).

- **Step III:** Combining the above two steps together, we upper-bound $\mathbb{P}(\mathcal{E} = 0)$ as

$$\mathbb{P}\left(\mathcal{E} = 0\right) \leq \frac{\log 2 + \mathrm{I}\left(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}); \boldsymbol{y}, \mathbf{X}\right)}{\log \zeta} \overset{①}{=} \frac{\log 2 + \mathrm{I}\left(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}); \boldsymbol{y} \mid \mathbf{X}\right)}{\log \zeta},$$

  where ① is because $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$ and $\boldsymbol{y}$ are independent given $\mathbf{X}$. Invoking Lemma 1, we complete the proof.

Then we present the computational details.

Step I. We expand $H(\mathcal{E}, \mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X})$ via the chain rule (Cover and Thomas, 2012, Theorem 2.5.1) as

$$\begin{aligned}
& \mathsf{H}(\mathcal{E}, \mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X}) \\
= \ & \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X}) + \mathsf{H}(\mathcal{E} \mid \mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}), \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X}) \\
= \ & \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \boldsymbol{y}, \mathbf{X}),
\end{aligned}$$

where in the last equation we have used that ① $\mathcal{E}$ is deterministic conditional on $\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}), \widehat{\mathbf{\Pi}}, \boldsymbol{y}, \mathbf{X}$, and ② $(\mathbf{\Pi}^{\natural}, \boldsymbol{\beta}^{\natural})$ and $(\widehat{\mathbf{\Pi}}, \widehat{\boldsymbol{\beta}})$ are independent given $\mathbf{X}$ and $\boldsymbol{y}$.

Step II. Define $\mathbb{B}\left[\left(\mathbf{\Pi}^{\natural}, \boldsymbol{\beta}^{\natural}\right); \mathsf{D}\right]$ as

$$\mathbb{B}\left[\left(\mathbf{\Pi}^{\natural}, \boldsymbol{\beta}^{\natural}\right); \mathsf{D}\right] \triangleq \left\{ (\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) \middle| \begin{array}{l} \mathsf{d_H}(\widehat{\mathbf{\Pi}}, \mathbf{\Pi}^{\natural}) = i, \\ \mathsf{d_H}(\mathrm{supp}(\boldsymbol{\beta}^{\natural}), \mathrm{supp}(\widehat{\boldsymbol{\beta}})) = j, \\ \text{s.t.} \quad i + j \leq \mathsf{D} \end{array} \right\}$$

which denotes the set of all possible pairs $(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}))$ given $\mathcal{E} = 0$. Easily we can verify that its cardinality can be upper bounded by

$$\left|\mathbb{B}\left[\left(\mathbf{\Pi}^{\natural}, \boldsymbol{\beta}^{\natural}\right); \mathsf{D}\right]\right| \leq \sum_{i=1}^{\mathsf{D}} \sum_{j=1}^{(\mathsf{D}-i)\wedge k} \binom{n}{i} i! \cdot \binom{k}{j} \binom{p-k}{j}.$$

Then we expand $H(\mathcal{E}, \mathbf{\Pi}^* | \widehat{\mathbf{\Pi}}, \boldsymbol{y}, \mathbf{X})$ as

$$\begin{aligned}
& \mathsf{H}(\mathcal{E}, \mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X}) \\
= \; & \mathsf{H}(\mathcal{E}|\widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X}) + \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})|\mathcal{E}, \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X}) \\
\overset{②}{\leq} \; & \log 2 + \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}) \mid \mathcal{E}, \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}}), \boldsymbol{y}, \mathbf{X}) \\
\overset{③}{\leq} \; & \log 2 + \; \mathbb{P}(\mathcal{E}=1) \, \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})|\mathcal{E}=1, \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}})) + \mathbb{P}(\mathcal{E}=0) \, \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})|\mathcal{E}=0, \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}})) \\
\overset{④}{\leq} \; & \log 2 + [1 - \mathbb{P}(\mathcal{E}=0)] \, \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) + \mathbb{P}(\mathcal{E}=0) \log \left[\sum_{i=1}^{\mathsf{D}} \sum_{j=1}^{(\mathsf{D}-i)\wedge k} \frac{n!}{(n-i)!} \binom{k}{j} \binom{p-k}{j} \right] \\
\overset{⑤}{=} \; & \log 2 + \mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) - \mathbb{P}(\mathcal{E}=0) \log \zeta,
\end{aligned}$$

where in ② we use the fact that $\mathcal{E}$ is binary and hence $\mathsf{H}(\mathcal{E}|\cdot) \leq \log 2$, in ③ we use the property that conditioning reduces entropy (Cover and Thomas, 2012, Equation (2.157)), in ④ we use the property

$$\mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})|\mathcal{E}=0, \widehat{\mathbf{\Pi}}, \mathrm{supp}(\widehat{\boldsymbol{\beta}})) \leq \log \left| \mathbb{B}\left[\left(\mathbf{\Pi}^{\natural}, \boldsymbol{\beta}^{\natural} \right); \mathsf{D} \right] \right|,$$

and in ⑤ we use the fact that $\mathsf{H}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})) = \log \left(n! \cdot \binom{p}{k} \right)$. □

A.3 Supporting lemmas for Section 3

Lemma 1. *Assume that $\boldsymbol{\beta}^{\natural} \in \{0, 1\}^p$. we have*

$$\mathrm{I}\left(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}); \boldsymbol{y}|\mathbf{X} \right) \leq \frac{n}{2} \log\left(1 + \mathsf{SNR}\right),$$

where $\mathrm{I}(\cdot;\cdot)$ denotes the mutual information.

Proof. Denote $\mathsf{h}(\cdot)$ as the differential entropy. We have

$$\begin{aligned}
\mathrm{I}(\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}); \boldsymbol{y} \mid \mathbf{X}) & \overset{①}{=} \mathbb{E}_{\mathbf{X}, \boldsymbol{w}, \mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural})} \left[\mathsf{h}\left(\boldsymbol{y}|\mathbf{X}=\boldsymbol{x} \right) - \mathsf{h}\left(\boldsymbol{y}|\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}), \mathbf{X}=\boldsymbol{x} \right) \right] \\
& \overset{②}{\leq} \mathbb{E}_{\mathbf{X}} \frac{1}{2} \log\det\left(\mathbb{E}_{\boldsymbol{w}, \mathbf{\Pi}^{\natural} \mid \mathbf{X}=\boldsymbol{x}} \boldsymbol{y}\boldsymbol{y}^{\top} \right) - \frac{n}{2} \log \sigma^2 \\
& \overset{③}{\leq} \frac{1}{2} \log\det\left[\mathbb{E}_{\mathbf{X}, \mathbf{\Pi}^{\natural}} (\sigma^2 \mathbf{I}_{n\times n} + \mathbf{\Pi}^{\natural}\mathbf{X}\boldsymbol{\beta}^{\natural} \; \boldsymbol{\beta}^{\natural\top}\mathbf{X}^{\top}\mathbf{\Pi}^{\natural\top}) \right] - \frac{n}{2} \log \sigma^2 \\
& \overset{④}{=} \frac{n}{2} \log\left(\sigma^2 + \|\boldsymbol{\beta}^{\natural}\|_2^2 \right) - \frac{n}{2} \log \sigma^2 = \frac{n}{2} \log(1 + \mathsf{SNR}),
\end{aligned}$$

where ① is because of the definition of conditional mutual information; ② is due to the property (Cover and Thomas, 2012, Theorem 8.6.5)

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

for a random variable $\mathbf{Z}$ with finite covariance matrix $\mathrm{Cov}(\mathbf{Z})$, and $\mathsf{h}(\boldsymbol{y}|\mathbf{\Pi}^{\natural}, \mathrm{supp}(\boldsymbol{\beta}^{\natural}), \mathbf{X}=\boldsymbol{x}) = \mathsf{h}(\boldsymbol{w})$ as $\boldsymbol{\beta}^{\natural}$'s information is fully encoded in $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$; in ③ we use the concavity of $\log\det(\cdot)$, i.e., $\mathbb{E}\log\det(\cdot) \leq \log\det\mathbb{E}(\cdot)$; and in ④ we have

$$\mathbb{E}_{\mathbf{X}, \mathbf{\Pi}^{\natural}} \left(\mathbf{\Pi}^{\natural}\mathbf{X}\boldsymbol{\beta}^{\natural} \; \boldsymbol{\beta}^{\natural\top}\mathbf{X}^{\top}\mathbf{\Pi}^{\natural\top} \right) = \|\boldsymbol{\beta}^{\natural}\|_2^2 \cdot \mathbf{I}_{n\times n}.$$

□

B Analysis of ML estimator

This section analyzes the ML estimator.

B.1 Notation definition

We denote $\mathrm{supp}(\boldsymbol{\beta}^{\natural})$ and $\mathrm{supp}(\boldsymbol{\beta})$ as T and S, respectively. In addition, we define $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ as $\mathcal{T}_1 \triangleq T \setminus S$, $\mathcal{T}_2 \triangleq T \bigcap S$, and $\mathcal{T}_3 \triangleq S \setminus T$, respectively. An illustration is available in Figure 5. In

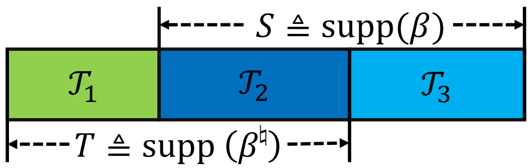

Figure 5: Illustration of $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$: $\mathcal{T}_1 \triangleq T \setminus S$, $\mathcal{T}_2 \triangleq T \bigcap S$, and $\mathcal{T}_3 \triangleq S \setminus T$, respectively.

addition, we define the following events before we proceed.

$$
\begin{aligned}
\mathcal{E}_1 &\triangleq \left\{\boldsymbol{\Pi}^{\natural} = \widehat{\boldsymbol{\Pi}},\ \boldsymbol{\beta}^{\natural} \neq \widehat{\boldsymbol{\beta}}\right\};\\
\mathcal{E}_2 &\triangleq \left\{\boldsymbol{\Pi}^{\natural} \neq \widehat{\boldsymbol{\Pi}},\ \boldsymbol{\beta}^{\natural} = \widehat{\boldsymbol{\beta}}\right\};\\
\mathcal{E}_3 &\triangleq \left\{\boldsymbol{\Pi}^{\natural} \neq \widehat{\boldsymbol{\Pi}},\ \boldsymbol{\beta}^{\natural} \neq \widehat{\boldsymbol{\beta}}\right\};\\
\mathcal{E}_4(\delta;\boldsymbol{\Pi}, S) &\triangleq \left\{\left\|\mathsf{P}^{\perp}_{\boldsymbol{\Pi}\mathbf{X}_S}\boldsymbol{y}\right\|_2^2 - \left\|\mathsf{P}^{\perp}_{\boldsymbol{\Pi}\mathbf{X}_S}\boldsymbol{w}\right\|_2^2 < 2\delta\right\};\\
\mathcal{E}_5(\delta;\boldsymbol{\Pi}, S) &\triangleq \left\{\left|\left\|\mathsf{P}^{\perp}_{\boldsymbol{\Pi}^{\natural}\mathbf{X}_T}\boldsymbol{w}\right\|_2^2 - \left\|\mathsf{P}^{\perp}_{\boldsymbol{\Pi}\mathbf{X}_S}\boldsymbol{w}\right\|_2^2\right| \geq \delta\right\};\\
\mathcal{E}_6(t;h) &\triangleq \left\{\left\|\mathsf{P}^{\perp}_{\boldsymbol{\Pi}\mathbf{X}_S}\boldsymbol{\Pi}^{\natural}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T\right\|_2^2 \geq t\|\boldsymbol{\beta}^{\natural}_T\|_2^2,\ \forall\ S, \boldsymbol{\Pi} \text{ s.t. } \mathsf{d_H}(\mathbf{I},\boldsymbol{\Pi}) = h\right\};\\
\mathcal{E}_7(\delta;\boldsymbol{\Pi}, S) &\triangleq \left\{\left|\left\|\mathsf{P}^{\perp}_{\boldsymbol{\Pi}^{\natural}\mathbf{X}_T}\boldsymbol{w}\right\|_2^2 - \mathbb{E}\left\|\mathsf{P}^{\perp}_{\boldsymbol{\Pi}^{\natural}\mathbf{X}_T}\boldsymbol{w}\right\|_2^2\right| \geq \frac{\delta}{2}\right\};\\
\mathcal{E}_8(\delta;\boldsymbol{\Pi}, S) &\triangleq \left\{\left|\left\|\mathsf{P}^{\perp}_{\boldsymbol{\Pi}\mathbf{X}_S}\boldsymbol{w}\right\|_2^2 - \mathbb{E}\left\|\mathsf{P}^{\perp}_{\boldsymbol{\Pi}\mathbf{X}_S}\boldsymbol{w}\right\|_2^2\right| \geq \frac{\delta}{2}\right\}.
\end{aligned}
$$

B.2 Proof of Theorem 3

We prove a more specific version of Theorem 3, which is

Theorem. *Consider the noiseless case, i.e.,* $\boldsymbol{w} = \mathbf{0}$. *Suppose* $n = \Omega(k\log p)$, *we have* $\mathbb{P}((\widehat{\boldsymbol{\Pi}}_{\mathsf{ML}}, \widehat{\boldsymbol{\beta}}_{\mathsf{ML}}) \neq (\boldsymbol{\Pi}^{\natural},\ \boldsymbol{\beta}^{\natural})) \lesssim n^{-2}$.

Proof. We upper bound the error probability $\mathbb{P}(\widehat{\boldsymbol{\Pi}}_{\mathsf{ML}} \neq \boldsymbol{\Pi}^{\natural})$ can be decomposed as $\sum_{\ell=1}^{3}\mathbb{E}\mathbb{1}(\mathcal{E}_{\ell})$, whose definitions are stated in Subsection B.1.

Bounding $\mathbb{E}\mathbb{1}(\mathcal{E}_1)$. Conditional on $\mathcal{E}_1$, easily we can prove that $\mathcal{T}_1$ is not empty, since otherwise we have $\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T = \mathbf{X}_T\widehat{\boldsymbol{\beta}}$, which leads to contradiction. Given the support set S, we can write $\boldsymbol{\beta}_S$ as $(\mathbf{X}_S^{\top}\mathbf{X}_S)^{-1}\mathbf{X}_S^{\top}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T$. Then we have

$$
\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T = \mathbf{X}_S\boldsymbol{\beta}_S = \mathsf{P}_{\mathbf{X}_S}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T,
$$

which implies $\mathbf{X}_T\boldsymbol{\beta}^\natural_T$ lies within the linear space spanned by the columns of $\mathbf{X}_S$. Then we obtain $\mathsf{P}^\perp_{\mathbf{X}_S}(\mathbf{X}_T\boldsymbol{\beta}^\natural_T) = \mathsf{P}^\perp_{\mathbf{X}_S}(\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^\natural_{\mathcal{T}_1}) = 0$ and can upper-bound $\mathbb{E}\mathbb{1}(\mathcal{E}_1)$ as $\binom{p}{k}\cdot\mathbb{P}(\|\mathsf{P}^\perp_{\mathbf{X}_S}(\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^\natural_{\mathcal{T}_1})\|_2 = 0)$. Recalling the definition of $\mathcal{T}_1$, we conclude $\mathbf{X}_{\mathcal{T}_1}$ to be independent of $\mathbf{X}_S$. Then, with the rotational invariance of Gaussian distribution, we can rewrite $\mathbb{P}(\|\mathsf{P}^\perp_{\mathbf{X}_S}(\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^\natural_{\mathcal{T}_1})\|_2 = 0)$ as $\mathbb{P}(\left\|\mathsf{P}^\perp_{\mathbf{X}_S}\boldsymbol{z}\right\|_2\|\boldsymbol{\beta}^\natural_{\mathcal{T}_1}\|_2 = 0) = \mathbb{P}(\left\|\mathsf{P}^\perp_{\mathbf{X}_S}\boldsymbol{z}\right\|_2 = 0)$, where $\boldsymbol{z}\in\mathbb{R}^n$ is a Gaussian random variable satisfying $\boldsymbol{z}\sim\mathsf{N}(0,\mathbf{I}_{n\times n})$.

In the following proof, we will view $\boldsymbol{z}$ as a fixed vector as it is independent of $\mathbf{X}_S$. Afterwards, we can upper-bound $\mathbb{E}\mathbb{1}(\mathcal{E}_1)$ as

$$\mathbb{E}\mathbb{1}(\mathcal{E}_1) \leq \binom{p}{k}\mathbb{P}\left(\left\|\mathsf{P}^\perp_{\mathbf{X}_S}\boldsymbol{z}\right\|_2 \leq \gamma\|\boldsymbol{z}\|_2\right) \overset{①}{\leq} \left(\frac{ep}{k}\right)^k \exp\left[\frac{n-k}{n}\left(1-\gamma^2+\log\gamma^2\right)\right] \overset{②}{\leq} n^{-2},$$

where in ① we use $\binom{p}{k}\leq (ep/k)^k$ and (Dasgupta and Gupta, 2003, Lemma 2.2(a)), and in ② we pick γ as $\frac{k^k}{\sqrt{en^2(ep)^k}}$.

Bounding $\mathbb{E}\mathbb{1}\left(\mathcal{E}_2\right)$. Event $\mathcal{E}_2$ suggests that there exists another permutation matrix $\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^\natural$ such that $\widehat{\mathbf{\Pi}}\mathbf{X}_T\boldsymbol{\beta}^\natural_T = \mathbf{\Pi}^\natural\mathbf{X}_T\boldsymbol{\beta}^\natural_T$. Then, we can equate $\mathbb{E}\mathbb{1}(\mathcal{E}_2)$ with $\mathbb{P}(\|(\widehat{\mathbf{\Pi}}-\mathbf{\Pi}^\natural)\boldsymbol{z}\|_2 = 0,\ \exists\ \widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^\natural)$, where $\boldsymbol{z}\sim\mathsf{N}(\mathbf{0},\mathbf{I}_{n\times n})$. This leads to

$$\mathbb{E}\mathbb{1}(\mathcal{E}_2) \leq \mathbb{P}(\|(\widehat{\mathbf{\Pi}}-\mathbf{\Pi}^\natural)\boldsymbol{z}\|_2 = 0,\ \exists\ \widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^\natural) \leq \sum_{h\geq 2}\binom{n}{h}h!\cdot\mathbb{P}(\|(\mathbf{I}-\mathbf{\Pi})\boldsymbol{z}\|_2 = 0,\ \text{s.t. } \mathsf{d_H}(\mathbf{I},\mathbf{\Pi}) = h).$$

With Lemma 5, we have

$$\begin{aligned}
\mathbb{E}\mathbb{1}(\mathcal{E}_2) &\leq \sum_{h\geq 2}\binom{n}{h}h!\mathbb{P}(\|(\mathbf{I}-\mathbf{\Pi})\boldsymbol{z}\|_2 \leq \frac{4}{en^{20}},\ \text{s.t. } \mathsf{d_H}(\mathbf{I},\mathbf{\Pi}) = h) \\
&\overset{③}{\leq} 6\sum_{h\geq 2} n^h\cdot\exp\left[\frac{h}{10}\left(\log\left(2/(ehn^{20})\right) - 2/(ehn^{20}) + 1\right)\right] \overset{④}{\leq} 6\sum_{h\geq 2} n^{-h} \leq \frac{6}{n(n-1)},
\end{aligned}$$

where ③ is because $n!/(n-h)!\leq n^h$, and ④ is due to $\exp\left[\frac{h}{10}\left(\log\left(2/(ehn^{20})\right) - 2/(ehn^{20}) + 1\right)\right] \leq n^{-2h}$.

Bounding $\mathbb{E}\mathbb{1}(\mathcal{E}_3)$. Adopting the similar argument as in bounding $\mathbb{E}\mathbb{1}(\mathcal{E}_1)$, for a fixed permutation matrix $\mathbf{\Pi}$ and support set S, we have

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_S = \left(\mathbf{\Pi}\mathbf{X}_S\right)^\dagger\ \mathbf{\Pi}^\natural\mathbf{X}_T\boldsymbol{\beta}^\natural_T.$$

Based on the optimality of the objective function in (5), we conclude that

$$\mathbf{\Pi}^\natural\mathbf{X}_T\boldsymbol{\beta}^\natural_T = \mathbf{\Pi}\mathbf{X}_S\widehat{\boldsymbol{\beta}} = \mathsf{P}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^\natural\mathbf{X}_T\boldsymbol{\beta}^\natural_T,$$

which suggests $\mathbf{\Pi}^\natural\mathbf{X}_T\boldsymbol{\beta}^\natural_T$ lies within the linear space spanned by the columns in $\mathbf{\Pi}\mathbf{X}_S$. Following a similar procedure as in bounding $\mathbb{E}\mathbb{1}(\mathcal{E}_1)$, we have $\mathbb{E}\mathbb{1}(\mathcal{E}_3)\leq\frac{8}{n(n-1)}$. Combining the discussions thereof then completes the proof.

□

B.3 Proof of Theorem 4

Proof. The proof consists of two stages

- **Stage I.** The permutation matrix can be obtained with high probability;
- **Stage II.** Given the correct permutation, we can detect the support set with high probability.

The proof of Stage I is in Lemma 2 while the proof of Stage II is in Lemma 3.

$\square$

B.4 Supporting lemmas for Section 4

Lemma 2. *Assume $n = \Omega\left(k \log\left(\frac{ep}{k}\right)\right)$, and*

$$\log(\mathsf{SNR}) \geq \ [128 \log n + {}^{64k}/_{n} \cdot \log\left({}^{ep}/_{k}\right)] \vee \left[34 \log n + 2\log(4e^6)\right] \vee \left[148 \log n + 4\log(2e^6)\right], \quad (12)$$

we conclude that $\mathbb{P}(\widehat{\mathbf{\Pi}}_{\mathrm{ML}} \neq \mathbf{\Pi}^{\natural}) \leq \frac{13}{n(n-1)}$.

Proof. Fixing the support set S and permutation matrix $\mathbf{\Pi}$, we have $\min_{\mathrm{supp}(\widehat{\beta})=S} \|\boldsymbol{y} - \mathbf{\Pi}\mathbf{X}\boldsymbol{\beta}\|_2$ be expressed as $\left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\boldsymbol{y}\right\|_2$. Here we only study the permutation reconstruction error (i.e., $\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^{\natural}$) and define error event $\mathcal{E}$ as

$$\mathcal{E} \triangleq \left\{\exists\ (\mathbf{\Pi}, S)\ \text{ s.t. } \mathbf{\Pi} \neq \mathbf{\Pi}^{\natural},\ \left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\boldsymbol{y}\right\|_2 \leq\ \left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}^{\natural}\mathbf{X}_T}\boldsymbol{y}\right\|_2\right\}.$$

Then we would like to show $\mathcal{E}$ holds with probability close to zero.

First, we would like to show $\mathcal{E} \subseteq \cup_{\mathbf{\Pi},S}\mathcal{E}_4(\delta;\mathbf{\Pi},S) \cup \mathcal{E}_5(\delta;\mathbf{\Pi},S)$. This is because

$$\begin{aligned}
&\bigcap_{\mathbf{\Pi},S}\left[\overline{\mathcal{E}}_4(\delta;\mathbf{\Pi},S)\bigcap\overline{\mathcal{E}}_5(\delta;\mathbf{\Pi},S)\right] \\
=&\left\{\forall\ (\mathbf{\Pi},S) \text{ s.t. } \mathbf{\Pi}\neq\mathbf{\Pi}^{\natural},\ \left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\boldsymbol{y}\right\|_2^2 - \left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\boldsymbol{w}\right\|_2^2 \geq 2\delta, \left|\left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}^{\natural}\mathbf{X}_T}\boldsymbol{w}\right\|_2^2 - \left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\boldsymbol{w}\right\|_2^2\right| < \delta\right\} \\
\subseteq&\left\{\forall\ (\mathbf{\Pi},S) \text{ s.t.} \mathbf{\Pi}\neq\mathbf{\Pi}^{\natural},\ \left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\boldsymbol{y}\right\|_2^2 - \left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}^{\natural}\mathbf{X}_T}\boldsymbol{w}\right\|_2^2 \geq \delta\right\} \subseteq \overline{\mathcal{E}}.
\end{aligned}$$

Then we can bound $\mathbb{E}\mathbb{1}(\mathcal{E})$ as

$$\mathbb{E}\mathbb{1}(\mathcal{E}) \leq \sum_{h\geq 2}\binom{n}{h}h!\left[\sum_{S}(\zeta_1+\zeta_2)+\zeta_3\right], \quad (13)$$

where ζ_1, ζ_2, and ζ_3 are defined as

$$\begin{aligned}
\zeta_1 &\triangleq\ \ \mathbb{E}\mathbb{1}\left(\mathcal{E}_4\left(\delta;\mathbf{\Pi},S\right)\bigcap\mathcal{E}_6(t_h;h)\ \middle|\ \mathsf{d}_{\mathsf{H}}(\mathbf{I},\mathbf{\Pi}) = h\right); \\
\zeta_2 &\triangleq\ \ \mathbb{E}\mathbb{1}\left(\mathcal{E}_5\left(\delta;\mathbf{\Pi},S\right)\bigcap\mathcal{E}_6(t_h;h)\ \middle|\ \mathsf{d}_{\mathsf{H}}(\mathbf{I},\mathbf{\Pi}) = h\right); \\
\zeta_3 &\triangleq\ \ \mathbb{E}\mathbb{1}\left(\overline{\mathcal{E}}_6(t_h;h)\right).
\end{aligned}$$

The following analysis can be roughly divided into 2 steps

- **Step I.** Setting $\delta = \left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^{\natural}\mathbf{X}_T\beta^{\natural}_T\right\|_2^2/4$, we separately bound ζ_1, ζ_2, and ζ_3.
- **Step II.** We pick t_h as $(n\cdot\log\mathsf{SNR})/\mathsf{SNR}$ and show $\mathbb{P}(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^{\natural}) \leq 13/n(n-1)$ provided assumption (12) holds.

Step I. Picking δ as $\left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^{\natural}\mathbf{X}_T\beta^{\natural}_T\right\|_2^2/4$, we have

$$\begin{aligned}
\zeta_1 =&\ \mathbb{E}_{\mathbf{X},\boldsymbol{w}}\mathbb{1}\left[\langle \mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^{\natural}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T, \boldsymbol{w}\rangle \leq -\left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^{\natural}\mathbf{X}_T\beta^{\natural}_T\right\|_2^2/4 \cap \mathcal{E}_6(t_h;h)\right] \\
\overset{①}{\leq}&\ \mathbb{E}_{\mathbf{X}}\mathbb{1}\left[\Phi\left(-\left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^{\natural}\mathbf{X}_T\beta^{\natural}_T\right\|_2^2/4\sigma\left\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^{\natural}\mathbf{X}_T\beta^{\natural}_T\right\|_2\right)\bigcap\mathcal{E}_6(t_h;h)\right] \\
\overset{②}{\leq}&\ \mathbb{E}\Phi\left(-\sqrt{t_h}\|\beta^{\natural}_T\|_2/4\sigma\right) \overset{③}{\leq} \exp\left(-t_h\|\beta^{\natural}_T\|_2^2/32\sigma^2\right) = \exp\left(-t_h\cdot\mathsf{SNR}/32\right),
\end{aligned}$$

where in ① we condition on $\mathbf{X}$ and have $2\langle \mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^{\natural}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T, \boldsymbol{w}\rangle$ to be a Gaussian random variable with zero mean and $4\sigma^2\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^{\natural}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T\|_2^2$ variance, i.e., $\mathsf{N}(0, 4\sigma^2\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\mathbf{\Pi}^{\natural}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T\|_2^2)$, $\Phi(\cdot)$ denotes the CDF for the standard normal distribution, ② is because of event $\mathcal{E}_6(t_h;h)$, and ③ is due to the tail bound $\Phi(-x)\leq e^{-x^2/2}$, $x\geq 0$ (c.f. Proposition 2.12 in Vershynin (2018)).

For ζ_2, we notice the relation such that $\mathbb{E}_{\boldsymbol{w},\mathbf{X}}\|\mathsf{P}^{\perp}_{\mathbf{\Pi}^{\natural}\mathbf{X}_T}\boldsymbol{w}\|_2^2 = \mathbb{E}_{\boldsymbol{w},\mathbf{X}}\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\boldsymbol{w}\|_2^2 = (n-k)\sigma^2$ and perform the decomposition

$$\mathcal{E}_5(\delta;\mathbf{\Pi},S)\subseteq\mathcal{E}_7(\delta;\mathbf{\Pi},S)\bigcup\mathcal{E}_8(\delta;\mathbf{\Pi},S).$$

This leads to

$$\zeta_2 \leq \mathbb{E}\mathbb{1}(\mathcal{E}_7(\delta;\mathbf{\Pi},S)\bigcap\mathcal{E}_6(t_h;h)) + \mathbb{E}\mathbb{1}(\mathcal{E}_8(\delta;\mathbf{\Pi},S)\bigcap\mathcal{E}_6(t_h;h)).$$

Exploiting the independence between $\mathbf{X}$ and $\boldsymbol{w}$, we can condition on $\mathbf{X}$ view $\|\mathsf{P}^{\perp}_{\mathbf{\Pi}^{\natural}\mathbf{X}_T}\boldsymbol{w}\|_2^2/\sigma^2$ ($\|\mathsf{P}^{\perp}_{\mathbf{\Pi}\mathbf{X}_S}\boldsymbol{w}\|_2^2/\sigma^2$ resp.) as χ^2 random variable with freedom $n-k$. Plugging δ into the tail bound for χ^2 as in Wainwright (2019, Example 2.11, P. 29), we conclude

$$\zeta_2 \leq 4\exp\left[-\left(\frac{t_h\cdot\mathsf{SNR}}{64}\wedge\frac{t_h^2\cdot\mathsf{SNR}^2}{512(n-k)}\right)\right].$$

Ultimately, we invoke Lemma 4 and bound ζ_3 as

$$\zeta_3 \leq \quad 2n^{-2h} + 6\exp\left[\frac{h}{10}\left(\log\left(\frac{4e^6n^{16h/n}t_h}{h}\right) - \frac{4e^6n^{16h/n}t_h}{h} + 1\right)\right],$$

where $t_h < h/(4e^6n^{16h/n})$.

Step II. Let t_h be $(n\cdot\log\mathsf{SNR})/\mathsf{SNR}$, we will show ζ_ℓ $(1\leq\ell\leq 3)$ all shrink to zero with the assumption (12). For ζ_1, we use the assumption $\log(\mathsf{SNR}) \geq 128\log n + 64k\log(\frac{ep}{k})/n \geq 128h\log n/n + 64k\log(\frac{ep}{k})/n$ in (12) and have $\zeta_1 \leq \exp\left(-t_h\cdot\mathsf{SNR}/32\right) \leq n^{-2h}\left(k/ep\right)^k$.

Then we turn to ζ_2. If $\frac{t_h\cdot\mathsf{SNR}}{64}\wedge\frac{t_h^2\cdot\mathsf{SNR}^2}{512(n-k)} = \frac{t_h\cdot\mathsf{SNR}}{64}$, we can apply the same strategy to bound ζ_2. Otherwise, we need to bound ζ_2 as

$$\exp\left(-\frac{t_h^2\cdot\mathsf{SNR}^2}{512(n-k)}\right) = \exp\left(-\frac{n^2\cdot\log^2\mathsf{SNR}}{512(n-k)}\right) \overset{⑤}{\leq} \exp\left(-\frac{n\cdot\log\mathsf{SNR}}{64}\right) \leq n^{-2h}\left(\frac{k}{ep}\right)^k, \tag{14}$$

where in ⑤ we use the relation $\log \mathsf{SNR} \geq 8$ and redo the above analysis. For ζ_3, we first need to check the condition $t_h < h/(4e^6 n^{16h/n})$ is not violated as

$$\log t_h < -16\log n - \log(4e^6) < \log h - {}^{16h\log n}/_{n} - \log(4e^6).$$

Then we consider $\exp[\frac{h}{10}(\log {}^{(4e^6 n^{16h/n} t_h)}/_{h} - {}^{(4e^6 n^{16h/n} t_h)}/_{h} + 1)]$, which reads as

$$\begin{aligned}
& \exp\left[\frac{h}{10}\left(\log\left(\frac{4e^6 n^{16h/n} t_h}{h}\right) - \frac{4e^6 n^{16h/n} t_h}{h} + 1\right)\right] \\
\leq\ & \exp\left[\frac{h}{10}\log\left(\frac{4e^6 n^{16h/n+1}}{h}\right)\right] \times \exp\left[\frac{h}{10}\left(\log\frac{\log \mathsf{SNR}}{\mathsf{SNR}} - \frac{\log \mathsf{SNR}}{\mathsf{SNR}} + 1\right)\right] \\
\overset{⑥}{\leq}\ & \exp\left[\frac{h}{10}\log\left(\frac{4e^6 n^{16h/n+1}}{h}\right)\right] \times \exp\left(-\frac{h\cdot\log \mathsf{SNR}}{40}\right) \overset{⑦}{\leq} n^{-2h},
\end{aligned} \tag{15}$$

where in ⑥ we use $\frac{\log z}{z} - 1 - \log\frac{\log z}{z} \geq \frac{\log z}{4}$ when $z \geq 1.25$, and in ⑦ we use the relation $\log \mathsf{SNR} \geq 148\log n + 4\log(2e^6)$ in (12).

Combining (13), (14), and (15), we complete the proof as

$$\mathbb{E}\mathbb{1}(\mathcal{E}) \leq \sum_{h\geq 2}\binom{n}{h} h! \cdot \left[\sum_{S}\left(5n^{-2h}\,({}^{k}/_{ep})^k\right) + 8n^{-2h}\right] \leq 13\sum_{h\geq 2} n^{-h} \leq \frac{13}{n(n-1)}.$$

□

Lemma 3. *Consider the case where the ground-truth permutation matrix $\mathbf{\Pi}^\natural$ is given a prior. Provided that (i) $n \gtrsim k\log p$, and (ii) $\min_{i\in T} {}^{|\beta_i^\natural|^2}/_{\sigma^2} \gtrsim 1$, we have $\mathbb{P}(S \neq T) \leq c_0 e^{-c_1 k\log p}$, where $c_0, c_1 > 0$ are some positive constants.*

Proof. We assume $\mathbf{\Pi}^\natural = \mathbf{I}$ w.l.o.g. Recalling the definition of our estimator in (5), we ought to have

$$\|\mathsf{P}_{\mathbf{X}_S}^{\perp}\boldsymbol{y}\|_{\mathrm{F}}^2 \leq \|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{y}\|_{\mathrm{F}}^2 = \|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2,$$

which is equivalent to

$$\|\mathsf{P}_{\mathbf{X}_S}^{\perp}\mathbf{X}_T\boldsymbol{\beta}_T^\natural\|_{\mathrm{F}}^2 + 2\left\langle \mathsf{P}_{\mathbf{X}_S}^{\perp}\mathbf{X}_T\boldsymbol{\beta}_T^\natural, \boldsymbol{w}\right\rangle \leq \|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2 - \|\mathsf{P}_{\mathbf{X}_S}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2.$$

To begin with, we consider a fixed index set S and have

$$\mathbb{P}\left(\|\mathsf{P}_{\mathbf{X}_S}^{\perp}\boldsymbol{y}\|_{\mathrm{F}} \leq \|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{y}\|_{\mathrm{F}}\right) \leq \zeta_1 + \zeta_2 + \zeta_3, \tag{16}$$

where ζ_1, ζ_2, and ζ_3 are defined as

$$\begin{aligned}
\zeta_1 &\triangleq \mathbb{P}(\|\mathsf{P}_{\mathbf{X}_S}^{\perp}\mathbf{X}_{T_1}\boldsymbol{\beta}_{T_1}^\natural\|_2^2 \leq {}^{(n-k)}/_{4}\|\boldsymbol{\beta}_{T_1}^\natural\|_2^2); \\
\zeta_2 &\triangleq \mathbb{P}(2\langle \mathsf{P}_{\mathbf{X}_S}^{\perp}\mathbf{X}_T\boldsymbol{\beta}_T^\natural, \boldsymbol{w}\rangle \lesssim -c\sigma\sqrt{(k\log p)(n-k)}\|\boldsymbol{\beta}_{T_1}^\natural\|_2); \\
\zeta_3 &\triangleq \mathbb{P}(\|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2 - \|\mathsf{P}_{\mathbf{X}_S}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2 \geq {}^{(n-k)}/_{4}\|\boldsymbol{\beta}_{T_1}^\natural\|_2^2 - c\sigma\sqrt{(k\log p)(n-k)}\|\boldsymbol{\beta}_{T_1}^\natural\|_2).
\end{aligned}$$

Analysis of ζ_1. Recalling the fact such that $\mathsf{P}_{\mathbf{X}_S}^{\perp}$ is the projection onto the orthogonal complement of column space spanned by $\mathbf{X}_T$, we can verify $\mathsf{P}_{\mathbf{X}_S}^{\perp}\mathbf{X}_T\boldsymbol{\beta}_T^\natural = \mathsf{P}_{\mathbf{X}_S}^{\perp}\mathbf{X}_{T_1}\boldsymbol{\beta}_{T_1}^\natural$. Then, we decompose ζ_1 as

$$\zeta_1 \leq \zeta_{1,1} + \zeta_{1,2}, \tag{17}$$

where $\zeta_{1,1}$ and $\zeta_{1,2}$ are defined as

$$\begin{aligned}
\zeta_{1,1} &\triangleq \mathbb{P}\Big(\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2 \le \sqrt{n-k}/2\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2, \|\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2 \ge \sqrt{n/2}\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2\Big); \\
\zeta_{1,2} &\triangleq \mathbb{P}\left(\|\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2 \le \sqrt{n/2}\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2\right).
\end{aligned}$$

For $\zeta_{1,1}$, we follow the same procedure as in Theorem 3. Conditional on $\mathbf{X}_{\mathcal{T}_1}$, we view $\mathsf{P}^{\perp}_{\mathbf{X}_S}$ as a random projection from a linear space of dimension n to a linear space of dimension $n-k$, which yields

$$\zeta_{1,1} \le \mathbb{P}\left(\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2 \le (n-k)/2n\|\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2\right) \overset{①}{\le} \exp\left[\frac{n-k}{2}(\log 1/2 - 1/2 + 1)\right] = e^{-c_0 n}, \tag{18}$$

where ① is due to Lemma 11. For $\zeta_{1,2}$, we exploit the fact such that $\|\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2/\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2$ is a χ^2 random variable with freedom n. Then we have

$$\zeta_{1,2} \le \exp\left[\frac{n}{2}(\log 1/2 - 1/2 + 1)\right] = e^{-c_1 n}. \tag{19}$$

Analysis of ζ_2. With the union bound, we have

$$\begin{aligned}
\zeta_2 \le\ & \underbrace{\mathbb{P}\left(2\langle \mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}, \boldsymbol{w}\rangle \lesssim -\sigma\sqrt{k\log p}\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T\|_2\right)}_{\triangleq\ \zeta_{2,1}} \\
&+ \underbrace{\mathbb{P}\Big(\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2 \ge \sqrt{3(n-k)}\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2\Big)}_{\triangleq\ \zeta_{2,2}}.
\end{aligned} \tag{20}$$

For $\zeta_{2,1}$, we exploit the independence between $\mathbf{X}$ and $\boldsymbol{w}$. Conditional on $\mathbf{X}$, we can view $2\langle \mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}, \boldsymbol{w}\rangle$ as a Gaussian random variable with zero mean and $4\sigma^2\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2$ variance, i.e., $\mathsf{N}\left(0, 4\sigma^2\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2\right)$. Then, we have

$$\zeta_{2,1} \le \exp\left(-\frac{c\sigma^2(k\log p)\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2}{4\sigma^2\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2}\right) = e^{-ck\log p}. \tag{21}$$

For $\zeta_{2,2}$, we follow a similar proof as in bounding ζ_1 and have

$$\begin{aligned}
\zeta_{2,2} \le\ & \mathbb{P}(\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2 \ge 3(n-k)\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2, \|\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2 \le \sqrt{3n/2}\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2) \\
&+ \mathbb{P}\left(\|\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2 \ge \sqrt{3n/2}\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2\right) \\
\le\ & \mathbb{P}\left(\|\mathsf{P}^{\perp}_{\mathbf{X}_S}\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2 \ge \frac{2(n-k)}{n}\|\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2\right) + \mathbb{P}\left(\|\mathbf{X}_{\mathcal{T}_1}\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2 \ge \sqrt{3n/2}\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2\right) \le 2e^{-c_1 n}.
\end{aligned} \tag{22}$$

Analysis of ζ_3. Due to the assumptions in Lemma 3, we can verify

$$(n-k)/4\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2^2 - c\sigma\sqrt{(k\log p)(n-k)}\|\boldsymbol{\beta}^{\natural}_{\mathcal{T}_1}\|_2 \ge \min_{i\in T}|\boldsymbol{\beta}^{\natural}_i|\left((n-k)/4\min_{i\in T}|\boldsymbol{\beta}^{\natural}_i| - c\sigma\sqrt{(k\log p)(n-k)}\right)$$

$$\geq c^{'}\sigma^2\sqrt{k\log(ep/k)}\left[\sqrt{k\log(ep/k)} \vee \sqrt{n-k}\right].$$

Thus, we have

$$\zeta_3 \leq \mathbb{P}\left(\left|\|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2 - \|\mathsf{P}_{\mathbf{X}_S}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2\right| \geq \Delta\right),$$

where Δ is set as $c^{'}\sigma^2\sqrt{k\log(ep/k)}\left[\sqrt{k\log(ep/k)} \vee \sqrt{n-k}\right]$. Noticing the relation $\mathbb{E}\|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2 = \mathbb{E}\|\mathsf{P}_{\mathbf{X}_S}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2$, we obtain

$$\mathbb{P}\left(\left|\|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2 - \|\mathsf{P}_{\mathbf{X}_S}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2\right| \geq \Delta\right) \leq 2\mathbb{P}\left(\left|\|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2 - \mathbb{E}\|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2\right| \geq \Delta/2\right).$$

Due to the independence between $\mathbf{X}$ and $\boldsymbol{w}$, we can view $\|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2/\sigma^2$ as a χ^2 random variable with freedom $n-k$. Thus, we conclude

$$\zeta_3 \leq 2\mathbb{P}\left(\left|\|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2 - \mathbb{E}\|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{w}\|_{\mathrm{F}}^2\right| \geq \Delta/2\right) \leq 4\exp\left(-\left(\frac{\Delta}{16\sigma^2} \wedge \frac{\Delta^2}{32\sigma^4(n-k)}\right)\right) \leq 4e^{-ck\log p}. \tag{23}$$

Combining (16), (17), (18), (19), (20), (21), (22), and (23) yields

$$\mathbb{P}\left(\|\mathsf{P}_{\mathbf{X}_S}^{\perp}\boldsymbol{y}\|_{\mathrm{F}} \leq \|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{y}\|_{\mathrm{F}}\right) \leq c_0 e^{-c_1 k\log p}, \tag{24}$$

where the assumption $n \gtrsim k\log p$ is invoked. Notice that (24) is w.r.t. a fixed index set S. In the end, we iterate over all possible index set $S \neq T$ and complete the proof with the union bound, which reads as

$$\mathbb{P}(S \neq T) \leq \sum_{S\neq T}\mathbb{P}\left(\|\mathsf{P}_{\mathbf{X}_S}^{\perp}\boldsymbol{y}\|_{\mathrm{F}} \leq \|\mathsf{P}_{\mathbf{X}_T}^{\perp}\boldsymbol{y}\|_{\mathrm{F}}\right) \lesssim \binom{p}{k}\cdot e^{-c_1 k\log p} \leq (ep/k)^k e^{-c_1 k\log p} = e^{-c_2 k\log p}.$$

□

Lemma 4. *We have*

$$\begin{aligned}&\mathbb{P}\left(\|\mathsf{P}_{\mathbf{\Pi}\mathbf{X}_S}^{\perp}\mathbf{\Pi}^{\natural}\mathbf{X}_T\boldsymbol{\beta}_T^{\natural}\|_2^2 < t\|\boldsymbol{\beta}_T^{\natural}\|_2^2,\ \ \exists\ \mathbf{\Pi}, S\ \ \text{s.t.}\ \ \mathsf{d_H}(\mathbf{\Pi},\mathbf{\Pi}^{\natural}) = h\right)\\ &\leq\ 2n^{-2h} +\ 6\exp\left(\frac{h}{10}\left(\log\left(\frac{4e^6 n^{16h/n}t}{h}\right) - \frac{4e^6 n^{16h/n}t}{h} + 1\right)\right),\end{aligned}$$

where $t < h/(4e^6 n^{16h/n})$, and $h \geq 2$.

Proof. We assume $\mathbf{\Pi}^{\natural} = \mathbf{I}$ w.l.o.g. With some simple algebraic manipulations, we have

$$\|\mathsf{P}_{\mathbf{X}_{S\bigcup\mathcal{T}_1}}^{\perp}\mathbf{\Pi}^{\top}\mathbf{X}_T\boldsymbol{\beta}_T^{\natural}\|_2^2 =\ \|\mathsf{P}_{\mathbf{\Pi}\mathbf{X}_{S\bigcup\mathcal{T}_1}}^{\perp}\mathbf{X}_T\boldsymbol{\beta}_T^{\natural}\|_2^2 \leq \|\mathsf{P}_{\mathbf{\Pi}\mathbf{X}_S}^{\perp}\mathbf{X}_T\boldsymbol{\beta}_T^{\natural}\|_2^2.$$

Then, we obtain

$$\begin{aligned}&\mathbb{P}(\|\mathsf{P}_{\mathbf{\Pi}\mathbf{X}_S}^{\perp}\mathbf{X}_T\boldsymbol{\beta}_T^{\natural}\|_2^2 < t\|\boldsymbol{\beta}_T^{\natural}\|_2^2,\ \ \exists\ \mathbf{\Pi}, S\ \ \text{s.t.}\ \ \mathsf{d_H}(\mathbf{\Pi},\mathbf{\Pi}^{\natural}) = h)\\ &\leq \mathbb{P}(\|\mathsf{P}_{\mathbf{X}_{S\bigcup\mathcal{T}_1}}^{\perp}\mathbf{\Pi}^{\top}\mathbf{X}_T\boldsymbol{\beta}_T^{\natural}\|_2^2 \leq t\|\boldsymbol{\beta}_T^{\natural}\|_2^2,\ \exists\ \mathbf{\Pi}, S\ \text{s.t.}\ \mathsf{d_H}(\mathbf{\Pi},\mathbf{\Pi}^{\natural}) = h) \leq \zeta_1 + \zeta_2,\end{aligned} \tag{25}$$

where ζ_1 and ζ_2 are defined as

$$\zeta_1 \triangleq \mathbb{P}\left(\left\|\mathsf{P}^{\perp}_{\mathbf{X}_{T\bigcup \mathcal{T}_3}} \mathsf{P}^{\perp}_{\mathbf{X}_T}\mathbf{\Pi}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T\right\|_2^2 \leq \gamma_h\left\|\mathsf{P}^{\perp}_{\mathbf{X}_T}\mathbf{\Pi}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T\right\|_2^2, \mathsf{d}_{\mathsf{H}}(\mathbf{I},\mathbf{\Pi}) = h, \exists\, S\right);$$
$$\zeta_2 \triangleq \mathbb{P}\left(\left\|\mathsf{P}^{\perp}_{\mathbf{X}_T}\mathbf{\Pi}\mathbf{X}_T\boldsymbol{\beta}^{\natural}_T\right\|_2^2 \leq \frac{t}{\gamma_h}\|\boldsymbol{\beta}^{\natural}_T\|_2^2,\ \mathsf{d}_{\mathsf{H}}(\mathbf{I},\mathbf{\Pi}) = h, \exists\, S\right).$$

Here we set γ_h as $2^{-1}e^{-5}n^{-8h/n} < 2e^{5^{-1}} < \frac{n-2k}{n-k}$. The following context separately bound ζ_1 and ζ_2.

Analysis of ζ_1. When $\mathcal{T}_3 = \emptyset$, we can verify $\zeta_1 = 0$. Then we turn to the case where $\mathcal{T}_3 \neq \emptyset$. Conditional on $\mathbf{X}_T$, we can view $\mathsf{P}^{\perp}_{\mathbf{X}_{T\bigcup\mathcal{T}_3}}$ as a random projection from a linear space with dimension $n-k$ to a linear space with dimension $n - |T\bigcup\mathcal{T}_3| \geq n-2k$. Invoking Lemma 2.2 in Dasgupta and Gupta (2003) (listed as Lemma 11), we have

$$\zeta_1 \overset{①}{\leq} \binom{p}{k}\exp\left[\frac{n-2k}{2}\left(\log\left(\frac{\gamma_h(n-k)}{n-2k}\right)+1\right)\right] \overset{②}{\leq} \left(\frac{ep}{k}\right)^k \exp\left(\frac{n}{4}\log\left(\frac{n-k}{n-2k}\frac{n^{-8h/n}}{2e^4}\right)\right) \overset{③}{\leq} n^{-2h}, \tag{26}$$

where ① is because of the union bound; ② is due to $\binom{p}{k} \leq (ep/k)^k$ and the definition of γ_h; and ③ is because of the assumption $n \geq k\log(ep/k) \geq 4k$.

Analysis of ζ_2. Due to the independence between S and T, we can safely drop $\exists S$ in ζ_2. The following analysis is a replication of the proof of Lemma 3 in Pananjady et al. (2018) with the only difference in the parameter setting, namely, t/γ_h. We present it only for the sake of self-containing without claiming any novelties.

Without loss of generality, we assume T to be the first k entries. With the union bound, we obtain

$$\begin{aligned}
\zeta_2 \leq\ & \mathbb{P}\left(\|\mathsf{P}^{\perp}_{\mathbf{X}_{1:k}}\mathbf{\Pi}\mathbf{X}_1\|_2^2 \leq t/\gamma_h,\ \mathsf{d}_{\mathsf{H}}(\mathbf{I},\mathbf{\Pi}) = h\right) \\
\leq\ & \underbrace{\mathbb{P}\left(\ \left\|\mathsf{P}^{\perp}_{\mathbf{X}_{1:k}}\mathsf{P}^{\perp}_{\mathbf{X}_1}\mathbf{\Pi}\mathbf{X}_1\right\|_2^2 \leq \vartheta_h\left\|\mathsf{P}^{\perp}_{\mathbf{X}_1}\mathbf{\Pi}\mathbf{X}_1\right\|_2^2,\ \mathsf{d}_{\mathsf{H}}(\mathbf{I},\mathbf{\Pi}) = h\right)}_{\triangleq\ \zeta_{2,1}} \\
& + \underbrace{\mathbb{P}\left(\left\|\mathsf{P}^{\perp}_{\mathbf{X}_1}\mathbf{\Pi}\mathbf{X}_1\right\|_2^2 \leq \frac{t}{\gamma_h\vartheta_h},\ \mathsf{d}_{\mathsf{H}}(\mathbf{I},\mathbf{\Pi}) = h\right)}_{\triangleq\ \zeta_{2,2}},
\end{aligned}$$

where γ_h is a positive constant set as $n^{-8h/n}/(2e)$.

For $\zeta_{2,1}$, we notice the relation $\mathsf{P}^{\perp}_{\mathbf{X}_{1:k}} = \mathsf{P}_{\mathbf{X}_1^{\perp}\bigcap\mathbf{X}_{2:k}^{\perp}}$. Condition on $\mathbf{X}_1$, we can view $\mathsf{P}^{\perp}_{\mathbf{X}_{1:k}}$ as a random projection from a $(n-1)$-dimensional linear space to a $(n-k)$-dimensional linear space, which yields

$$\zeta_{2,1} \leq \exp\left[\frac{n-k}{2}\left(\log\left(\frac{(n-1)\vartheta_h}{n-k}\right) - \frac{(n-1)\vartheta_h}{n-k}+1\right)\right] \leq n^{-2h}, \tag{27}$$

where $\vartheta_h \leq \frac{n-k}{n-1}$ is due to Lemma 2.2 in Dasgupta and Gupta (2003) (also listed as Lemma 11).

For $\zeta_{2,2}$, we first perform decomposition

$$\left\|\mathsf{P}^{\perp}_{\mathbf{X}_1}\mathbf{\Pi}\mathbf{X}_1\right\|_2^2 = \|\mathbf{X}_1\|_2^2 - \frac{\langle\mathbf{X}_1,\mathbf{\Pi}\mathbf{X}_1\rangle^2}{\|\mathbf{X}_1\|_2^2} \overset{④}{\geq} \|\mathbf{X}_1\|_2^2 - |\langle\mathbf{X}_1,\mathbf{\Pi}\mathbf{X}_1\rangle| = \frac{1}{2}\left[\|\mathbf{X}_1-\mathbf{\Pi}\mathbf{X}_1\|_2^2 \wedge \|\mathbf{X}_1+\mathbf{\Pi}\mathbf{X}_1\|_2^2\right],$$

where in ④ we use the Cauchy inequality such that $|\langle \mathbf{X}_1, \mathbf{\Pi}\mathbf{X}_1\rangle| \leq \|\mathbf{X}_1\|_2\|\mathbf{\Pi}\mathbf{X}_1\|_2 = \|\mathbf{X}_1\|_2^2$. When $\|\mathbf{X}_1 - \mathbf{\Pi}\mathbf{X}_1\|_2 \wedge \|\mathbf{X}_1 + \mathbf{\Pi}\mathbf{X}_1\|_2 = \|\mathbf{X}_1 - \mathbf{\Pi}\mathbf{X}_1\|_2$, we can directly invoke Lemma 5 to bound $\zeta_{2,2}$. Following a similar procedure, we can show

$$\zeta_{2,2} \leq 6\exp\left(\frac{h}{10}\left(\log\left(\frac{t}{h\gamma_h\vartheta_h}\right) - \frac{t}{h\gamma_h\vartheta_h} + 1\right)\right), \tag{28}$$

provided that we have $t < h\gamma_h\vartheta_h$. The proof for ζ is hence completed by combining (25), (26), (27), and (28).

□

Lemma 5. *Denote h as the Hamming distance between $\mathbf{I}$ and $\mathbf{\Pi}$, i.e., $h \triangleq \mathsf{d_H}(\mathbf{I}, \mathbf{\Pi})$. Assume $\boldsymbol{x} \in \mathbb{R}^n$ be a random vector satisfying $\boldsymbol{x} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_{n\times n})$, then we have*

$$\mathbb{P}\left(\|(\mathbf{I} - \mathbf{\Pi})\boldsymbol{x}\|_2^2 \leq \vartheta\right) \leq 6\exp\left[\frac{h}{10}\left(\log\left(\frac{\vartheta}{2h}\right) - \frac{\vartheta}{2h} + 1\right)\right],$$

for $\vartheta \leq 2h$.

Proof. Adopting the similar proof tricks as in Pananjady et al. (2018), we separately consider the two cases where $h = 2$ and $h \geq 3$.

(Case I) $h = 2$. We assume the first two rows are switched w.l.o.g. Then we have

$$\mathbb{P}\left(\|(\mathbf{I} - \mathbf{\Pi})\boldsymbol{x}\|_2^2 \leq \vartheta\right) = \mathbb{P}\left[(x_1 - x_2)^2 \leq \vartheta/2\right] \overset{①}{\leq} \exp\left(-1/2\left(\vartheta/4 - \log\left(\vartheta/4\right) - 1\right)\right),$$

where in ① we use the tail bounds for the χ^2 random variable $(x_1 - x_2)^2/2$ with freedom 1.

(Case II) $h \geq 3$. We decompose the non-zero rows of $\left(\mathbf{\Pi} - \mathbf{\Pi}^{\natural}\right)$ into three disjoint sets $\mathcal{I}_\ell$ $(1 \leq \ell \leq 3)$ such that (i) the cardinality of each set $\mathcal{I}_\ell$ is lower bounded by $\lfloor h/3\rfloor$, i.e., $|\mathcal{I}_\ell| = h_\ell \geq \lfloor h/3\rfloor$; and (ii) we have j and $\pi(j)$ reside within different sets for an arbitrary index j, where $\pi(\cdot)$ denotes the permutation map pertaining to $\mathbf{\Pi}$.

Define $Z_\ell = \sum_{j\in\mathcal{I}_\ell} (x_j - x_{\pi(j)})^2/2$, which is a χ^2 random variable with freedom h_ℓ. Then, we can decompose $\|(\mathbf{I} - \mathbf{\Pi})\boldsymbol{x}\|_2^2 = 2(Z_1 + Z_2 + Z_3)$ and obtain

$$\begin{aligned}\mathbb{P}\left(\|(\mathbf{I} - \mathbf{\Pi})\boldsymbol{x}\|_2^2 \leq \vartheta\right) \leq \sum_{\ell=1}^{3}\mathbb{P}\left(Z_\ell \leq h_\ell\vartheta/(2h)\right) &\leq \sum_{\ell=1}^{3}\exp\left(-\frac{h_\ell}{2}\left(\frac{\vartheta}{2h} - \log\left(\frac{\vartheta}{2h}\right) - 1\right)\right)\\ &\overset{②}{\leq} 6\exp\left(-\frac{h}{10}\left(\frac{\vartheta}{2h} - \log\left(\frac{\vartheta}{2h}\right) - 1\right)\right),\end{aligned}$$

where in ② we use the relation $h_i \geq \lfloor h/3\rfloor$. The proof is completed by summarizing the above two cases. □

C Proof of Theorem 5

First, we restate the definition of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Xi}})$, which is written as

$$(\widehat{\boldsymbol{\Xi}}, \widehat{\boldsymbol{\beta}}) = \operatorname{argmin}_{\boldsymbol{\Xi},\boldsymbol{\beta}} \frac{1}{2n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \sqrt{n}\cdot\boldsymbol{\Xi}\|_2^2 + \lambda_{\boldsymbol{\Xi}}\|\boldsymbol{\Xi}\|_1 + \lambda_{\beta}\|\boldsymbol{\beta}\|_1.$$

Then, we would like to prove Theorem 5.

Proof. The proof is a combination of (Nguyen and Tran, 2013) and (Slawski and Ben-David, 2019). Define $\boldsymbol{u} \triangleq \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\natural}$ and $\boldsymbol{v} \triangleq \widehat{\boldsymbol{\Xi}} - {(\mathbf{I}-\boldsymbol{\Pi}^{\natural})\mathbf{X}\beta^{\natural}}/{\sqrt{n}}$. In addition, we define the support set of $\boldsymbol{\beta}^{\natural}$ and $(\mathbf{I} - \boldsymbol{\Pi}^{\natural})\mathbf{X}\boldsymbol{\beta}^{\natural}$ as T and S, respectively. According to our definition, their cardinality is bounded by k and h, respectively, namely, $|T| \le k$ and $|S| \le h$. Before delving into the technical details, we first illustrate the proof outline.

- **Step I.** According to the optimality of (6), we show

$$\|\boldsymbol{u}\|_1 \le 4\sqrt{k}\|\boldsymbol{u}\|_2 + {3\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/{\lambda_{\beta}}\|\boldsymbol{v}\|_2; \tag{29}$$

$$\|\boldsymbol{v}\|_1 \le {3\sqrt{k}\lambda_{\beta}}/{\lambda_{\boldsymbol{\Xi}}}\|\boldsymbol{u}\|_2 + 4\sqrt{h}\|\boldsymbol{v}\|_2. \tag{30}$$

- **Step II.** We establish the inequality

$$\left(4\sqrt{k}\|\boldsymbol{u}\|_2 + {3\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/{\lambda_{\beta}}\|\boldsymbol{v}\|_2\right)^2 \lesssim (k\lambda_{\beta} \vee h\lambda_{\boldsymbol{\Xi}})\cdot\left[4\sqrt{k}\|\boldsymbol{u}\|_2 + {3\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/{\lambda_{\beta}}\|\boldsymbol{v}\|_2\right], \tag{31}$$

  and then obtain the upper-bound $c\cdot k\lambda_{\beta}$ for the reconstruction error $4\sqrt{k}\|\boldsymbol{u}\|_2 + {3\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/{\lambda_{\beta}}\|\boldsymbol{v}\|_2$.

- **Step III.** We upper-bound $\|\mathbf{X}\boldsymbol{u}\|_{\infty}$ as $k\sigma\sqrt{{(\log p)(\log np)}/{n}}$ and complete the proof by invoking Lemma 10.

The technical details are presented as follows. According to the definition of (6), we have

$$\frac{1}{2n}\left\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \sqrt{n}\cdot\widehat{\boldsymbol{\Xi}}\right\|_{\mathrm{F}}^2 + \lambda_{\beta}\|\widehat{\boldsymbol{\beta}}\|_1 + \lambda_{\boldsymbol{\Xi}}\|\widehat{\boldsymbol{\Xi}}\|_1 \le \frac{1}{2n}\left\|\boldsymbol{y} - \mathbf{X}\beta^{\natural} - \sqrt{n}\boldsymbol{\Xi}^{\natural}\right\|_{\mathrm{F}}^2 + \lambda_{\beta}\left\|\boldsymbol{\beta}^{\natural}\right\|_1 + \lambda_{\boldsymbol{\Xi}}\left\|\boldsymbol{\Xi}^{\natural}\right\|_1.$$

With some standard algebraic manipulation, we obtain

$$\begin{aligned}
\frac{1}{2n}\|\mathbf{X}\boldsymbol{u} + \sqrt{n}\boldsymbol{v}\|_{\mathrm{F}}^2 &\le \frac{\langle \boldsymbol{w}, \mathbf{X}\boldsymbol{u} + \sqrt{n}\boldsymbol{v}\rangle}{n} + \underbrace{\lambda_{\beta}(\left\|\boldsymbol{\beta}^{\natural}\right\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1)}_{\triangleq\ \theta_1} + \underbrace{\lambda_{\boldsymbol{\Xi}}(\left\|\boldsymbol{\Xi}^{\natural}\right\|_1 - \|\widehat{\boldsymbol{\Xi}}\|_1)}_{\triangleq\ \theta_2} \\
&\le \frac{\left\|\mathbf{X}^{\top}\boldsymbol{w}\right\|_{\infty}}{n}\cdot\|\boldsymbol{u}\|_1 + \frac{\|\boldsymbol{w}\|_{\infty}}{\sqrt{n}}\cdot\|\boldsymbol{v}\|_1 + \theta_1 + \theta_2.
\end{aligned} \tag{32}$$

For θ_1, we exploit the fact $\|\boldsymbol{\beta}^{\natural}_T\|_1 = \|\boldsymbol{\beta}^{\natural}\|_1$ and have

$$\theta_1 = \lambda_{\beta}(\|\boldsymbol{\beta}^{\natural}_T\|_1 - \|\widehat{\boldsymbol{\beta}}_T\|_1 - \|\widehat{\boldsymbol{\beta}}_{T^c}\|_1) \le \lambda_{\beta}(\|\boldsymbol{\beta}^{\natural}_T - \widehat{\boldsymbol{\beta}}_T\|_1 - \|\widehat{\boldsymbol{\beta}}_{T^c}\|_1) = \lambda_{\beta}(\|\boldsymbol{u}_T\|_1 - \|\boldsymbol{u}_{T^c}\|_1).$$

Similarly, we have $\theta_2 \le \lambda_{\boldsymbol{\Xi}}(\|\boldsymbol{v}_S\|_1 - \|\boldsymbol{v}_{S^c}\|_1)$. According to Lemma 7 and Lemma 8, we have $\lambda_{\beta} \ge \frac{2\|\mathbf{X}^{\top}\boldsymbol{w}\|_{\infty}}{n}$ and $\lambda_{\boldsymbol{\Xi}} \ge \frac{2\|\boldsymbol{w}\|_{\infty}}{\sqrt{n}}$. Summing the above together, we have

$$\frac{1}{2n}\|\mathbf{X}\boldsymbol{u} + \sqrt{n}\boldsymbol{v}\|_{\mathrm{F}}^2 \le \frac{3\lambda_{\beta}}{2}\|\boldsymbol{u}_T\|_1 + \frac{3\lambda_{\boldsymbol{\Xi}}}{2}\|\boldsymbol{v}_S\|_1 - \frac{\lambda_{\beta}}{2}\|\boldsymbol{u}_{T^c}\|_1 - \frac{\lambda_{\boldsymbol{\Xi}}}{2}\|\boldsymbol{v}_{S^c}\|_1. \tag{33}$$

Step I. Notice that the left-hand size of (33) is non-negative, we obtain

$$\begin{aligned}\|\boldsymbol{u}\|_1 = & \ \|\boldsymbol{u}_T\|_1 + \|\boldsymbol{u}_{T^c}\|_1 \le 4\|\boldsymbol{u}_T\|_1 + \frac{3\lambda_{\boldsymbol{\Xi}}}{\lambda_\beta}\|\boldsymbol{v}_S\|_1 - \underbrace{\frac{\lambda_{\boldsymbol{\Xi}}}{\lambda_\beta}\|\boldsymbol{v}_{S^c}\|_1}_{\ge 0} \\ & \overset{①}{\le} 4\sqrt{k}\|\boldsymbol{u}_T\|_2 + {}^{3\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/_{\lambda_\beta}\|\boldsymbol{v}_S\|_2 \le 4\sqrt{k}\|\boldsymbol{u}\|_2 + {}^{3\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/_{\lambda_\beta}\|\boldsymbol{v}\|_2,\end{aligned}$$

where in ① we exploit the fact such that $\boldsymbol{u}_T$ and $\boldsymbol{v}_S$ are k-sparse and h-sparse respectively. As for (30), we follow a similar approach and finish its proof.

Step II. Without loss of generality, we assume $\lambda_\beta\sqrt{k} \ge \lambda_{\boldsymbol{\Xi}}\sqrt{h}$ and have

$$\left(4\sqrt{k}\|\boldsymbol{u}\|_2 + {}^{3\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/_{\lambda_\beta}\|\boldsymbol{v}\|_2\right)^2 \asymp \lambda_\beta^{-2}\left(\lambda_\beta\sqrt{k}\|\boldsymbol{u}\|_2 + \sqrt{h}\lambda_{\boldsymbol{\Xi}}\|\boldsymbol{v}\|_2\right)^2 \le k\left(\|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2\right)^2. \tag{34}$$

On one hand, we can invoke Lemma 9 and obtain an upper-bound for $(\|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2)^2$ reading as

$$\frac{1}{2n}\|\mathbf{X}\boldsymbol{u} + \sqrt{n}\boldsymbol{v}\|_{\mathrm{F}}^2 \gtrsim \left(\|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2\right)^2. \tag{35}$$

On the other hand, (33) yields the upper-bound for $\frac{1}{2n}\|\mathbf{X}\boldsymbol{u} + \sqrt{n}\boldsymbol{v}\|_{\mathrm{F}}^2$, which can be written as

$$\frac{1}{2n}\|\mathbf{X}\boldsymbol{u} + \sqrt{n}\boldsymbol{v}\|_{\mathrm{F}}^2 \le \ \frac{3\lambda_\beta}{2}\|\boldsymbol{u}_T\|_1 + \frac{3\lambda_{\boldsymbol{\Xi}}}{2}\|\boldsymbol{v}_S\|_1 \le \frac{3\lambda_\beta}{2}\sqrt{k}\|\boldsymbol{u}\|_2 + \frac{3\lambda_{\boldsymbol{\Xi}}}{2}\sqrt{h}\|\boldsymbol{v}\|_2. \tag{36}$$

Combining (34), (35), and (36) then yields the relation

$$\left(4\sqrt{k}\|\boldsymbol{u}\|_2 + {}^{3\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/_{\lambda_\beta}\|\boldsymbol{v}\|_2\right)^2 \lesssim \frac{k}{n}\|\mathbf{X}\boldsymbol{u} + \sqrt{n}\boldsymbol{v}\|_{\mathrm{F}}^2 \lesssim k\left(4\lambda_\beta\sqrt{k}\|\boldsymbol{u}\|_2 + 3\lambda_{\boldsymbol{\Xi}}\sqrt{h}\|\boldsymbol{v}\|_2\right).$$

Dividing both sides by $4\sqrt{k}\|\boldsymbol{u}\|_2 + {}^{3\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/_{\lambda_\beta}\|\boldsymbol{v}\|_2$ then completes the proof of (31).

Step III. Our goal is upper-bound $\|\mathbf{X}\boldsymbol{u}\|_\infty$, which reads as

$$\begin{aligned}\|\mathbf{X}\boldsymbol{u}\|_\infty = & \ \max_i |\langle \mathbf{X}_{i,:}, \boldsymbol{u}\rangle| \le \max_{ij}|\mathbf{X}_{ij}| \cdot \|\boldsymbol{u}\|_1 \overset{②}{\lesssim} \sqrt{\log np}\left(\sqrt{k}\|\boldsymbol{u}\|_2 + {}^{\sqrt{h}\lambda_{\boldsymbol{\Xi}}}/_{\lambda_\beta}\|\boldsymbol{v}\|_2\right) \\ & \overset{③}{\lesssim} \sigma\sqrt{\log np}\left(k\sqrt{\frac{\log p}{n}} \vee h\sqrt{\frac{\log n}{n}}\right),\end{aligned}$$

where in ② we condition on the event $\max_{ij}|\mathbf{X}_{ij}| \lesssim \sqrt{\log np}$, and in ③ we use the inequality in (31). Provided that

$$\mathsf{SNR} \gtrsim \frac{n^{2(1+\varepsilon)}(n-1)^2}{4\pi}\left[\sqrt{\log np}\left(k\sqrt{\frac{\log p}{n}} \vee h\sqrt{\frac{\log n}{n}}\right) + 2\log(n^{1+\varepsilon}(n-1))\right]^2,$$

we invoke Lemma 10 and complete the proof such that ground-truth permutation $\mathbf{\Pi}^\natural$ can be obtained with probability exceeding $1 - 2n^{-\varepsilon}$. □

C.1 Supporting Lemmas

This subsection collects the supporting lemmas and useful facts used in the proof thereof.

Lemma 6. *We have* $\max_{1\leq i\leq n, 1\leq j\leq p}|\mathbf{X}_{ij}| \leq 2\sqrt{\log np}$ *with probability exceeding* $1-(np)^{-1}$.

Lemma 7. *We have* $\left\|\mathbf{X}^\top \boldsymbol{w}\right\|_\infty \lesssim \sigma\sqrt{n\log p}$ *with probability exceeding* $1-c_0 p^{-c_1}$.

Lemma 8. *We have* $\|\boldsymbol{w}\|_\infty \lesssim \sigma\sqrt{\log n}$ *with probability* $1-c_0 n^{-c_1}$.

Since the above results are quite standard, we list them without giving a detailed proof.

Lemma 9 (Lemma 1 In Nguyen and Tran (2013))**.** *Consider the optimal solution* $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Xi}})$ *to* (6) *with regularizer coefficients being set as* $\lambda_{\boldsymbol{\beta}} \asymp \sigma\sqrt{\log p/n}$ *and* $\lambda_{\boldsymbol{\Xi}} \asymp \sigma\sqrt{\log n/n}$*, respectively. Assuming that* $n \gtrsim k\log p$ *and* $h \lesssim \frac{n}{\log n}$*, we have*

$$\frac{1}{\sqrt{n}}\left\|\mathbf{X}\boldsymbol{u} + \sqrt{n}\boldsymbol{v}\right\|_{\mathrm{F}} \gtrsim \|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2$$

hold with probability at least $1-c_0 e^{-c_1 n}$*, where* $c_0, c_1 > 0$ *are some fixed positive constants.*

Lemma 10 (Theorem 3 (Part (a)) in Slawski and Ben-David (2019))**.** *Conditional on the event* $\mathcal{E}_{\widetilde{\boldsymbol{\beta}}}$ *such that* $\mathcal{E}_{\widetilde{\boldsymbol{\beta}}} \triangleq \left\{\|\mathbf{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\natural)\|_\infty \leq \sigma\Delta\right\}$*, we reconstruct the permutation matrix via* $\widehat{\boldsymbol{\Pi}} = \operatorname{argmax}_{\boldsymbol{\Pi}}\langle \mathbf{Y}, \boldsymbol{\Pi}\mathbf{X}\widetilde{\boldsymbol{\beta}}\rangle$*. Provided that*

$$\mathsf{SNR} > \frac{n^2(n-1)^2}{4\delta^2\pi}\left[\Delta + 2\log\frac{n(n-1)}{\delta}\right]^2,$$

Then, we can obtain the ground-truth permutation with probability exceeding $1-2\delta$*, i.e.,* $\mathbb{P}(\widehat{\boldsymbol{\Pi}} = \boldsymbol{\Pi}^\natural | \mathcal{E}_{\widetilde{\boldsymbol{\beta}}}) \geq 1-2\delta$*.*

D Useful Facts About Probability Inequalities

For the self-containing of this paper, we list some useful facts about probability inequalities in this section.

Lemma 11 (Lemma 2.2 In Dasgupta and Gupta (2003))**.** *For a projection matrix* $\mathsf{P}_{d_1\to d_2}$ *which projects a fixed vector* $\mathbf{Z} \in \mathbb{R}^{d_1}$ *to a uniformly random subspace with dimension* d_2*, we have*

$$\mathbb{P}\left(\|\mathsf{P}_{d_1\to d_2}\mathbf{Z}\|_2^2 \leq \frac{\alpha d_2}{d_1}\|\mathbf{Z}\|_2^2\right) \leq \exp\left(\frac{d_2}{2}\left(\log\alpha - \alpha + 1\right)\right), \alpha < 1;$$
$$\mathbb{P}\left(\|\mathsf{P}_{d_1\to d_2}\mathbf{Z}\|_2^2 \geq \frac{\alpha d_2}{d_1}\|\mathbf{Z}\|_2^2\right) \leq \exp\left(\frac{d_2}{2}\left(\log\alpha - \alpha + 1\right)\right), \alpha > 1.$$