Permutation Recovery from Multiple Measurement Vectors in Unlabeled Sensing

Hang Zhang, Martin Slawski, and Ping Li
Cognitive Computing Lab
Baidu Research USA
10900 NE 8th ST, Bellevue, WA 98004, USA

***Abstract*—[1] In "Unlabeled Sensing", one observes a set of linear measurements of an underlying signal with incomplete or missing information about their ordering, which can be modeled in terms of an unknown permutation. Previous work on the case of a single noisy measurement vector has exposed two main challenges: 1) a high requirement concerning the *signal-to-noise ratio* (snr), i.e., approximately of the order of n^5, and 2) a massive computational burden in light of NP-hardness in general. In this paper, we study the case of *multiple* noisy measurement vectors (MMVs) resulting from a *common* permutation and investigate to what extent the number of MMVs m facilitates permutation recovery by "borrowing strength". The above two challenges have at least partially been resolved within our work. First, we show that a large stable rank of the signal significantly reduces the required snr which can drop from a polynomial in n for $m = 1$ to a constant for $m = \Omega(\log n)$, where m denotes the number of MMVs and n denotes the number of measurements per MV. This bound is shown to be sharp and is associated with a phase transition phenomenon. Second, we propose computational schemes for recovering the unknown permutation in practice. For the "oracle case" with the known signal, the maximum likelihood (ML) estimator reduces to a linear assignment problem whose global optimum can be obtained efficiently. For the case in which both the signal and permutation are unknown, the problem is reformulated as a bi-convex optimization problem with an auxiliary variable, which can be solved by the Alternating Direction Method of Multipliers (ADMM). Numerical experiments based on the proposed computational schemes confirm the tightness of our theoretical analysis.**

I. Introduction

Noisy linear sensing with m measurement vectors is described by the relation

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

where $\mathbf{Y} \in \mathbb{R}^{n\times m}$ represents the observed n measurements, $\mathbf{X} \in \mathbb{R}^{n\times p}$ represents the sensing matrix, and the columns of $\mathbf{B}^* \in \mathbb{R}^{p\times m}$ contain m signals of interest with dimension p each, and $\mathbf{W} \in \mathbb{R}^{n\times m}$ represents additive noise. Model (1) also arises in linear regression modeling with m response variables and p explanatory variables [1]. Least squares regression yields the estimator $\widehat{\mathbf{B}} = (\mathbf{X})^\dagger\mathbf{Y}$, where $(\cdot)^\dagger$ denotes the Moore-Penrose inverse. The properties of $\widehat{\mathbf{B}}$ under various assumptions on the noise $\mathbf{W}$ are well-known. In this paper, we consider the more challenging situation in which we observe n measurements with missing or incomplete information about their ordering, i.e., the correspondence between the rows of $\mathbf{Y}$ and the rows of $\mathbf{X}$ has been lost. Put differently, we observe data according to (1) up to an unknown permutation:

$$\mathbf{Y} = \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^* + \mathbf{W}, \tag{2}$$

where $\mathbf{\Pi}^*$ is an n-by-n permutation matrix. Ignoring the unknown permutation can significantly impair performance with regard to the estimation of $\mathbf{B}^*$. We herein consider recovery of $\mathbf{\Pi}^*$ given $(\mathbf{X}, \mathbf{Y})$. The latter suffices for signal recovery since with restored correspondence the setup becomes standard. In addition, recovery of $\mathbf{\Pi}^*$ may be of its own interest, e.g., in *record linkage* in which two data sets containing different pieces of information about a common set of entities are integrated into a single comprehensive data set [2].

A. Related Work

The work [3] discusses *signal recovery* under setup (2) referred to as "Unlabeled Sensing" therein for the case of a single measurement vector ($m = 1$) and no noise ($\mathbf{W} = 0$). It is shown that if the entries of the sensing matrix $\mathbf{X}$ are drawn from a continuous distribution over $\mathbb{R}$, the condition $n \geq 2p$ is required for signal recovery by means of exhaustive search over all permutation matrices. The authors also motivate the problem from a variety of applications, including the reconstruction of spatial fields using mobile sensors, time-domain sampling in the presence of clock jitter, and multi-target tracking in radar. Alternative proofs of the main result in [3] are shown in [4], [5].

A number of recent papers discuss the case $m = 1$ and Gaussian $\mathbf{W}$. The paper [6] establishes the statistical limits of exact and approximate permutation recovery based on the ratio of signal energy and noise variance henceforth referred to as "snr". In [6], it is also demonstrated that the least squares estimation of $\mathbf{\Pi}^*$ is NP-hard in general. In [7], a polynomial-time approximation algorithm is proposed, and lower bounds on the required snr for approximate signal recovery in the noisy case are shown; related results can be found in [8], [9]. The works [9]–[12] discuss both signal and permutation recovery if $\mathbf{\Pi}^*$ only permutes a small fraction of the rows of the sensing matrix. An interesting variation of (2) in which $\mathbf{\Pi}^*$ is an unknown selection matrix that selects a fraction measurements in an order-preserving fashion is studied in [13]. The paper [14] develops the approach in [13] further by combining it with a careful branch-and-bound scheme to solve general unlabeled sensing problems.

Several papers [10], [11], [15], [16] have studied the setting of multiple measurement vectors ($m > 1$) and associated potential benefits for permutation recovery. The paper [15] discusses a practical branch-and-bound scheme for permutation recovery but does not provide theoretical insights. The

[1]Partial preliminary results appeared in 2019 IEEE International Symposium on Information Theory (ISIT'19), Paris, France.

work [16] analyzes the *denoising problem*, i.e., recovery of $\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*$, rather than individual recovery of $\mathbf{\Pi}^*$ and $\mathbf{B}^*$. In [10], [11] the number of permuted rows in the sensing matrix is assumed to be small, and are treated as outliers. Methods for robust regression and outlier detection are proposed to perform signal recovery. While both [10], [11] also contain achievability results for permutation recovery given an estimate of the signal, none of these works provides information-theoretic lower bounds to assess the sharpness of the results. Moreover, the method in [10] limits the fraction of permuted rows to a constant multiple of the reciprocal of the signal dimension p, while the method in [11] requires the number of MMVs m to be of the same order of p and additionally exhibits an unfavorable running time that is cubic in the number of measurements. In the present paper, we eliminate the limitations in [10], [11] to a good extent.

B. Contribution

Results in [6] on the case $m = 1$ indicate that the maximum likelihood (ML) estimator in Eq. (5) can be regarded as impractical from both statistical and computational viewpoints. On one hand, successful recovery of $\mathbf{\Pi}^*$ requires snr $\sim n^c$, where $c > 0$ is a constant that is approximately equal to 5 according to simulations. As n grows, this requirement becomes prohibitively strong. On the other hand, the ML estimator Eq. (5) has been proven to be NP-hard except for the special case $m = 1$ and $p = 1$. To the best of our knowledge, no efficient algorithm has been proposed yet. In this paper, by contrasting $m = 1$ and $m \gg 1$, our goal is to tackle both obstacles. Before giving a detailed account of our contribution, we first define a crucial quantity, the signal-to-noise-ratio (snr)

$$\text{snr} = \|\mathbf{B}^*\|_{\text{F}}^2/(m \cdot \sigma^2). \tag{3}$$

- We improve the requirement snr $\sim n^c$ to roughly snr $\sim n^{c/\varrho(\mathbf{B}^*)}$, where $\varrho(\mathbf{B}^*) = \|\mathbf{B}^*\|_{\text{F}}^2/\|\mathbf{B}^*\|_{\text{OP}}^2$ is the stable rank of $\mathbf{B}^*$. Once $\varrho(\mathbf{B}^*)$ is of the order $\Omega(\log n)$, we conclude that the requirement on the snr is well-controlled even when n approaches infinity. The underlying intuition is that larger values of m lead to relaxed requirements on the snr since (i) the overall signal energy increases, (ii) all MMVs result from the same permutation matrix $\mathbf{\Pi}^*$, which is expected to yield extra information. In our analysis, (i) is reflected by conditions on permutation recovery involving dependence on the overall signal energy, while (ii) enters via a dependence on the stable rank of the signal matrix $\mathbf{B}^*$.
- We propose practical algorithms for recovery of Π^* and B^* via least squares fitting, which is an NP-hard problem except for the special case with $p = m = 1$. In our approach, we introduce an auxiliary variable, which prompts a bi-convex optimization problem that can be tackled via an efficient ADMM (alternating direction method of multipliers) scheme [17]. To achieve computational speed-ups, we propose two initialization estimators, the "averaging estimator" and the "eigenvalue estimator", as warm-starts for ADMM. The empirical study suggests that convergence will be obtained within 10 steps in most cases.

C. Outline

The rest of the paper is organized as follows. In Section II, we review the sensing model. In Section III, we consider the "oracle case", where the signal matrix $\mathbf{B}^*$ is known, and study the conditions for exact recovery of $\mathbf{\Pi}^*$. In Section IV, the practical case in which $\mathbf{B}^*$ is unknown as well is investigated. Practical algorithms for approximate recovery of $\mathbf{\Pi}^*$ are developed in Section V. Simulations and concluding remarks are provided in Section VI and Section VII, respectively.

II. System Model

We recall that the sensing model under consideration reads

$$\mathbf{Y} = \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^* + \mathbf{W}, \tag{4}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times m}$ represents the results of the sensing process, $\mathbf{\Pi}^* \in \mathbb{R}^{n\times n}$ denotes the unknown permutation matrix, $\mathbf{X} \in \mathbb{R}^{n\times p}$ ($n \geq 2p$) is the sensing matrix, $\mathbf{B}^* \in \mathbb{R}^{p\times m}$ is the matrix of signals, and $\mathbf{W} \in \mathbb{R}^{n\times m}$ is the sensing noise. For what follows, we assume that the entries (X_{ij}) of $\mathbf{X}$ are i.i.d. standard Gaussian random variables, i.e., $X_{ij} \sim \mathscr{N}(0, 1)$, $1 \leq i \leq n,\ 1 \leq j \leq p$. Likewise, we assume that the entries of $\mathbf{W}$ are i.i.d. $\mathscr{N}(0, \sigma^2)$-random variables, where $\sigma^2 > 0$ denotes the noise variance. The maximum likelihood (ML) estimator of $(\mathbf{\Pi}^*, \mathbf{B}^*)$ then results as the least squares solution

$$(\widehat{\mathbf{\Pi}}, \widehat{\mathbf{B}}) = \text{argmin}_{(\mathbf{\Pi},\mathbf{B})}\ \|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}\|_{\text{F}}^2. \tag{5}$$

Note that for a fixed permutation matrix $\mathbf{\Pi}$, we obtain

$$\widehat{\mathbf{B}}(\mathbf{\Pi}) = (\mathbf{\Pi}\mathbf{X})^\dagger\mathbf{Y}, \tag{6}$$

From the above, we can see the importance of accurate estimation of $\mathbf{\Pi}^*$ in a least squares approach since errors may significantly degrade the quality of the corresponding estimator $\widehat{\mathbf{B}}$, while exact permutation recovery, i.e., $\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*$ yields the same quality as the usual least squares estimator in the absence of $\mathbf{\Pi}^*$. In the following, we put estimation of $\mathbf{B}^*$ aside and concentrate on analyzing the determining factors for estimation of $\mathbf{\Pi}^*$. Broadly speaking, permutation recovery involves two main sources of difficulty.

- *Sensing noise* $\mathbf{W}$. Even in the oracle case in which $\mathbf{B}^*$ is known and computation of the ML estimator of $\mathbf{\Pi}^*$ reduces to the *linear assignment problem* [18]

$$\widehat{\mathbf{\Pi}} = \text{argmax}_{\mathbf{\Pi}}\ \left\langle \mathbf{\Pi},\ \mathbf{Y}\mathbf{B}^{*\top}\mathbf{X}^\top \right\rangle, \tag{7}$$

  whose solution can be obtained efficiently, recovery of $\mathbf{\Pi}^*$ is still likely to fail if the noise level σ^2 is large enough.
- *Unknown* $\mathbf{B}^*$. Compared with the oracle case above, we have no access to $\mathbf{B}^*$ in practice, which makes recovery more challenging.

In the sequel, we will show that the sensing noise $\mathbf{W}$ constitutes the major difficulty in recovering $\mathbf{\Pi}^*$ rather than missing knowledge of $\mathbf{B}^*$. But first, we define the following notations.

Notations: Positive constants are denoted by c, c', c_0, c_1, etc. We write $a \lesssim b$ if there is a constant c_0 such that $a \leq c_0 b$. Similarly, we define $\gtrsim$. If both $a \lesssim b$ and $a \gtrsim b$ hold, we write $a \asymp b$. The maximum of a and b is denoted by $a \vee b$ while the minimum of a and b is denoted by $a \wedge b$. The Frobenius norm

of a matrix is represented as $\|\cdot\|_{\mathrm{F}}$ while the operator norm is denoted as $\|\cdot\|_{\mathrm{OP}}$, whose definition can be found in [19] (Section 2.3, P71). The ratio $\varrho(\cdot) = \|\cdot\|_{\mathrm{F}}^2/\|\cdot\|_{\mathrm{OP}}^2$ represents the stable-rank while $r(\cdot)$ represents the rank. The *signal-to-noise-ratio* (snr) is defined as $\mathrm{snr} = \|\mathbf{B}^*\|_{\mathrm{F}}^2/(m\sigma^2)$. Additional notation can be found in Appendix A.

III. Failure of recovery

This section presents conditions under which exact or approximate recovery of $\mathbf{\Pi}^*$ is expected to fail with high probability. Investigation of this case is intended to provide valuable insights into the fundamental statistical limits. We consider the "oracle case" with information of $\mathbf{B}^*$, whose limits apply to the case of unknown $\mathbf{B}^*$ as well, since it is hopeless to recover $\mathbf{\Pi}^*$ even if the knowledge of $\mathbf{B}^*$ does not suffice for recovery.

Compared with the case $m = 1$ in which snr is the only prominent factor in determining the recovery performance [6], our analysis uncovers another very important factor, namely, the energy distribution over singular values. Our work shows that a more uniform spread of the signal energy over singular values can greatly facilitate the recovery of $\mathbf{\Pi}^*$.

A. Recovery of $\mathbf{\Pi}^*$

For our first result, $\mathbf{\Pi}^*$ is supposed to be a random variable independent of $\mathbf{X}$ and $\mathbf{W}$, drawn from a probability distribution supported on a subset $\mathscr{H}$ of the set of n-by-n permutation matrices. Using Fano's inequality, the following inachievability result can be shown.

Theorem 1. *Consider the condition*

$$\frac{1}{2}\sum_i \log\left(1+\frac{\lambda_i^2}{\sigma^2}\right) + \frac{\log\left(|\mathscr{H}|\right)}{2n} < \frac{H(\mathbf{\Pi}^*)-1}{n}, \tag{8}$$

where λ_i denotes the i-th singular value of $\mathbf{B}^*$, and $\mathscr{H}$ and $H(\mathbf{\Pi}^*)$ denote the support and the entropy of the random permutation matrix $\mathbf{\Pi}^*$, respectively. Under (8)*, for any estimator $\widehat{\mathbf{\Pi}}$ of $\mathbf{\Pi}^*$, it holds that* $\Pr(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*) > 1/2$.

In Thm. 1, the set $\mathscr{H}$ and the entropy $H(\mathbf{\Pi}^*)$ capture the amount of prior information about $\mathbf{\Pi}^*$. In the absence of prior knowledge, $\mathbf{\Pi}^*$ can be regarded as uniformly distributed among all possible permutation matrices, which corresponds to maximal entropy $\log(n!) \approx n\log(n)$. The availability of prior information leads to reduced entropy. For example, it may be known that $d_H(\mathbf{I}, \mathbf{\Pi}^*) \leq D$, where $d_H(\mathbf{I}, \mathbf{\Pi}) \triangleq \sum_{i=1}^n \mathbb{1}(\mathbf{\Pi}_{i,i} \neq 1)$ denotes the Hamming distance between a permutation matrix $\mathbf{\Pi}$ and the identity. In this case, the entropy is upper bounded by $\log(n!/(n-D)!)$, which means that the inachievability condition (8) is less likely to be fulfilled.

The second major ingredient in condition (8) is the term $\sum_i \log\left(1+\lambda_i^2/\sigma^2\right)$. Since each singular value λ_i is determined by the matrix $\mathbf{B}^*$ as whole rather than by individual columns, we conclude that linear independence among multiple measurements can positively impact the recovery of $\mathbf{\Pi}^*$, which implies extra benefits apart from mere energy accumulation.

When maximizing the term $\sum_i \log\left(1+\lambda_i^2/\sigma^2\right)$ given fixed signal energy $\|\mathbf{B}^*\|_{\mathrm{F}}^2 = \sum_i \lambda_i^2$, it is easy to determine the most favorable configuration to avoid failure of recovery: the signal energy is evenly spread over all singular values. In contrast, if $\mathbf{B}^*$ has rank one with all signal energy concentrated on the principal singular value, condition (8) reduces to the same as for a single MV ($m = 1$) with signal energy $\|\mathbf{B}^*\|_{\mathrm{F}}^2$ since

$$\sum_i \log\left(1+\frac{\lambda_i^2}{\sigma^2}\right) = \log\left(1+\frac{\lambda_1^2}{\sigma^2}\right) = \log\left(1+\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2}\right).$$

This indicates that in accordance with the intuition of "borrowing strength" across different sets of measurements, performance is expected to improve as the rank of $\mathbf{B}^*$ increases. Moreover, the result of Thm. 1 constitutes a fundamental limit as it applies to *any* estimator.

The statement below provides a condition for failure of recovery when using the ML estimator in Eq. (5), which is computationally feasible if $\mathbf{B}^*$ is known. In this statement, $\mathbf{\Pi}^*$ is considered as fixed (non-random) as is the case throughout the paper with the exception of Thm. 1 and Corol. 3.

Proposition 2. *Given knowledge of $\mathbf{B}^*$, the ML estimator $\widehat{\mathbf{\Pi}}$ in Eq. (7) satisfies* $\Pr\left(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*\right) \geq \frac{49(1-n^{-1})}{64}$ *for $n \geq 10$ if*

$$\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2} \leq \frac{2\log n}{4\left(1+2\sqrt{\frac{\log 2}{c\varrho(\mathbf{B}^*)}}\right)^2}, \tag{9}$$

where $\varrho(\mathbf{B}^*) = \|\mathbf{B}^*\|_{\mathrm{F}}^2/\|\mathbf{B}^*\|_{\mathrm{OP}}^2$ is the stable rank *of $\mathbf{B}^*$.*

The proposition states that the total signal energy given by $m \times \mathrm{snr}$ should be at least of the order $\log n$ to avoid failure in recovery. This is in agreement with what can be concluded from Thm. 1 in the full-rank case and a uniform prior for $\mathbf{\Pi}^*$.

B. Approximate recovery of $\mathbf{\Pi}^*$

Exact recovery of $\mathbf{\Pi}^*$ may not always be necessary. The following corollary of Thm. 1 yields a condition under which even a close approximation w.r.t. the Hamming distance, i.e., $d_H(\widehat{\mathbf{\Pi}},\ \mathbf{\Pi}^*) \leq D$, cannot be guaranteed.

Corollary 3. *Provided that*

$$\frac{1}{2}\sum_i \log\left(1+\frac{\lambda_i^2}{\sigma^2}\right) + \frac{\log 2}{n} \leq \frac{\log(n-D+1)!}{2n},$$

where λ_i denotes the i-th singular value of $\mathbf{B}^*$, we have $\Pr\left(d_H(\widehat{\mathbf{\Pi}},\ \mathbf{\Pi}^*) \geq D\right) \geq \frac{1}{2}$ *for any estimator $\widehat{\mathbf{\Pi}}$ of $\mathbf{\Pi}^*$.*

Comparing the above result with Thm. 1, one can see that the essentially only difference is the replacement of the term $H(\mathbf{\Pi}^*)$ by $\log(n-D+1)!$. Here is an intuitive interpretation:

- The set of n-by-n permutation matrices under consideration can be covered by a subset $\{\mathbf{\Pi}^{(1)}, \mathbf{\Pi}^{(2)}, \cdots,\ \mathbf{\Pi}^{((n-D+1)!)}\}$ such that for any permutation matrix $\mathbf{\Pi}$, there exists an element $\mathbf{\Pi}^\dagger \in \{\mathbf{\Pi}^{(1)}, \mathbf{\Pi}^{(2)}, \cdots,\ \mathbf{\Pi}^{((n-D+1)!)}\}$ such that $d_H(\mathbf{\Pi},\ \mathbf{\Pi}^\dagger) \leq D$.
- We would like to recover $\mathbf{\Pi}^\dagger$ from data $(\mathbf{X}, \mathbf{Y})$.

As a result, since the cardinality of the covering is $(n-D+1)!$, we encounter the term $\log(n-D+1)!$ in place of $H(\mathbf{\Pi}^*) \leq \log(n!)$; setting $D = 0$ or 1 gives back Thm. 1.

To conclude this section, we would like to emphasize that the above conditions reflect the price to compensate for the

uncertainty induced by the sensing noise $\mathbf{W}$, as there is no uncertainty in $\mathbf{B}^*$ involved. In the next section, we will study conditions for the successful recovery.

IV. Successful Recovery

In the previous section, we have studied conditions under which recovery is expected to fail. In this section, we state conditions under which the true permutation $\mathbf{\Pi}^*$ can be recovered with high probability, for the oracle case with known $\mathbf{B}^*$ as well as the realistic case with unknown $\mathbf{B}^*$. For the sake of transparency, we provide explicit values for numerical constants in most cases even though no specific effort was made to optimize these constants. We believe these constants can be improved further.

A. Oracle case: known $\mathbf{B}^*$

In this case, the ML estimator in Eq. (5) is re-written as Eq. (7). The condition on the snr in the following statement can serve both as an upper bound for the failure of permutation recovery and as a lower bound for the more challenging case with unknown $\mathbf{B}^*$.

Theorem 4. *Given knowledge of $\mathbf{B}^*$, the ML estimator in Eq. (5) satisfies*

$$\Pr\left(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*\right) \leq \frac{2\alpha_0^{2\kappa\varrho(\mathbf{B}^*)}}{n^2},$$

provided

$$\log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2}\right) \geq \frac{8}{\kappa\varrho(\mathbf{B}^*)}\log(n) + 4\log(\alpha_0^{-1}) + \log\left(32\left(2\log n + \kappa\varrho(\mathbf{B}^*)\log\alpha_0^{-1}\right)\right), \tag{10}$$

where $0 < \alpha_0 < 1$, $\kappa > 0$ are universal constants.

We would like to compare this result with the bound on incorrect recovery in Thm. 1. First, we consider the full-rank case with constant singular values, i.e., $\mathbf{B}^{*\top}\mathbf{B}^* = \gamma\mathbf{I}$, where γ is a positive constant; in particular, $\varrho(\mathbf{B}^*) = m$. Then a simple term re-arrangement of Eq. (10) suggests that having

$$\log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\varrho(\mathbf{B}^*)\sigma^2}\right) = \log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m\sigma^2}\right) \gtrsim \frac{\log(n)}{\varrho(\mathbf{B}^*)} \tag{11}$$

ensures success, while Thm. 1 suggests that

$$\log\left(1 + \frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m\sigma^2}\right) \lesssim \frac{\log(n)}{\varrho(\mathbf{B}^*)} \tag{12}$$

implies failure. Conditions (11) and (12) thus match up to multiplicative factors.

Next, we consider the rank-1 case. Eq. (10) suggests that $\log(\|\mathbf{B}^*\|_{\mathrm{F}}^2/\sigma^2) \gtrsim \log n$ ensures success, while Thm. 1 suggests that $\log(1+\|\mathbf{B}^*\|_{\mathrm{F}}^2/\sigma^2) \lesssim \log n$ leads to failure. Putting them together, we conclude tightness for this case.

Finally, we would like to provide an illustration of the benefits brought by large stable rank $\varrho(\mathbf{B}^*)$. We compare the snr requirement for different $\varrho(\mathbf{B}^*)$ and list them in Tab. I. To obtain successful recovery, $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2\right)}{\log(n)}$ should be roughly 5. We can see that as $\varrho(\mathbf{B}^*)$ increases from 1 to 100, the requirement on the snr drops from 10^{15} to 0.41.

$\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2\right)}{\log(n)}$	1	2	3	4	5	6
$\varrho(\mathbf{B}^*) = 1$	10^3	10^6	10^9	10^{12}	10^{15}	10^{18}
$\varrho(\mathbf{B}^*) = 10$	1	2.98	6.94	14.85	30.62	62.10
$\varrho(\mathbf{B}^*) = 20$	0.41	1.00	1.82	2.98	4.62	6.94
$\varrho(\mathbf{B}^*) = 50$	0.15	0.32	0.51	0.74	1.00	1.29
$\varrho(\mathbf{B}^*) = 100$	0.07	0.15	0.23	0.32	0.41	0.51

TABLE I: snr requirement when $n = 1000$, $p = 100$, and $\mathbf{B}^*_{:,i} = \mathbf{e}_i$, where $\mathbf{e}_i$ denotes the i-th canonical basis vector.

B. Realistic case: unknown $\mathbf{B}^*$

For this case of unknown $\mathbf{B}^*$, we first present a basic result that will be improved upon later under additional assumptions.

Theorem 5. *Fix $\epsilon > 0$. Provided that* $\mathrm{snr} \times n^{-\frac{2n}{n-p}} \geq 1$, *if*

$$\frac{\log(m \times \mathrm{snr})}{380} \geq \left(1 + \epsilon + \frac{n}{190(n-p)}\right)\log(n) + \frac{1}{2}\log r(\mathbf{B}^*), \tag{13}$$

then the ML estimator (5) *satisfies*

$$\Pr\left(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*\right) \leq 10.36\left(\frac{1}{n^\epsilon(n^\epsilon-1)} \vee \frac{1}{n^\epsilon}\right),$$

as long as $n > C(\epsilon)$, where $C(\epsilon) > 0$ is a positive constant depending only on ϵ.

Note that since we have $r(\mathbf{B}^*) \leq (m \wedge p)$ and $p \leq n/2$, the bound in Eq. (13) can be simplified to $\log(m \times \mathrm{snr}) \geq 380\left(\frac{3}{2} + \epsilon + \frac{n}{190(n-p)}\right)\log(n)$, which suggests that perfect recovery can be achieved with high probability if $\log(m \times \mathrm{snr}) \gtrsim \log n$.

Comparing the above result with Thm. 1, we can see that our bound is tight for the rank-1 case, since both theorems show that the total energy should be of the order $\log(1 + m \times \mathrm{snr}) \gtrsim \log n$ to obtain correct recovery. However, the above theorem fails to capture potential improvement brought by higher measurement diversity, namely, larger stable rank $\varrho(\mathbf{B}^*)$. Thm. 5 suggests that multiple measurements behave like a single measurement with the same energy level, which can be far from the actual behavior beyond the rank-1 case.

In the sequel, we state an improved bound under additional assumptions on $d_H(\mathbf{I}, \mathbf{\Pi}^*)$ and $\varrho(\mathbf{B}^*)$.

Theorem 6. *For a fixed $\epsilon > 0$, provided that* $\mathrm{snr} > 26.2$, $d_H(\mathbf{I}, \mathbf{\Pi}^*) \leq h_{\max}$, $\varrho(\mathbf{B}^*) \geq 5(1+\epsilon)\log(n)/c_0$, *and $h_{\max} r(\mathbf{B}^*) \leq n/8$, then if*

$$\log(\mathrm{snr}) \geq \frac{288(1+\epsilon)\log(n)}{\varrho(\mathbf{B}^*)} + 33.44, \tag{14}$$

the constrained ML estimator (5) *subject to the constraint $d_H(\mathbf{I}, \mathbf{\Pi}) \leq h_{\max}$ satisfies*

$$\Pr\left(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*\right) \leq 10\left(\frac{1}{n^\epsilon(n^\epsilon-1)} \vee \frac{1}{n^\epsilon}\right),$$

as long as $n > C(\epsilon)$, where $C(\epsilon) > 0$ is a positive constant depending only on ϵ.

We here comment on the new constraint $d_H(\mathbf{I}, \mathbf{\Pi}^*) \leq h_{\max}$. To ensure that signal diversity as quantified by $\varrho(\mathbf{B}^*)$ improves the recovery performance, we require $\varrho(\mathbf{B}^*) = \Omega(\log n)$. In this case, we obtain the condition $\mathrm{snr} \geq C$ for some constant

$C > 0$, which then also matches the assertion in Thm. 4. At the same time, $h_{\max}$ is required to be of the order

$$h_{\max} \lesssim \frac{n}{\log(n)},$$

which is only slightly sub-optimal compared to the order $\mathscr{O}(n)$. Simulation results ($p = 1, m = 1$) imply the upper bound $h_{\max}$ in the constraint $d_H(\mathbf{I}, \mathbf{\Pi}^*) \leq h_{\max}$ can be safely relaxed to a linear fraction of n. We believe these simulation results hold universally and the constraint on $d_H(\mathbf{I}, \mathbf{\Pi}^*)$ can be avoided with more advanced tools.

Since the order for the required snr to achieve correct recovery remains the same as in Thm. 4, we can draw the conclusion that the major difficulty in recovering $(\mathbf{\Pi}^*, \mathbf{B}^*)$ is due to the sensing noise $\mathbf{W}$ while the fact that $\mathbf{B}^*$ is not given a priori does not change the level of difficulty significantly.

Additionally, comparison with the condition for failure in Thm. 1 lets us conjecture a *phase transition* since $\log(\text{snr}) \gtrsim \log(n)/\varrho(\mathbf{B}^*)$ leads to success while $\log(\text{snr}) \lesssim \log(n)/\varrho(\mathbf{B}^*)$ leads to failure, where the right hand sides only differ in multiplicative constants.

Having established the statistical limits, we will then turn to numerical aspects and present several computational schemes to recover the permutation matrix $\mathbf{\Pi}^*$ in the next section.

V. Computational Approach

In this section, we discuss the computational aspects of the problem. Note that for $p = 1,\ m = 1$, the ML estimator $\widehat{\mathbf{\Pi}}$ in (5) can be computed by solving a linear assignment problem [6]. For all other cases, namely $p \geq 2$, the computation has been proved to be NP-hard in general [6]. To the best of our knowledge, no efficient algorithm is known.

A. Oracle recovery

In this part, we discuss how to recover the permutation matrix $\mathbf{\Pi}^*$ with known $\mathbf{B}^*$. Note that in this case the ML estimator in Eq. (7) reduces to a linear assignment problem [18] that can be solved by the Hungarian algorithm [20] or the auction algorithm [21].

B. Sorting-based method for $p = 1$

In this part, we restrict ourselves to the case $p = 1$ and propose to recover $\widehat{\mathbf{\Pi}}$ by sorting. Note that when $p = 1$ and $m = 1$, the solution of Eq. (5) can be found by sorting $\mathbf{X}$ with respect to $\mathbf{Y}$ as in [6]. For the case $m \geq 2$, we relax the original problems and obtain approximate solutions via sorting. Two estimators are proposed whose details are summarized in Alg. 1 and Alg. 2, respectively. In the following, we present the intuition underlying the respective designs.

a) Averaging estimator: With a fixed $\mathbf{\Pi}$ given, we can estimate $\mathbf{B}_{:,i}$ as $\langle \mathbf{\Pi}\mathbf{X}, \mathbf{Y}_{:,i}\rangle / \|\mathbf{X}\|_{\mathrm{F}}^2$. Back-substitution gives us

$$\|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}\|_{\mathrm{F}}^2 = \sum_{i=1}^{m} \left(\|\mathbf{Y}_{:,i}\|_2^2 - \frac{\left(\langle \mathbf{Y}_{:,i}, \mathbf{\Pi}\mathbf{X}\rangle\right)^2}{\|\mathbf{X}\|_2^2} \right),$$

which means that we can recover $\widehat{\mathbf{\Pi}}$ by

$$\widehat{\mathbf{\Pi}} \in \operatorname{argmax}_{\mathbf{\Pi}} \ \sum_{i=1}^{m} \langle \mathbf{Y}_{:,i},\ \mathbf{\Pi}\mathbf{X}\rangle^2. \tag{15}$$

Given that $\mathbf{B}^*_{:,i}$ is positive, we assume $\langle \mathbf{Y}_{:,i}, \mathbf{\Pi}\mathbf{X}\rangle \approx \mathbb{E}\langle \mathbf{Y}_{:,i}, \mathbf{\Pi}\mathbf{X}\rangle = \mathbb{E}\langle \mathbf{\Pi}^*\mathbf{X}, \mathbf{\Pi}\mathbf{X}\rangle \mathbf{B}^*_{:,i} > 0$, and relax (15) to

$$\sum_{i=1}^{m} |\langle \mathbf{Y}_{:,i},\ \mathbf{\Pi}\mathbf{X}\rangle|^2 \leq \left\langle \sum_{i=1}^{m} \mathbf{Y}_{:,i},\ \mathbf{\Pi}\mathbf{X} \right\rangle^2. \tag{16}$$

We then compute an estimator $\widehat{\mathbf{\Pi}}$ by maximizing the above upper bound in Eq. (16), which can be formulated as a linear assignment problem as in Eq. (7). The computational complexity of this estimator is $\Omega(m + n\log n)$, since only averaging and sorting are needed [22].

Algorithm 1 Averaging estimator.

- Compute the average $\frac{1}{m}\sum_{i=1}^{m} \mathbf{Y}_{:,i}$.
- Compute $\widehat{\mathbf{\Pi}}$ by maximizing $\left(\left\langle m^{-1}\sum_{i=1}^{m} \mathbf{Y}_{:,i}, \mathbf{\Pi}\mathbf{X}\right\rangle\right)^2$.

b) Eigenvalue estimator: We consider $\frac{1}{m}\sum_{i=1}^{m} \mathbf{Y}_{:,i}\mathbf{Y}_{:,i}^\top$, which can be expressed as

$$\frac{\sum_i \mathbf{Y}_{:,i}\mathbf{Y}_{:,i}^\top}{m} = \frac{\sum_{i=1}^{m} \left\|\mathbf{B}^*_{:,i}\right\|_{\mathrm{F}}^2}{m} \left(\mathbf{\Pi}^*\mathbf{X}\mathbf{X}^\top\mathbf{\Pi}^{*\top}\right) + \mathscr{T}_1 + \mathscr{T}_2 + \mathscr{T}_3,$$

where $\mathscr{T}_1$, $\mathscr{T}_2$, and $\mathscr{T}_3$ are defined as

$$\mathscr{T}_1 \triangleq \mathbf{\Pi}^*\mathbf{X}\left(\frac{\sum_{i=1}^{m} \mathbf{B}^*_{:,i}\mathbf{W}_{:,i}}{m}\right)^\top$$
$$\mathscr{T}_2 \triangleq \left(\frac{\sum_{i=1}^{m} \mathbf{B}^*_{:,i}\mathbf{W}_{:,i}}{m}\right)\mathbf{X}^\top\mathbf{\Pi}^{*\top}$$
$$\mathscr{T}_3 \triangleq \frac{\sum_{i=1}^{m} \mathbf{W}_{:,i}\mathbf{W}_{:,i}^\top}{m},$$

respectively. As $m \to \infty$, we have $\mathscr{T}_1, \mathscr{T}_2 \to 0$ and $\mathscr{T}_3 \to \sigma^2\mathbf{I}$. Hence we can approximate $m^{-1}\left(\sum_{i=1}^{m} \mathbf{Y}_{:,i}\mathbf{Y}_{:,i}^\top\right)$ as

$$\frac{\sum_{i=1}^{m} \mathbf{Y}_{:,i}\mathbf{Y}_{:,i}^\top}{m} \approx \frac{\sum_{i=1}^{m} \left\|\mathbf{B}^*_{:,i}\right\|_{\mathrm{F}}^2}{m}\left(\mathbf{\Pi}^*\mathbf{X}\mathbf{X}^\top\mathbf{\Pi}^{*\top}\right) + \frac{\sigma^2\mathbf{I}}{m}. \tag{17}$$

Since the principal eigenvector $\mathbf{u}$ of the matrix on the right-side of Eq. (17) is aligned with $\mathbf{\Pi}^*\mathbf{X}$, we can find $\widehat{\mathbf{\Pi}}$ by maximizing $\langle \mathbf{u},\ \mathbf{\Pi}\mathbf{X}\rangle^2$.

Algorithm 2 Eigenvalue estimator.

- Compute the principal eigenvector $\mathbf{u}$ of $m^{-1}\left(\sum_{i=1}^{m} \mathbf{Y}_{:,i}\mathbf{Y}_{:,i}^\top\right)$.
- Recover $\widehat{\mathbf{\Pi}}$ by maximizing $(\langle \mathbf{u},\ \mathbf{\Pi}\mathbf{X}\rangle)^2$.

C. ADMM algorithm for $p \geq 2$

In this subsection, we relax the ML estimation problem (5) to a bi-convex problem and solve it via an ADMM algorithm (cf. Alg. 3). A detailed derivation is given in the sequel.

a) ADMM formulation: First, in light of Eq. (6) we have

$$\min_{\mathbf{\Pi},\mathbf{B}} \ \|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}\|_{\mathrm{F}}^2 = \left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 \tag{18}$$

where $P^{\perp}_{\mathbf{\Pi}\mathbf{X}}$ is defined as $\mathbf{I} - \mathbf{\Pi}\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{\Pi}^\top$. Note that we can decompose $\mathbf{Y}$ as $P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y} + P_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}$. Since $\|\mathbf{Y}\|_{\mathrm{F}}^2 = \left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 + \|P_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\|_{\mathrm{F}}^2$ can be treated as a constant, minimizing $\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2$ is equivalent to maximizing $\|P_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\|_{\mathrm{F}}^2$.

By introducing two redundant variables $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$, we formulate Eq. (18) as

$$\min_{\mathbf{\Pi}_1,\ \mathbf{\Pi}_2} -\text{trace}\left(\mathbf{\Pi}_1 P_{\mathbf{X}} \mathbf{\Pi}_2^\top \mathbf{Y}\mathbf{Y}^\top\right), \quad \text{s.t. } \mathbf{\Pi}_1 = \mathbf{\Pi}_2, \tag{19}$$

where $P_{\mathbf{X}} \triangleq \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$. We propose to solve Eq. (19) with the ADMM algorithm [17] and present the details of the algorithm in Alg. 3.

Algorithm 3 ADMM algorithm for the recovery of $\mathbf{\Pi}$.

- **Input**: Initial estimate for the permutation matrix $\mathbf{\Pi}^{(0)}$.
- **For time** $t+1$**:** Update $\mathbf{\Pi}_1^{(t+1)}, \mathbf{\Pi}_2^{(t+1)}$ as

$$\mathbf{\Pi}_1^{(t+1)} = \text{argmin}_{\mathbf{\Pi}_1} \left\langle \mathbf{\Pi}_1, -\mathbf{Y}\mathbf{Y}^\top \mathbf{\Pi}_2^{(t)} P_{\mathbf{X}}^\top + \boldsymbol{\mu}^{(t)} - \rho \mathbf{\Pi}_2^{(t)} \right\rangle$$

$$\mathbf{\Pi}_2^{(t+1)} = \text{argmin}_{\mathbf{\Pi}_2} \left\langle \mathbf{\Pi}_2, \ \mathbf{Y}\mathbf{Y}^\top \mathbf{\Pi}_1^{(t+1)} P_{\mathbf{X}} - \boldsymbol{\mu}^{(t)} - \rho \mathbf{\Pi}_1^{(t+1)} \right\rangle$$

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \rho\left(\mathbf{\Pi}_1^{(t+1)} - \mathbf{\Pi}_2^{(t+1)}\right).$$

- **Termination**: Stop the ADMM algorithm once $\mathbf{\Pi}_1^{(t+1)}$ is identical to $\mathbf{\Pi}_2^{(t+1)}$.

b) Acceleration: Since ADMM may exhibit slow convergence [17], we adopt a *warm start* strategy to accelerate the algorithm, which consists of two steps:

- Compute the average value $\overline{\mathbf{X}} = \frac{1}{p}\sum_{i=1}^{p} \mathbf{X}_{:,i}$.
- Obtain a rough estimate $\mathbf{\Pi}^{(0)}$ by substituting $\mathbf{X}$ in Alg. 1 or Alg. 2 with $\overline{\mathbf{X}}$.

Then we choose $\mathbf{\Pi}^{(0)}$ as the starting point of the ADMM scheme in Alg. 3 to obtain an acceleration.

VI. Numerical Results

In this section, we present simulation results, which can be divided into two parts: 1) the oracle case (known $\mathbf{B}^*$) and 2) the realistic case with $\mathbf{B}^*$ being unknown. In virtue of Thm. 1, we here plot

$$\frac{\log\det(\mathbf{I} + \mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2)}{\log n} = \frac{\sum_i \log\left(1+\lambda_i^2/\sigma^2\right)}{\log n}$$

on the horizontal axis, and the empirical probability of permutation recovery on the vertical axis.

A. Oracle case

In this subsection, we study the relation between the error probability $\Pr\left(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*\right)$ and the snr in the oracle case with known $\mathbf{B}^*$, which is shown in Fig. 1. The simulation results confirm our theoretical results in Thm. 1 and Prop. 2.

a) Rank-1 case: In this case, we assume that all columns $\{\mathbf{B}^*_{:,i}\}_{i=1}^m$ of $\mathbf{B}^*$ are identical.

b) Full-rank case: In this case, we consider the case $\mathbf{B}^*_{:,i} \perp \mathbf{B}^*_{:,j}$, $1 \le i \neq j \le m$. For simplicity, we set $\mathbf{B}^*_{:,i} \parallel \mathbf{e}_i$, where $\{\mathbf{e}_i\}$ denotes the canonical basis.

B. Comparison between different estimators

In this subsection, we compare the performance of four different estimators: the *averaging estimator* (Alg. 1), the *eigenvalue estimator* (Alg. 2), averaging ADMM (ADMM estimator with averaging estimator as warm start, Alg. 3), and eigenvalue ADMM (ADMM estimator with eigenvalue estimator as warm start, Alg. 3).

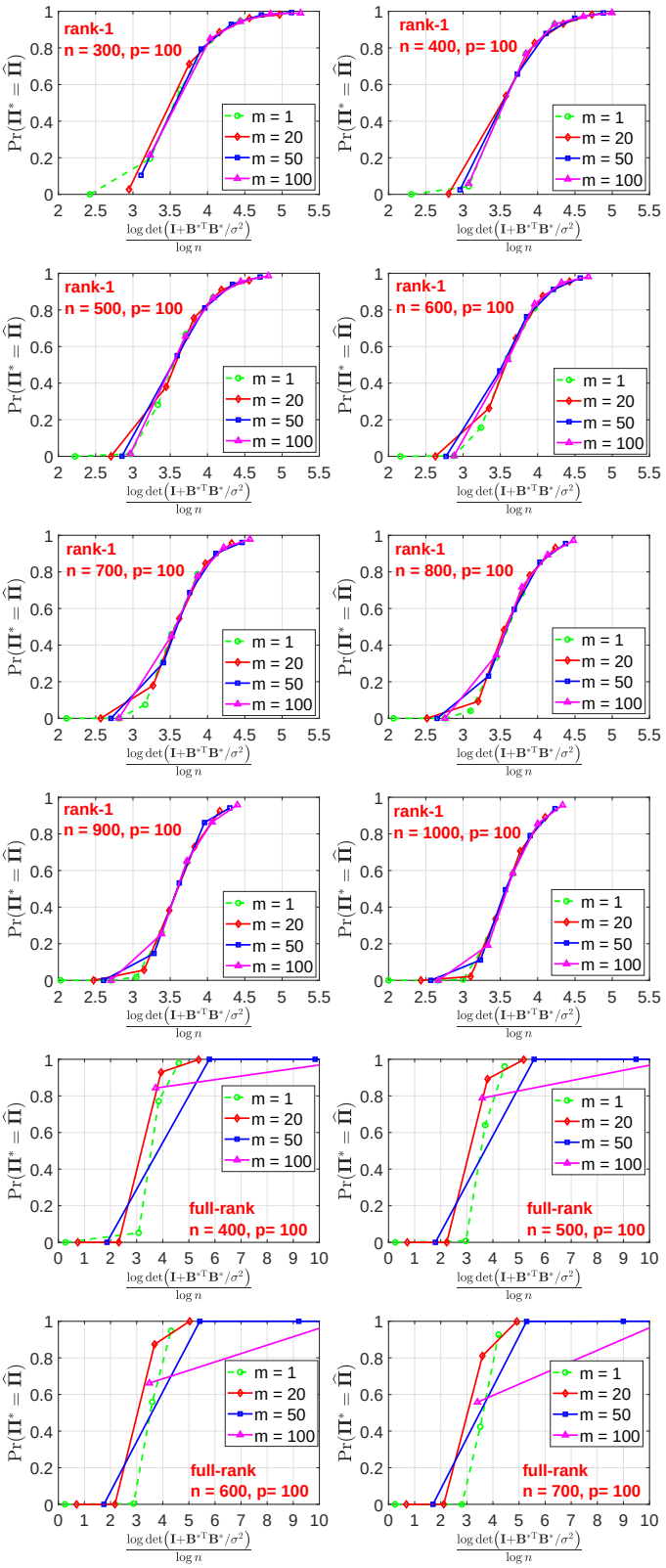

Fig. 1: Oracle case: relation between correct probability $\Pr\left(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\right)$ and $\frac{\log\det(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2)}{\log n}$ for different n values.

a) One-dimensional case ($p = 1$): The simulation result is shown in Fig. 2. The empirical study suggests that both averaging ADMM and eigenvalue ADMM converge in one step in most cases, which suggests that the rough initial

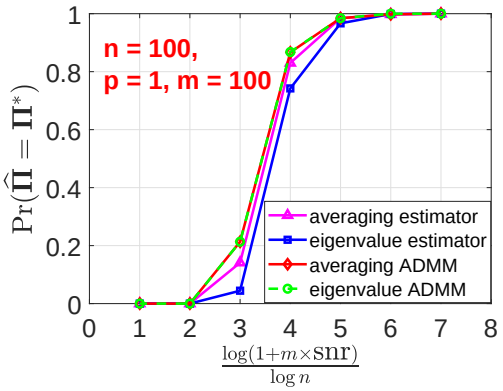

Fig. 2: Relation between correct probability $\Pr\left(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\right)$ and $\frac{\log(1+m\times\mathrm{snr})}{\log n}$ for four estimators ($n = 100$, $p = 1$, $m = 100$).

estimate is already a local optimum. We also observe that:

- The averaging estimator performs better than the eigenvalue estimator.
- Averaging ADMM exhibits a similar performance as eigenvalue ADMM when $p = 1$. As $p \geq 2$, averaging ADMM outperforms eigenvalue ADMM as shown in Fig. 3.
- The relative frequency of the event $\{\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\}$ becomes positive when $\log(1 + m \times \mathrm{snr}) \geq 2$, which is much larger than that shown in Fig. 3 and Fig. 4. One potential reason is that our estimator can only find a spurious local optimum. In the low-dimensional case ($p = 1$), ADMM is more likely to get trapped in such bad local optima, and hence a higher snr is required to reach the global optimum.

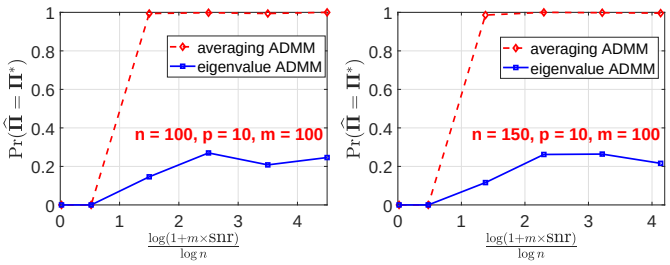

Fig. 3: Relation between correct probability $\Pr\left(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\right)$ and $\frac{\log(1+m\times\mathrm{snr})}{\log n}$ for different estimators ($p = 10$ and $m = 100$).

b) Dimension larger than one ($p \geq 2$)*:* The simulation results are shown in Fig. 3. We notice that averaging ADMM outperforms eigenvalue ADMM. A potential reason for this observation is that the eigenvalue estimator heavily relies on the fact that $p = 1$. When $p \geq 2$, the principal eigenvector may not align with the direction of $\mathbf{\Pi}^*\mathbf{X}$, which implies a poor start and hence a slow convergence rate.

C. Realistic case

In Fig. 4, we consider averaging ADMM and study the relation between the relative frequency of the event $\{\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\}$ and snr for different n. We observe a transition in the region $[0.5,\ 1.5]$, which is consistent with our results in Thm. 5 which asserts that $\log(m \times \mathrm{snr}) \sim \log(n)$ is needed to ensure successful recovery. However, it needs to be emphasized that $\min_{\mathbf{\Pi},\mathbf{B}} \|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}\|_{\mathrm{F}}^2$ is NP-hard [6] when $p \geq 1$ and hence the computed solution may not be the global optimum. Less energy, i.e., $\varrho(\mathbf{B}^*)\log(1+\mathrm{snr}) \sim \log n$, may be sufficient to achieve correct recovery.

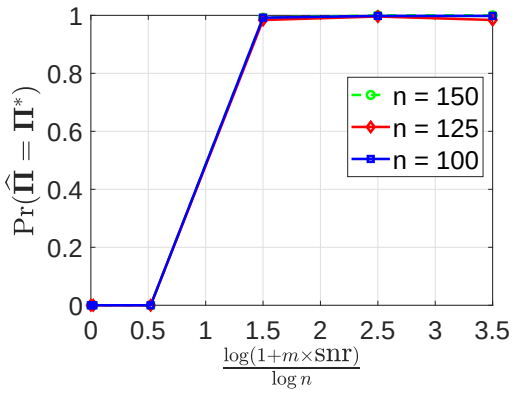

Fig. 4: Realistic case: relation between correct probability $\Pr\left(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\right)$ and the signal snr for different n values ($m = 100$ and $p = 10$).

VII. Conclusion

In this paper, we have studied the unlabeled sensing problem given multiple measurement vectors. First, we established the statistical limits in terms of conditions on the snr implying failure of recovery with high probability, namely, $\varrho(\mathbf{B}^*)\log(\mathrm{snr}) \lesssim \log(n)$. The tightness of these conditions is consolidated by the corresponding condition for correct recovery with $\mathbf{B}^*$ being known. Without knowledge of $\mathbf{B}^*$, we needed $\log(m \times \mathrm{snr}) \gtrsim \log n$ for correct recovery, which matches the lower bound for the oracle case with $\varrho(\mathbf{B}^*) = 1$. By imposing the additional assumption $d_H(\mathbf{I}, \mathbf{\Pi}^*) \leq h_{\max}$, it can be proved that $\varrho(\mathbf{B}^*)\log(\mathrm{snr}) \gtrsim \log(n)$ is sufficient for correct recovery. Moreover, we proposed a computational framework based on ADMM to tackle the computational difficulties associated with the computation of the ML estimator. Simulation results largely corroborated our theoretical findings with the exception of gaps that are likely attributable to the fact that ADMM may deliver spurious local optima with potentially different statistical properties compared to the global optimum that is the object of our theoretical analysis. In future work, we aim to bridge this gap by designing improved computational schemes, e.g., based on more reliable initialization procedures to avoid spurious local optima.

References

[1] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. Academic Press, 1979.

[2] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.

[3] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, "Unlabeled Sensing: solving a linear system with unordered measurements," in *Communication, Control, and Computing (Allerton)*, 2015, pp. 786–793.

[4] I. Domankic, "Permutations unlabeled beyond sampling unknown," December 2018, arXiv:1812.00498.

[5] M. Tsakiris, "Eigenspace conditions for homomorphic sensing," December 2018, arXiv:1812.07966.

[6] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Linear regression with an unknown permutation: Statistical and computational limits," in *Communication, Control, and Computing (Allerton)*, 2016, pp. 417–424.

[7] D. J. Hsu, K. Shi, and X. Sun, "Linear regression without correspondence," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 1530–1539.

[8] A. Abid, A. Poon, and J. Zou, "Linear regression with shuffled labels," May 2017, arXiv:1705.01342.

[9] M. Slawski and E. Ben-David, "Linear Regression with Sparsely Permuted Data," *Electronic Journal of Statistics*, vol. 1, pp. 1–36, 2019.

[10] M. Slawski, E. Ben-David, and P. Li, "A Two-Stage Approach to Multivariate Linear Regression with Sparsely Mismatched Data," July 2019, arXiv:1907.07148.

[11] M. Slawski, M. Rahmani, and P. Li, "A Robust Subspace Recovery Approach to Linear Regression with Partially Shuffled Labels," May 2019, to appear in Uncertainty in Artificial Intelligence (UAI).

[12] X. Shi, X. Lu, and T. Cai, "Spherical regresion under mismatch corruption with application to automated knowledge translation," October 2018, arXiv:1810.05679.

[13] S. Haghighatshoar and G. Caire, "Signal recovery from unlabeled samples," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 451–455.

[14] M. Tsakiris and L. Peng, "Homomorphic sensing," in *International Conference on Machine Learning (ICML)*, 2019, pp. 6335–6344.

[15] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval, "Compressed sensing with unknown sensor permutation," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1040–1044.

[16] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Denoising linear models with permuted data," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 446–450.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[18] R. Burkard, M. Dell'Amico, and S. Martello, *Assignment problems, revised reprint*. SIAM, 2012, vol. 106.

[19] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[20] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[21] D. P. Bertsekas and D. A. Castanon, "A forward/reverse auction algorithm for asymmetric assignment problems," *Computational Optimization and Applications*, vol. 1, no. 3, pp. 277–297, 1992.

[22] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2009.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[24] R. A. Horn, R. A. Horn, and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[25] S. Boucheron and M. Thomas, "Concentration inequalities for order statistics," *Electronic Communications in Probability*, vol. 17, pp. 1–12, 2012.

[26] M. Rudelson, R. Vershynin *et al.*, "Hanson-Wright inequality and sub-Gaussian concentration," *Electronic Communications in Probability*, vol. 18, 2013.

[27] R. Latala, P. Mankiewicz, K. Oleszkiewicz, and N. Tomczak-Jaegermann, "Banach-mazur distances and projections on random subgaussian polytopes," *Discrete & Computational Geometry*, vol. 38, no. 1, pp. 29–50, 2007.

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

[29] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

APPENDIX A
NOTATIONS

For an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we denote by $\mathbf{A}_{:,i} \in \mathbb{R}^n$ the i-th column of $\mathbf{A}$ while $\mathbf{A}_{i,:} \in \mathbb{R}^m$ denotes the i-th row, treated as column vector. Moreover, A_{ij} denotes the (i,j)-th element of the matrix $\mathbf{A}$. The pseudo-inverse $\mathbf{A}^\dagger$ of the matrix $\mathbf{A}$ is defined as $\left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$. We define $P_{\mathbf{A}} = \mathbf{A}\mathbf{A}^\dagger$ as the projection onto the column space of $\mathbf{A}$, while $P_{\mathbf{A}}^\perp = \mathbf{I} - P_{\mathbf{A}}$ denotes the projection onto its orthogonal complement. The *singular value decomposition* (SVD) of the matrix $\mathbf{A}$ [19] (Section 2.4, P76) is represented by $\mathrm{SVD}(\mathbf{A})$, such that $\mathrm{SVD}(\mathbf{A}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{m\times m}$, $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$, and $\mathbf{V} \in \mathbb{R}^{n\times n}$, where $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}_{m\times m}$, $\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_{n\times n}$. The operator $\mathrm{vec}(\mathbf{A})$ denotes the vectorization of $\mathbf{A}$ that is obtained by concatenating the columns of $\mathbf{A}$ into a vector.

We write $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm while $\|\cdot\|_{\mathrm{OP}}$ is used for the operator norm, whose definitions can be found in [19] (Section 2.3, P71). The ratio $\varrho(\cdot) = \|\cdot\|_{\mathrm{F}}^2/\|\cdot\|_{\mathrm{OP}}^2$ represents the stable rank while $r(\cdot)$ represents the usual rank of a matrix.

We write $\pi(\cdot)$ for a permutation of $\{1, 2, \cdots, n\}$ that moves index i to $\pi(i)$, $1 \le i \le n$. The corresponding permutation matrix is denoted by Π. We use $d_H(\cdot,\cdot)$ to denote the Hamming distance between two permutation matrices, *i.e.*, $d_H(\Pi_1, \Pi_2) = \sum_{i=1}^n \mathbb{1}\left(\pi_1(i) \neq \pi_2(i)\right)$.

For an event $\mathscr{E}$, we denote its complement by $\overline{\mathscr{E}}$. In addition, we use $a \vee b$ to denote the maximum of a and b while $a \wedge b$ to denote the minimum of a and b.

APPENDIX B
PROOF OF THM. 1

Proof. Without loss of generality, we assume that $\mathbf{B}^*$ is known. Note that if we cannot recover $\mathbf{\Pi}^*$ even when $\mathbf{B}^*$ is known, it is hopeless to recover $\mathbf{\Pi}^*$ with unknown $\mathbf{B}^*$. We can reformulate the sensing relation Eq. (4)

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

as the following transmission process

$$\mathbf{\Pi}^* \overset{(i)}{\to} \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^* \overset{(ii)}{\to} \underbrace{\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^* + \mathbf{W}}_{\mathbf{Y}}, \tag{20}$$

where in (i) the signal $\mathbf{\Pi}^*$ is encoded to the code word $\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*_{:,i}$, and in (ii) the n codewords $\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*_{:,i}$ are transmitted through n i.i.d Gaussian channels. With this reformulation, we can treat the recovery of $\mathbf{\Pi}^*$ as a decoding problem. Denote the recovered permutation matrix as $\widehat{\mathbf{\Pi}}$. Following a similar approach as in [23] (Section 7.9, P206), we have

$$\begin{aligned}
H(\mathbf{\Pi}^*) &\overset{(a)}{=} H(\mathbf{\Pi}^* \mid \mathbf{X}) \\
&\overset{(b)}{=} H(\mathbf{\Pi}^* \mid \widehat{\mathbf{\Pi}}, \mathbf{X}) + I(\mathbf{\Pi}^*;\ \widehat{\mathbf{\Pi}} \mid \mathbf{X}) \\
&\overset{(c)}{\le} H(\mathbf{\Pi}^* \mid \widehat{\mathbf{\Pi}}) + I(\mathbf{\Pi}^*;\ \widehat{\mathbf{\Pi}} \mid \mathbf{X}) \\
&\overset{(d)}{\le} 1 + \log\left(|\mathscr{H}|\right)\Pr(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*) + I(\mathbf{\Pi}^*;\ \widehat{\mathbf{\Pi}} \mid \mathbf{X}) \\
&\overset{(e)}{\le} 1 + \log\left(|\mathscr{H}|\right)\Pr(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*) + I(\mathbf{\Pi}^*;\ \mathbf{Y} \mid \mathbf{X}) \\
&\overset{(f)}{\le} 1 + \log\left(|\mathscr{H}|\right)\Pr(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*) + \frac{n}{2}\sum_i \log\left(1 + \frac{\lambda_i^2}{\sigma^2}\right),
\end{aligned}$$

where λ_i denote the i-th singular value of $\mathbf{B}^*$, in (a) we use the fact that $\mathbf{X}$ and $\mathbf{\Pi}^*$ are independent, in (b) we use the definition of the conditional mutual information $I(\mathbf{\Pi}^*; \widehat{\mathbf{\Pi}} \mid \mathbf{X})$, in (c) we use $H(\mathbf{\Pi}^* \mid \widehat{\mathbf{\Pi}}, \mathbf{X}) \le H(\mathbf{\Pi}^* \mid \widehat{\mathbf{\Pi}})$, in (d) we use Fano's inequality, in (e) we use the data-processing inequality, noting that $\mathbf{\Pi}^* \to \mathbf{\Pi}^*\mathbf{X}\mathbf{B} \to \mathbf{Y}$ forms a Markov chain [23], and in (f) we use Lemma 7 to upper bound the

conditional mutual information $I(\mathbf{\Pi}^*;\ \mathbf{Y}\mid\mathbf{X})$. Background on the tools used in steps (b)-(e) is provided in Appendix I for the convenience of the reader.

We thus obtain the following lower bound on $\Pr(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*)$

$$\Pr(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*)\geq\frac{H(\mathbf{\Pi}^*)-1-(n/2)\sum_i\log\left(1+\lambda_i^2/\sigma^2\right)}{\log\left(|\mathscr{H}|\right)},$$

which is bounded below by $1/2$ provided

$$H(\mathbf{\Pi}^*)>1+\frac{n}{2}\sum_i\log\left(1+\frac{\lambda_i^2}{\sigma^2}\right)+\frac{\log\left(|\mathscr{H}|\right)}{2},$$

and complete the proof. □

Lemma 7. *For the channel described in Eq. (20), we have*

$$I(\mathbf{\Pi}^*;\ \mathbf{Y}_{:,1},\mathbf{Y}_{:,2},\ldots,\mathbf{Y}_{:,m}|\ \mathbf{X})\leq\frac{n}{2}\sum_i\log\left(1+\frac{\lambda_i^2}{\sigma^2}\right),$$

where λ_i denotes the i-th singular value of $\mathbf{B}^*$.

Proof. Let $\mathrm{vec}\left(\mathbf{Y}\right)$ ($\mathrm{vec}\left(\mathbf{W}\right)$) be the vector formed by concatenating $\mathbf{Y}_{:,1},\mathbf{Y}_{:,2},\cdots,\mathbf{Y}_{:,m}$ ($\mathbf{W}_{:,1},\mathbf{W}_{:,2},\cdots,\mathbf{W}_{:,m}$), according to the definition in Appendix A. For simplicity of notation, we use $I\left(\mathbf{\Pi}^*;\ \mathrm{vec}\left(\mathbf{Y}\right)|\mathbf{X}\right)$ as a shortcut for $I\left(\mathbf{\Pi}^*;\ \mathbf{Y}_{:,1},\mathbf{Y}_{:,2},\ldots,\mathbf{Y}_{:,m}|\ \mathbf{X}\right)$. We then calculate the conditional mutual information $I\left(\mathbf{\Pi}^*;\ \mathrm{vec}\left(\mathbf{Y}\right)|\mathbf{X}\right)$ as

$$\begin{aligned}
&I\left(\mathbf{\Pi}^*;\mathrm{vec}\left(\mathbf{Y}\right)|\mathbf{X}\right)\overset{(i)}{=}h(\mathrm{vec}(\mathbf{Y})|\mathbf{X})-h(\mathrm{vec}(\mathbf{Y})|\mathbf{X},\mathbf{\Pi}^*)\\
&\overset{(ii)}{=}\ \mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}h\left(\mathrm{vec}(\mathbf{Y})|\mathbf{X}=\mathbf{x}\right)-h(\mathrm{vec}(\mathbf{W}))\\
&\overset{(iii)}{=}\ \mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}h\left(\mathrm{vec}(\mathbf{Y})|\mathbf{X}=\mathbf{x}\right)-\frac{mn}{2}\log\sigma^2\\
&\overset{(iv)}{\leq}\ \mathbb{E}_{\mathbf{X}}\frac{1}{2}\log\det\left(\mathbb{E}_{\mathbf{\Pi}^*,\mathbf{W}|\mathbf{X}=\mathbf{x}}\mathrm{vec}\left(\mathbf{Y}\right)\mathrm{vec}\left(\mathbf{Y}\right)^\top\right)-\frac{mn}{2}\log\sigma^2,\\
&\overset{(v)}{\leq}\ \frac{1}{2}\log\det\mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}\mathrm{vec}\left(\mathbf{Y}\right)\mathrm{vec}\left(\mathbf{Y}\right)^\top-\frac{mn}{2}\log\sigma^2,\qquad(21)
\end{aligned}$$

where in (i) we use the definition of the conditional mutual information $I(\mathbf{\Pi}^*;\ \mathrm{vec}\left(\mathbf{Y}\right)|\ \mathbf{X})$, in (ii) we have used that

$$\begin{aligned}
h(\mathrm{vec}\left(\mathbf{Y}\right)|\ \mathbf{X},\mathbf{\Pi}^*)&=h(\mathrm{vec}\left(\mathbf{\Pi}^*\mathbf{X}\mathbf{B}+\mathbf{W}\right)|\mathbf{X},\mathbf{\Pi}^*)\\
&=h(\mathrm{vec}\left(\mathbf{W}\right)|\mathbf{X},\mathbf{\Pi}^*),
\end{aligned}$$

in (iii) we use that the mn entries of $\mathrm{vec}(\mathbf{W})$ are i.i.d Gaussian distributed with entropy is $\frac{1}{2}\log(\sigma^2)$ each, in (iv) we use a result in [23] (Thm 8.6.5, P254) which yields

$$h(\mathbf{Z})\leq\frac{1}{2}\log\det\mathrm{cov}(\mathbf{Z})\leq\frac{1}{2}\log\det\mathbb{E}[\mathbf{Z}\mathbf{Z}^\top],$$

where $\mathbf{Z}$ is an arbitrary RV with finite covariance matrix $\mathrm{cov}(\mathbf{Z})$, and we use the concavity: $\mathbb{E}\log\det(\cdot)\leq\log\det\mathbb{E}(\cdot)$.

In the sequel, we compute the entries of the matrix $\mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}\mathrm{vec}\left(\mathbf{Y}\right)\mathrm{vec}\left(\mathbf{Y}\right)^\top$. For simplicity of notation, the latter matrix will henceforth be denoted by $\mathbf{\Sigma}$. First note that $\mathrm{vec}\left(\mathbf{Y}\right)$ equals the concatenation of $\mathbf{Y}_{:,1},\mathbf{Y}_{:,2},\cdots,\mathbf{Y}_{:,m}$. We decompose the matrix $\mathbf{\Sigma}$ into sub-matrices $\mathbf{\Sigma}_{i_1,i_2}=\mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}\mathbf{Y}_{:,i_1}\mathbf{Y}_{:,i_2}^\top$, $1\leq i_1,i_2\leq m$, which corresponds to the covariance matrix between $\mathbf{Y}_{:,i_1}$ and $\mathbf{Y}_{:,i_2}$. The (j_1,j_2)-th element of sub-matrix Σ_{i_1,i_2} is defined as Σ_{i_1,i_2,j_1,j_2}. The latter can be expressed as

$$\begin{aligned}
&\Sigma_{i_1,i_2,j_1,j_2}=\mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}\left(Y_{j_1,i_1}Y_{j_2,i_2}\right)\\
&=\ \mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}\left[\left(\mathbf{X}\mathbf{B}^*_{:,i_1}\right)_{\pi^*(j_1)}+W_{j_1,i_1}\right]\\
&\quad\times\left[\left(\mathbf{X}\mathbf{B}^*_{:,i_2}\right)_{\pi^*(j_2)}+W_{j_2,i_2}\right]\\
&=\ \mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X}}\left(\left\langle X_{\pi^*(j_1),:},\mathbf{B}^*_{:,i_1}\right\rangle\left\langle\mathbf{X}_{\pi^*(j_2),:},\mathbf{B}^*_{:,i_2}\right\rangle\right)\\
&\quad+\mathbb{E}_{\mathbf{W}}W_{j_1,i_1}W_{j_2,i_2},
\end{aligned}$$

where π^* is the permutation corresponding to the permutation matrix $\mathbf{\Pi}^*$ as defined in Appendix A. We then split the calculation into four sub-cases:

$$\begin{cases}
\textbf{Case } i_1=i_2,j_1=j_2\textbf{:} & \Sigma_{i_1,\ i_1,\ j_1,\ j_1}=\left\|\mathbf{B}^*_{:,i_1}\right\|_2^2+\sigma^2.\\
\textbf{Case } i_1\neq i_2,j_1=j_2\textbf{:} & \Sigma_{i_1,i_2,j_1,j_1}=\left\langle\mathbf{B}^*_{:,i_1},\mathbf{B}^*_{:,i_2}\right\rangle.\\
\textbf{Case } j_1\neq j_2\textbf{:} & \Sigma_{i_1,i_2,j_1,j_2}=0.
\end{cases}$$

In conclusion, the matrix $\mathbf{\Sigma}$ can be expressed as

$$\mathbf{\Sigma}=\underbrace{\begin{bmatrix}
\left\|\mathbf{B}^*_{:,1}\right\|_2^2+\sigma^2 & \left\langle\mathbf{B}^*_{:,1},\mathbf{B}^*_{:,2}\right\rangle & \cdots & \left\langle\mathbf{B}^*_{:,1},\mathbf{B}^*_{:,m}\right\rangle\\
\left\langle\mathbf{B}^*_{:,2},\mathbf{B}^*_{:,1}\right\rangle & \left\|\mathbf{B}^*_{:,2}\right\|_2^2+\sigma^2 & \cdots & \left\langle\mathbf{B}^*_{:,2},\mathbf{B}^*_{:,m}\right\rangle\\
\vdots & & \ddots & \vdots\\
\left\langle\mathbf{B}^*_{:,m},\mathbf{B}^*_{:,1}\right\rangle & \left\langle\mathbf{B}^*_{:,m},\mathbf{B}^*_{:,2}\right\rangle & \cdots & \left\|\mathbf{B}^*_{:,m}\right\|_2^2+\sigma^2
\end{bmatrix}}_{\mathbf{\Sigma}_1}\otimes\mathbf{I}_{n\times n}$$

where $\otimes$ denotes the Kronecker product [19] (Section 1.3.6, P27). According to [19] (Section 12.3.1, P709), we have

$$\begin{aligned}
\det\left(\mathbf{\Sigma}\right)&=\left(\det\left(\mathbf{\Sigma}_1\right)\right)^n\ \left(\det\left(\mathbf{I}_{n\times n}\right)\right)^m\\
&\overset{(a)}{=}\sigma^{2nm}\left(\det\left(\mathbf{I}+\frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right)\right)^n,\qquad(22)
\end{aligned}$$

where in (a) we have calculated $\det(\mathbf{\Sigma}_1)$ as

$$\det\left(\mathbf{\Sigma}_1\right)=\det\left(\sigma^2\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*\right)=\sigma^{2m}\det\left(\mathbf{I}+\frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right).$$

By combining Eq. (21) and Eq. (22), we have obtained the upper bound

$$\begin{aligned}
I\left(\mathbf{\Pi}^*;\ \mathrm{vec}\left(\mathbf{Y}\right)|\mathbf{X}\right)&\leq\ \frac{n}{2}\log\det\left(\mathbf{I}+\frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right)\\
&\overset{(b)}{=}\ \sum_i\log\left(1+\frac{\lambda_i^2}{\sigma^2}\right),
\end{aligned}$$

where (b) can be verified via the singular value decomposition $\mathrm{SVD}\left(\mathbf{B}^*\right)=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ as introduced in Appendix A) and by using basic properties of the matrix determinant [24] (Sec. 0.3, P8). □

Appendix C
Proof of Prop. 2

Proof. Observe that the sensing relation $\mathbf{Y}=\mathbf{\Pi}^*\mathbf{X}\mathbf{B}+\mathbf{W}$ is equivalent to $\mathbf{\Pi}^{*\top}\mathbf{Y}=\mathbf{X}\mathbf{B}+\mathbf{\Pi}^{*\top}\mathbf{W}$. As a consequence of the rotational invariance of the Gaussian distribution, $\mathbf{\Pi}^{*\top}\mathbf{W}$ follows the same distribution as $\mathbf{W}$. Since our proof applies to any instance of the permutation matrix $\mathbf{\Pi}^*$, we may assume $\mathbf{\Pi}^*=\mathbf{I}$ w.l.o.g. We begin the proof by first presenting the roadmap.

- **Stage I:** Define $\widetilde{W}_{i,j}$ as

$$\widetilde{W}_{i,j} = \left\langle \mathbf{W}_{j,:} - \mathbf{W}_{i,:},\ \frac{\mathbf{B}^{*\top}(\mathbf{X}_{i,:} - \mathbf{X}_{j,:})}{\|\mathbf{B}^{*\top}(\mathbf{X}_{i,:} - \mathbf{X}_{j,:})\|_2} \right\rangle,$$

for $1 \le i < j \le n$, we would like to prove that

$$\left\{ \exists\ (i,j), \text{s.t. } \widetilde{W}_{i,j} \ge \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2 \right\} \subseteq \left\{ \widehat{\mathbf{\Pi}} \ne \mathbf{I} \right\}.$$

We then lower bound the probability $\Pr\left(\widehat{\mathbf{\Pi}} \ne \mathbf{I}\right)$ as

$$\Pr\left(\widehat{\mathbf{\Pi}} \ne \mathbf{I}\right) \ge \Pr\left(\exists\ (i,j), \text{s.t. } \widetilde{W}_{i,j} \ge \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2\right).$$

- **Stage II:** We lower bound the probability $\Pr\left(\exists\ (i,j), \text{s.t. } \widetilde{W}_{i,j} \ge \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2\right)$ by two separate probabilities, namely

$$\begin{aligned} &\Pr\left(\exists\ (i,j), \text{s.t. } \widetilde{W}_{i,j} \ge \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2\right) \\ \ge\ &\Pr\left(\widetilde{W}_{1,j_0} \ge \rho_0\right) \Pr\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{1,:} - \mathbf{X}_{j_0,:}\right)\right\|_2 \le \rho_0\right), \end{aligned}$$

where j_0 is picked as $\text{argmax}_j \widetilde{W}_{1,j}$, and ρ_0 is one positive parameter waiting to be set.

- **Stage III**: Provided Condition (9) holds, we are allowed to set $\rho_0 = 2\sqrt{2\sigma^2 \log(n)}$ without violating the requirement of Lemma 9. We thereby conclude the proof by setting $\rho_0 = 2\sqrt{2\sigma^2 \log(n)}$ and invoking Lemma 8 and Lemma 9.

Detailed calculation comes as follows.

Stage I: We conclude the proof by showing if $\left\{\widetilde{W}_{i,j} \ge \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2\right\}$ holds, we would have

$$\begin{aligned} &\left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2^2 + \left\|\mathbf{Y}_{j,:} - \mathbf{B}^{*\top}\mathbf{X}_{i,:}\right\|_2^2 \\ \le\ &\left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{i,:}\right\|_2^2 + \left\|\mathbf{Y}_{j,:} - \mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2^2, \end{aligned}$$

which implies that $\min_{\mathbf{\Pi}} \|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}^2 \le \|\mathbf{Y} - \mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}^2$ since $\mathbf{\Pi}$ can be chosen as the transposition that swaps $\mathbf{Y}_{i,:}$ and $\mathbf{Y}_{j,:}$. This implies failure of recovery, i.e., the event $\{\widehat{\mathbf{\Pi}} \ne \mathbf{I}\}$.

Stage II: We lower bound the error probability $\Pr\left(\widehat{\mathbf{\Pi}} \ne \mathbf{I}\right)$ as

$$\begin{aligned} &\Pr\left(\widehat{\mathbf{\Pi}} \ne \mathbf{I}\right) \\ \ge\ &\Pr\left(\exists\ (i,j), \text{s.t. } \widetilde{W}_{i,j} \ge \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2\right) \\ \overset{(i)}{\ge}\ &\Pr\left(\widetilde{W}_{1,j_0} \ge \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{1,:} - \mathbf{X}_{j_0,:}\right)\right\|_2\right) \\ \ge\ &\Pr\left(\widetilde{W}_{1,j_0} \ge \rho_0 \ \middle|\ \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{1,:} - \mathbf{X}_{j_0,:}\right)\right\|_2 \le \rho_0\right) \\ &\quad \times \Pr\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{1,:} - \mathbf{X}_{j_0,:}\right)\right\|_2 \le \rho_0\right) \\ \overset{(ii)}{=}\ &\Pr\left(\widetilde{W}_{1,j_0} \ge \rho_0\right) \Pr\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{1,:} - \mathbf{X}_{j_0,:}\right)\right\|_2 \le \rho_0\right), \end{aligned}$$

where in (i) we pick j_0 as $\text{argmax}_j \widetilde{W}_{1,j}$ and in (ii) we use the independence between $\widetilde{W}_{i,j}$ and $\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2$. □

Lemma 8. *When n is large ($n \ge 10$), we have*

$$\Pr\left(\sup_j \widetilde{W}_{1,j} \ge 2\sqrt{2\sigma^2 \log(n)}\right) \ge 1 - n^{-1}.$$

Proof. This result is quite standard and can be easily proved by combining Sec. 2.5 (P31) and Thm 5.6 (P126) in [25]. We omit the details for the sake of brevity. □

Lemma 9. *Given that $\rho_0 \ge 2\left(1 + 2\sqrt{\frac{\log 2}{c_1 \varrho(\mathbf{B}^*)}}\right)\|\mathbf{B}^*\|_{\mathrm{F}}$, we have*

$$\Pr\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2 \le \rho_0\right) \ge \frac{49}{64},$$

where $c_1 > 0$ is some constant, and $\varrho(\mathbf{B}^*)$ is the stable rank of the matrix $\mathbf{B}^*$.

Proof. First we define $\mathscr{A}_{\rho_0}^{(i,j)}$ and the set $\mathbb{B}$ as

$$\begin{aligned} \mathscr{A}_{\rho_0}^{(i,j)} &\triangleq \left\{\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2 \le \rho_0\right\},\ 1 \le i < j \le n; \\ \mathbb{B} &\triangleq \left\{\mathbf{x} \middle| \left\|\mathbf{B}^{*\top}\mathbf{x}\right\|_2 \le \frac{\rho_0}{2}\right\}. \end{aligned}$$

We then define a RV u_i for each $\mathbf{X}_{i,:}$ via $u_i = \mathbb{1}(\mathbf{X}_{i,:} \in \mathbb{B})$. It is not hard to verify that $\{u_i = 1,\ u_j = 1\} \subseteq \mathscr{A}_{\rho_0}^{(i,j)}$ because

$$\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2 \le \left\|\mathbf{B}^{*\top}\mathbf{X}_{i,:}\right\|_2 + \left\|\mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2 \le \rho_0.$$

We hence have the relation

$$\begin{aligned} \Pr\left(\mathscr{A}_{\rho_0}^{(i,j)}\right) &\ge \Pr\left(u_i = 1, u_j = 1\right) \\ &\overset{(i)}{=} \Pr\left(u_i = 1\right)\Pr\left(u_j = 1\right) = \zeta^2, \end{aligned} \tag{23}$$

where ζ is defined as $\Pr(u_k = 1)$, $k \in \{1, \ldots, n\}$, is an arbitrary integer, and (i) is because of the independence between $\mathbf{X}_{i,:}$ and $\mathbf{X}_{j,:}$. It thus remains to lower bound the probability $\Pr\left(\left\|\mathbf{B}^{*\top}\mathbf{x}\right\|_2 \le \rho_0/2\right)$. Setting ρ_0 such that $\rho_0/2 \ge (1+t)\|\mathbf{B}^*\|_{\mathrm{F}}$, we need to establish that

$$\Pr\left(\left\|\mathbf{B}^{*\top}\mathbf{x}\right\|_2 \le \frac{\rho_0}{2}\right) \ge \Pr\left(\left\|\mathbf{B}^{*\top}\mathbf{x}\right\|_2 \le (1+t)\|\mathbf{B}^*\|_{\mathrm{F}}\right). \tag{24}$$

According to Thm 2.1 in [26], we have

$$\begin{aligned} &\Pr\left(\left\|\mathbf{B}^{*\top}\mathbf{x}\right\|_2 \ge (1+t)\|\mathbf{B}^*\|_{\mathrm{F}}\right) \\ \le\ &\Pr\left(\left|\left\|\mathbf{B}^{*\top}\mathbf{x}\right\|_2 - \|\mathbf{B}^*\|_{\mathrm{F}}\right| \ge t\|\mathbf{B}^*\|_{\mathrm{F}}\right) \\ \le\ &2e^{-c_1 t^2 \varrho(\mathbf{B}^*)} \quad \forall t \ge 0. \end{aligned}$$

Hence, we have

$$\begin{aligned} \zeta = \Pr\left(\left\|\mathbf{B}^{*\top}\mathbf{x}\right\|_2 \le \frac{\rho_0}{2}\right) &\ge \Pr\left(\left\|\mathbf{B}^{*\top}\mathbf{x}\right\|_2 \le (1+t)\|\mathbf{B}^*\|_{\mathrm{F}}\right) \\ &\ge 1 - 2e^{-c_1 t^2 \varrho(\mathbf{B}^*)}. \end{aligned}$$

Setting $t = 2\sqrt{\frac{\log 2}{c_1 \varrho(\mathbf{B}^*)}}$, we have $\zeta \ge 7/8$, which implies $\Pr\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2 \le \rho_0\right) \ge \zeta^2 \ge 49/64$ in view of Eq. (23) and completes the proof. □

Appendix D
Proof of Corol. 3

Proof. First we define $\mathscr{E} = \mathbb{1}\left\{d_H(\widehat{\mathbf{\Pi}}, \mathbf{\Pi}^*) \ge D\right\}$, which corresponds to the failure in obtaining an approximation of $\mathbf{\Pi}^*$ within a Hamming ball of radius D. Moreover, we suppose that $\mathbf{\Pi}^*$ follows a uniform distribution over the set of all $n!$ possible permutation matrices.

Then we consider the conditional entropy $H(\mathscr{E}, \mathbf{\Pi}^* | \ \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X})$. The latter can be decomposed as

$$\begin{aligned} & H(\mathscr{E}, \mathbf{\Pi}^* | \ \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \\ = \ & H(\mathbf{\Pi}^* \mid \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) + H(\mathscr{E} \mid \mathbf{\Pi}^*, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \\ \overset{(i)}{=} \ & H(\mathbf{\Pi}^* \mid \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \overset{(ii)}{=} H(\mathbf{\Pi}^* \mid \mathbf{Y}, \mathbf{X}), \end{aligned} \tag{25}$$

where in (i) we have used that $H\left(\mathscr{E} \mid \mathbf{\Pi}^*, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}\right) = 0$ since $\mathscr{E}$ is deterministic once $\mathbf{\Pi}^*, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}$ are given, and in (ii) we use the fact $I(\widehat{\mathbf{\Pi}}; \mathbf{\Pi}^* \mid \mathbf{Y}, \mathbf{X}) = 0$ since $\widehat{\mathbf{\Pi}}$ and $\mathbf{\Pi}^*$ are independent given $\mathbf{X}$ and $\mathbf{Y}$.

At the same time, we have

$$\begin{aligned} & H\left(\mathscr{E}, \mathbf{\Pi}^* | \ \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}\right) \\ = \ & H\left(\mathscr{E} | \ \widehat{\mathbf{\Pi}}\right) + H\left(\mathbf{\Pi}^* \mid \mathscr{E}, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}\right) \\ \overset{(a)}{\leq} \ & \log 2 + H\left(\mathbf{\Pi}^* \mid \mathscr{E}, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}\right) \\ \leq \ & \log 2 + \Pr(\mathscr{E}=1) H(\mathbf{\Pi}^* \mid \mathscr{E}=1, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \\ & + \Pr(\mathscr{E}=0) H(\mathbf{\Pi}^* \mid \mathscr{E}=0, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \\ \leq \ & \log 2 + \Pr(\mathscr{E}=1) H(\mathbf{\Pi}^* \mid \mathscr{E}=1, \ \widehat{\mathbf{\Pi}}) \\ & + \Pr(\mathscr{E}=0) H(\mathbf{\Pi}^* \mid \mathscr{E}=0, \ \widehat{\mathbf{\Pi}}) \\ \leq \ & \log 2 + (1 - \Pr(\mathscr{E}=0)) H(\mathbf{\Pi}^*) \\ & + \Pr(\mathscr{E}=0) \log \frac{n!}{(n-D+1)!} \\ \overset{(b)}{=} \ & \log 2 + H(\mathbf{\Pi}^*) - \Pr(\mathscr{E}=0) \log(n-D+1)!, \end{aligned} \tag{26}$$

where in (a) we use the fact that $\mathscr{E}$ is binary and hence $H(\mathscr{E}|\widehat{\mathbf{\Pi}}) \leq \log(2)$, and in (b) we use the fact that $H(\mathbf{\Pi}^*) = \log(n!)$. Combing Eq. (25) and Eq. (26) yields that

$$\begin{aligned} \Pr(\mathscr{E}=0) &\leq \frac{I(\mathbf{\Pi}^*; \ \mathbf{Y}, \ \mathbf{X}) + \log 2}{\log(n-D+1)!} \\ &\overset{(c)}{=} \frac{I(\mathbf{\Pi}^*; \mathbf{X}) + I(\mathbf{\Pi}^*; \mathbf{Y}|\mathbf{X}) + \log 2}{\log(n-D+1)!} \\ &\overset{(d)}{=} \frac{I(\mathbf{\Pi}^*; \mathbf{Y}|\mathbf{X}) + \log 2}{\log(n-D+1)!} \\ &\overset{(e)}{\leq} \frac{(n/2)\sum_i \log\left(1+\lambda_i^2/\sigma^2\right) + \log 2}{\log(n-D+1)!}, \end{aligned} \tag{27}$$

where (c) is because of the chain rule of $I(\mathbf{\Pi}^*; \mathbf{Y}, \mathbf{X})$, (d) is because $\mathbf{\Pi}^*$ and $\mathbf{X}$ are independent and hence $I(\mathbf{\Pi}^*; \mathbf{X}) = 0$, and (e) is because of Lemma 7. According to Eq. (27), if we have $n\sum_i \log\left(1+\lambda_i^2/\sigma^2\right) + \log 4 \leq \log(n-D+1)!$, we conclude that $\Pr(\mathscr{E}=1) \geq 1/2$. □

Appendix E
Proof of Thm. 4

Proof. Following a similar argument as in Appendix C, we assume that $\mathbf{\Pi}^* = \mathbf{I}$ w.l.o.g. and consider correct recovery $\{\widehat{\mathbf{\Pi}} = \mathbf{I}\}$. We start by providing a brief roadmap of the proof.

- **Stage I:** Define the event $\mathscr{E}$ as

$$\mathscr{E} \triangleq \bigcap_{i=1}^{n} \left\{ \left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{i,:}\right\|_2^2 < \min_{j\neq i} \left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2^2 \right\}.$$

We first show that $\left\{\widehat{\mathbf{\Pi}} \neq \mathbf{I}\right\} \subseteq \overline{\mathscr{E}}$.

- **Stage II:** We would like to upper bound the probability of error $\Pr(\widehat{\mathbf{\Pi}} \neq \mathbf{I})$ by $\Pr(\overline{\mathscr{E}})$. By re-arranging terms, we can upper bound $\overline{\mathscr{E}}$ by $\overline{\mathscr{E}} \subseteq \mathscr{E}_1 \bigcup \mathscr{E}_2$, where $\mathscr{E}_1$ and $\mathscr{E}_2$ are defined as

$$\mathscr{E}_1 \triangleq \bigcup_{i=1}^{n} \bigcup_{j\neq i} \left\{ 2 \left\langle \mathbf{W}_{i,:}, \ \frac{\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:} - \mathbf{X}_{i,:}\right)}{\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:} - \mathbf{X}_{i,:}\right)\right\|_2} \right\rangle \geq \delta \right\}$$

$$\mathscr{E}_2 \triangleq \bigcup_{i=1}^{n} \bigcup_{j\neq i} \left\{ \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:} - \mathbf{X}_{i,:}\right)\right\|_2 \leq \delta \right\},$$

where $\delta > 0$ is an arbitrary positive number. We then upper bound $\Pr(\mathscr{E}_1)$ and $\Pr(\mathscr{E}_2)$ separately.

- **Stage III:** Treating the above upper bounds as functions of δ, we complete the proof by choosing δ appropriately and invoking the Condition (10).

We now turn to the details of the proof.

Stage I: We first establish that $\left\{\widehat{\mathbf{\Pi}} \neq \mathbf{I}\right\} \subseteq \mathscr{E}^c$ by showing that $\mathscr{E} \subseteq \left\{\widehat{\mathbf{\Pi}} = \mathbf{I}\right\}$. Notice that $\mathscr{E}$ can be rewritten as

$$\mathscr{E} = \bigcap_{i=1}^{n} \bigcap_{j\neq i} \left\{ \left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{i,:}\right\|_2^2 < \left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2^2 \right\}.$$

Based on the definition of the ML estimator (5), we must have

$$\left\|\mathbf{Y} - \widehat{\mathbf{\Pi}}\mathbf{X}\mathbf{B}^*\right\|_2^2 \ \leq \|\mathbf{Y} - \mathbf{X}\mathbf{B}^*\|_2^2, \tag{28}$$

Assuming that $\widehat{\mathbf{\Pi}} \neq \mathbf{I}$, then for each term $\left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{i,:}\right\|_2$ we have

$$\begin{aligned} \left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{i,:}\right\|_2^2 &\leq \min_{j\neq i} \left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2^2 \\ &< \left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\left(\widehat{\mathbf{\Pi}}\mathbf{X}\right)_{i,:}\right\|_2^2, \end{aligned}$$

which leads to $\|\mathbf{Y} - \mathbf{X}\mathbf{B}^*\|_2^2 < \left\|\mathbf{Y} - \widehat{\mathbf{\Pi}}\mathbf{X}\mathbf{B}^*\right\|_2^2$, contradicting Eq. (28). Hence we have proved that $\mathscr{E} \subseteq \left\{\widehat{\mathbf{\Pi}} = \mathbf{I}\right\}$.

Stage II: In this stage, we will prove that $\overline{\mathscr{E}} \subseteq \mathscr{E}_1 \bigcup \mathscr{E}_2$. First, we expand $\overline{\mathscr{E}}$ as

$$\overline{\mathscr{E}} \triangleq \bigcup_{i=1}^{n} \bigcup_{j\neq i} \left\{ \left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{i,:}\right\|_2^2 \geq \left\|\mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2^2 \right\}.$$

Note that for each event in the union, the left hand side can be rewritten as $\|\mathbf{W}_{i,:}\|_2^2$ and the right hand side can be written as

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

Hence, the event $\overline{\mathscr{E}}$ is equivalent to

$$\begin{aligned} \overline{\mathscr{E}} = \ & \bigcup_{i=1}^{n} \bigcup_{j\neq i} \left\{ 2\left\langle \mathbf{W}_{i,:}, \frac{\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:} - \mathbf{X}_{i,:}\right)}{\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:} - \mathbf{X}_{i,:}\right)\right\|_2} \right\rangle \right. \\ & \left. \geq \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:} - \mathbf{X}_{i,:}\right)\right\|_2 \right\} \subseteq \mathscr{E}_1 \bigcup \mathscr{E}_2, \end{aligned}$$

since otherwise we will have the inequality reversed. Hence, we can upper bound $\Pr(\overline{\mathscr{E}})$ as

$$\begin{aligned}\Pr\left(\overline{\mathscr{E}}\right) \leq\ & \Pr(\mathscr{E}_1)+\Pr(\mathscr{E}_2)\\ \leq\ & \underbrace{\sum_{i=1}^{n}\sum_{j\neq i}\Pr\left\{2\left\langle \mathbf{W}_{i,:},\ \frac{\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)}{\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)\right\|_2}\right\rangle \geq \delta\right\}}_{\triangleq \mathscr{P}_1}\\ &+\underbrace{\sum_{i=1}^{n}\sum_{j\neq i}\Pr\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)\right\|_2\leq\delta\right)}_{\triangleq \mathscr{P}_2},\end{aligned}$$

where the terms $\mathscr{P}_1$ and $\mathscr{P}_2$ can be bounded by Lemma 10 and Lemma 11 (given below), respectively.

Stage III: Set δ^2 as $16\sigma^2\log\frac{n}{\epsilon_0}$, where $\epsilon_0=\alpha_0^{\kappa\varrho(\mathbf{B}^*)}/n$. We can bound $\mathscr{P}_1$ as

$$\mathscr{P}_1 \leq n^2\ \exp\left(-\frac{16\sigma^2}{8\sigma^2}\log\frac{n}{\epsilon_0}\right)=\epsilon_0^2. \tag{29}$$

At the same time, we can show that $\mathscr{P}_2$ is no greater than ϵ_0^2. To invoke Lemma 11, first we need to verify the condition $\delta^2<\alpha_0^2\|\mathbf{B}^*\|_{\mathrm{F}}^2/2$. This is proved by

$$\begin{aligned}\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2} &\overset{(i)}{\geq} 32\log\left(\frac{n}{\epsilon_0}\right)\left(\frac{n}{\epsilon_0}\right)^{4/\kappa\varrho(\mathbf{B}^*)}\\ &\overset{(ii)}{=} 32\log\left(\frac{n}{\epsilon_0}\right)\left(\frac{n^2}{\alpha_0^{\kappa\varrho(\mathbf{B}^*)}}\right)^{4/\kappa\varrho(\mathbf{B}^*)}\\ &\overset{(iii)}{\geq} \frac{32}{\alpha_0^2}\log\left(\frac{n}{\epsilon_0}\right),\end{aligned}$$

where in (i) we use condition (10), in (ii) we use the definition of $\epsilon_0=\alpha_0^{\kappa\varrho(\mathbf{B}^*)}/n$, and in (iii) we use $\alpha_0\in(0,1)$ and $n\geq 1$. We can then invoke Lemma 11 and bound $\mathscr{P}_2$ as

$$\begin{aligned}\mathscr{P}_2 &\leq n^2\left(\frac{2\delta^2}{\|\mathbf{B}^*\|_{\mathrm{F}}^2}\right)^{\kappa\varrho(\mathbf{B}^*)/2}\\ &\overset{(a)}{=} n^2\exp\left[-\frac{\kappa\varrho(\mathbf{B}^*)}{2}\left(\log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2}\right)-\log\left(32\log\left(\frac{n}{\epsilon_0}\right)\right)\right)\right]\\ &\overset{(b)}{\leq} n^2\exp\left[-\frac{\kappa\varrho(\mathbf{B}^*)}{2}\left(\frac{4}{\kappa\varrho(\mathbf{B}^*)}\log\frac{n}{\epsilon_0}\right)\right]=\epsilon_0^2,\end{aligned} \tag{30}$$

where in (a) we plug in the definition $\delta^2=16\sigma^2\log(n/\epsilon_0)$, and in (b) we use condition (10). Combining the bounds for $\mathscr{P}_1$ in Eq. (29) and $\mathscr{P}_2$ in Eq. (30) will complete the proof. □

Lemma 10. *It holds that*

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j\neq i}\Pr\left(2\left\langle \mathbf{W}_{i,:},\ \frac{\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)}{\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)\right\|_2}\right\rangle\geq\delta\right)\\ &\leq n^2e^{-\delta^2/8\sigma^2}.\end{aligned}$$

Proof. First, we consider a single term, namely $2\left\langle \mathbf{W}_{i,:}, \frac{\mathbf{B}^{*\top}(\mathbf{X}_{j,:}-\mathbf{X}_{i,:})}{\|\mathbf{B}^{*\top}(\mathbf{X}_{j,:}-\mathbf{X}_{i,:})\|_2}\right\rangle$, $(1\leq i<j\leq n)$. With $\mathbf{X}$ fixed, it is easy to check that this term is a Gaussian random variable with zero mean and variance $4\sigma^2$.

Then the probability $\Pr\left(2\left\langle \mathbf{W}_{i,:},\ \frac{\mathbf{B}^{*\top}(\mathbf{X}_{j,:}-\mathbf{X}_{i,:})}{\|\mathbf{B}^{*\top}(\mathbf{X}_{j,:}-\mathbf{X}_{i,:})\|_2}\right\rangle\geq\delta\right)$ can be bounded as

$$\begin{aligned}&\Pr\left(2\left\langle \mathbf{W}_{i,:},\ \frac{\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)}{\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)\right\|_2}\right\rangle\geq\delta\right)\\ &=\mathbb{E}_{\mathbf{X}}\Pr\left(2\left\langle \mathbf{W}_{i,:},\ \frac{\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)}{\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)\right\|_2}\right\rangle\geq\delta\mid\mathbf{X}\right)\\ &\overset{(i)}{\leq}\mathbb{E}_{\mathbf{X}}\ e^{-\delta^2/8\sigma^2}=e^{-\delta^2/8\sigma^2},\end{aligned}$$

where in (i) we use the tail bound for the Gaussian RV $\mathbf{W}_{i,:}$. Combining the above together, we show that $\mathscr{P}_1\leq n^2e^{-\delta^2/8\sigma^2}$ and complete the proof. □

Lemma 11. *Given that* $\delta^2<\frac{\alpha_0^2\|\mathbf{B}^*\|_{\mathrm{F}}^2}{2}$*, we have*

$$\sum_{i=1}^{n}\sum_{j\neq i}\Pr\left(\left\|\mathbf{B}^{*\top}(\mathbf{X}_{j,:}-\mathbf{X}_{i,:})\right\|_2\leq\delta\right)\leq n^2\left(\frac{2\delta^2}{\|\mathbf{B}^*\|_{\mathrm{F}}^2}\right)^{\kappa\varrho(\mathbf{B}^*)/2},$$

where $\alpha_0\in(0,1)$ *is a universal constant.*

Proof. We consider an arbitrary term $\left\|\mathbf{B}^{*\top}(\mathbf{X}_{i,:}-\mathbf{X}_{j,:})\right\|_2$, $(i<j)$, and define $\mathbf{Z}=\frac{\mathbf{X}_{i,:}-\mathbf{X}_{j,:}}{\sqrt{2}}$. It is easy to verify that $\mathbf{Z}$ is a p-dimensional random vector with i.i.d. $\mathscr{N}(0,1)$-entries. We then have

$$\Pr\left(\left\|\mathbf{B}^{*\top}(\mathbf{X}_{i,:}-\mathbf{X}_{j,:})\right\|_2\leq\delta\right)=\Pr\left(\left\|\mathbf{B}^{*\top}\mathbf{Z}\right\|_2^2\leq 2\delta^2\right).$$

According to Lemma 2.6 in [27] (which is re-stated in Appendix H herein), this probability can be bounded as

$$\begin{aligned}&\Pr\left(\left\|\mathbf{B}^{*\top}\mathbf{Z}\right\|_2^2\leq 2\delta^2\right)=\Pr\left(\left\|\mathbf{B}^{*\top}\mathbf{Z}\right\|_2\leq\sqrt{2}\delta\right)\\ &\leq\exp\left(-\kappa\varrho(\mathbf{B}^*)\log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}}{\sqrt{2}\delta}\right)\right)=\left(\frac{2\delta^2}{\|\mathbf{B}^*\|_{\mathrm{F}}^2}\right)^{\kappa\varrho(\mathbf{B}^*)/2},\end{aligned}$$

provided $\delta^2<\alpha_0^2\|\mathbf{B}^*\|_{\mathrm{F}}^2/2$, where $\alpha_0\in(0,1)$ is a universal constant. With the union bound, we complete the proof. □

Appendix F
Proof of Thm. 5

Proof. Before we proceed, we give an outline of our proof.

- **Stage I:** We decompose the event $\left\{\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*\right\}$ as

$$\left\{\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*\right\}=\bigcup_{\mathbf{\Pi}\neq\mathbf{\Pi}^*}\left\{\left\|P_{\mathbf{\Pi}\mathbf{X}}^{\perp}\mathbf{Y}\right\|_{\mathrm{F}}^2\leq\left\|P_{\mathbf{\Pi}^*\mathbf{X}}^{\perp}\mathbf{Y}\right\|_{\mathrm{F}}^2\right\}, \tag{31}$$

  and bound the probability of each individual event in Eq. (31).
- **Stage II:** For fixed Hamming distance $d_H(\mathbf{\Pi},\mathbf{\Pi}^*)=h$, we will prove

$$\begin{aligned}&\Pr\left(\left\|P_{\mathbf{\Pi}\mathbf{X}}^{\perp}\mathbf{Y}\right\|_{\mathrm{F}}^2\leq\left\|P_{\mathbf{\Pi}^*\mathbf{X}}^{\perp}\mathbf{Y}\right\|_{\mathrm{F}}^2,\ d_H(\mathbf{\Pi},\mathbf{\Pi}^*)=h\right)\\ &\leq\exp\left(-\frac{t\times\mathrm{snr}}{72}\right)+6r\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}}\\ &+r\exp\left(-n\log\frac{n}{2}\right)+2\exp\left(-\frac{1}{288}\left(\frac{t^2\times\mathrm{snr}^2}{mh}\wedge(t\times\mathrm{snr})\right)\right),\end{aligned} \tag{32}$$

where r denotes the rank of $\mathbf{B}^*$, and $t>0$ is an arbitrary positive number.

- **Stage III:** Under the condition specified by Eq. (13) and $\text{snr}\times n^{-\frac{2n}{n-p}} \geq 1$, we set t as $\sqrt{mh}\log\left(\text{snr}\times mn^{-\frac{2n}{n-p}}\right)/\text{snr}$ and show that

$$\begin{aligned}&\Pr\left(\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 \leq \left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2,\ d_H(\mathbf{\Pi},\mathbf{\Pi}^*)=h\right)\\ &\leq 9n^{-(1+\epsilon)h}+r\exp\left(-n\log\frac{n}{2}\right).\end{aligned}$$

- **Stage IV:** We prove that

$$\Pr\left(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*\right)\leq 10.36\left(\frac{1}{n^{\epsilon}\left(n^{\epsilon}-1\right)}\vee\frac{1}{n^{\epsilon}}\right),$$

when n is large, where $\epsilon>0$ is some positive constant.

As the outline of our proof, we start with providing the details of Stage I and Stage IV, while the proofs of Stage II and Stage III are given in Lemma 12 and Lemma 13, respectively.

Stage I: From the definition of ML estimator in Eq. (5), failure of recovery requires at least one pair $(\mathbf{\Pi},\mathbf{B})$ distinct from $(\mathbf{\Pi}^*,\mathbf{B}^*)$ such that

$$\|\mathbf{Y}-\mathbf{\Pi}\mathbf{X}\mathbf{B}\|_{\mathrm{F}}^2\leq\|\mathbf{Y}-\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}^2.$$

Note that the optimal $\mathbf{B}$ corresponding to $\mathbf{\Pi}$ can be expressed as $\mathbf{B}=(\mathbf{\Pi}\mathbf{X})^{\dagger}\mathbf{Y}$, where $(\mathbf{\Pi}\mathbf{X})^{\dagger}\triangleq\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{\Pi}^{\top}$. Back-substitution yields

$$\left\|\mathbf{Y}-\mathbf{\Pi}\mathbf{X}(\mathbf{\Pi}\mathbf{X})^{+}\mathbf{Y}\right\|_{\mathrm{F}}^2=\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2,$$

which proves the claim.

Stage II and Stage III: As stated above, the detailed proof can be found in Lemma 12 and Lemma 13.

Stage IV: We have

$$\begin{aligned}&\Pr\left(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*\right)\\ &\leq\sum_{h\geq 2}\binom{n}{h}h!\Pr\left(\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2\leq\left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2,\ d_H(\mathbf{\Pi},\mathbf{\Pi}^*)=h\right)\\ &\overset{(i)}{\leq}\sum_{h\geq 2}\binom{n}{h}h!\left(9n^{-(1+\epsilon)h}+r\exp\left(-n\log\frac{n}{2}\right)\right)\\ &\overset{(ii)}{\leq}9\sum_{h\geq 2}n^h n^{-(1+\epsilon)h}+r\sum_{h\geq 2}n!\exp\left(-n\log\frac{n}{2}\right)\\ &\overset{(iii)}{\leq}9\sum_{h\geq 2}n^{-\epsilon h}+\ r\sum_{h\geq 2}e\sqrt{n}\ \exp\left(n\log n-n\log\left(\frac{n}{2}\right)-n\right)\\ &\leq\frac{9}{n^{\epsilon}\left(n^{\epsilon}-1\right)}+e\sum_{h}rn^{\frac{1}{2}}\exp\left(-n\log\left(\frac{e}{2}\right)\right)\\ &\overset{(iv)}{\leq}\frac{9}{n^{\epsilon}\left(n^{\epsilon}-1\right)}+\frac{e}{2}\sum_{h}n^{\frac{3}{2}}\exp\left(-n\log\left(\frac{e}{2}\right)\right)\\ &\leq\frac{9}{n^{\epsilon}\left(n^{\epsilon}-1\right)}+\frac{e}{2}n^{\frac{5}{2}}\exp\left(-n\log\left(\frac{e}{2}\right)\right)\\ &\overset{(v)}{\leq}\frac{9}{n^{\epsilon}\left(n^{\epsilon}-1\right)}+\frac{e}{2}\exp\left(-\epsilon\log n\right)\\ &\leq 10.36\left(\frac{1}{n^{\epsilon}\left(n^{\epsilon}-1\right)}\vee\frac{1}{n^{\epsilon}}\right),\end{aligned}$$

where in (i) we use Eq. (32), in (ii) we use $\frac{n!}{(n-h)!}\leq n^h$ and $\frac{n!}{(n-h)!}\leq n!$, in (iii) we use *Stirling's approximation* in the form $n!\leq en^{n+0.5}e^{-n}$, in (iv) we use $r\leq\min(m,p)$ and $p\leq\frac{n}{2}$ (according to our assumption in Sec. II), and in (v), we use $n\log(\frac{e}{2})>\left(\frac{5}{2}+\epsilon\right)\log n$ when n is sufficiently large (e.g., when $\epsilon=0.5$, we require $n\geq 36$; when $\epsilon=1$, we require $n\geq 44$). The proof is hence complete. □

Lemma 12. *We have*

$$\begin{aligned}&\Pr\left(\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2\leq\left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2,\ d_H(\mathbf{\Pi},\mathbf{\Pi}^*)=h\right)\\ &\leq\ \exp\left(-\frac{t\times\text{snr}}{72}\right)\ +6r\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}}\\ &+r\exp\left(-n\log\left(\frac{n}{2}\right)\right)+2\exp\left(-\frac{1}{288}\left(\frac{t^2\times\text{snr}^2}{mh}\wedge(t\times\text{snr})\right)\right),\end{aligned}$$

where $t<mh$ is an arbitrary positive number.

Proof. Define the event $\mathscr{E}$ as

$$\mathscr{E}\triangleq\left\{\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2\leq\left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2,\ d_H(\mathbf{\Pi},\mathbf{\Pi}^*)=h\right\}.$$

we then separately bound the probability $\Pr(\mathscr{E})$ based on the whether $T_{\mathbf{\Pi}}\leq t\|\mathbf{B}^*\|_{\mathrm{F}}^2/m$ or not, where $T_{\mathbf{\Pi}}$ is defined as

$$T_{\mathbf{\Pi}}=\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}^2.\tag{33}$$

Case I: We assume $T_{\mathbf{\Pi}}\leq t\|\mathbf{B}^*\|_{\mathrm{F}}^2/m$ and obtain

$$\Pr\left(\mathscr{E},T_{\mathbf{\Pi}}\leq\frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right)\leq\ \Pr\left(T_{\mathbf{\Pi}}\leq\frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right).$$

We then use that $\text{SVD}(\mathbf{B}^*)=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$ (as defined in Appendix A), such that $\mathbf{\Sigma}=\text{diag}(\beta_1,\ \beta_2,\ \cdots,\beta_r,0,\cdots)$, where r denotes the rank of $\mathbf{B}^*$ ($r\leq\min(m,p)$), and β_i denotes the corresponding singular values.

Due to the rotational invariance of the Gaussian distribution and $\mathbf{V}$ being unitary, it is easy to check that $T_{\mathbf{\Pi}}$ has the same distribution as $\|P^{\perp}_{\mathbf{X}}\mathbf{\Pi}\mathbf{X}\mathbf{\Sigma}\|_F^2$. Therefore, we have

$$\begin{aligned}&\Pr\left(T_{\mathbf{\Pi}}\leq\frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right)\leq\ \sum_{i=1}^{r}\Pr\left(\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{\Pi}^*\mathbf{X}\beta_i\mathbf{e}_i\right\|_{\mathrm{F}}^2\leq\frac{t\beta_i^2}{m}\right)\\ &\overset{(i)}{\leq}\ r\left(\exp\left(-n\log\left(\frac{n}{2}\right)\right)+6\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}}\right),\end{aligned}$$

where (i) follows from Lemma 5 in [6].

Case II: We have $T_{\mathbf{\Pi}}>t\|\mathbf{B}\|_{\mathrm{F}}^2/m$. Defining the events $\mathscr{E}_1$ and $\mathscr{E}_2$ as

$$\begin{aligned}\mathscr{E}_1&\triangleq\left\{\ \left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2-\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2\leq\frac{2T_{\mathbf{\Pi}}}{3}\right\};\\ \mathscr{E}_2&\triangleq\left\{\ \left|\left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2-\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2\right|\geq\frac{T_{\mathbf{\Pi}}}{3}\right\},\end{aligned}\tag{34}$$

where we omit the condition $d_H(\mathbf{\Pi},\mathbf{\Pi}^*)=h, T_{\mathbf{\Pi}}>t\|\mathbf{B}^*\|_{\mathrm{F}}^2/m$ for simplicity. We first provide the basic proof structure.

- **Step II.(A)**: We show $\overline{\mathscr{E}_1}\bigcap\overline{\mathscr{E}_2}\subseteq\overline{\mathscr{E}}$, which implies $\mathscr{E}\subseteq\mathscr{E}_1\bigcup\mathscr{E}_2$. Invoking the union bound, we have

$$\Pr\left(\mathscr{E},T_{\mathbf{\Pi}}>\frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right)\leq\Pr(\mathscr{E}_1\bigcup\mathscr{E}_2)\leq\Pr(\mathscr{E}_1)+\Pr(\mathscr{E}_2).$$

- **Step II.(B)**: We separately upper bound $\Pr(\mathscr{E}_1)$ and $\Pr(\mathscr{E}_2)$.

We now turn to the proof details.

Step II.(A) Conditional on $\overline{\mathscr{E}_1} \bigcap \overline{\mathscr{E}_2}$, we have

$$\begin{aligned}
&\left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 \\
\overset{(a)}{=}\ & \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2 + \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2 \\
\geq\ & \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2 - \left|\left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2\right| \\
\overset{(b)}{>}\ & \frac{2T_{\mathbf{\Pi}}}{3} - \frac{T_{\mathbf{\Pi}}}{3} > 0,
\end{aligned}$$

where in (a) we use the fact $P^{\perp}_{\mathbf{\Pi}^*\underline{\mathbf{X}}}\mathbf{Y} = P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}$, and in (b) we use the definitions of $\overline{\mathscr{E}_1}$ and $\overline{\mathscr{E}_2}$.

Step II.(B) Then we separately bound $\Pr(\mathscr{E}_1)$ and $\Pr(\mathscr{E}_2)$. Regarding $\Pr(\mathscr{E}_1)$, we first expand

$$\begin{aligned}
&\left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2 \\
=\ & \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}^2 + 2\left\langle P^{\perp}_{\mathbf{\Pi X}}\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*,\ P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\rangle.
\end{aligned}$$

Conditional on the sensing matrix $\mathbf{X}$, we have that $\left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2$ follows a Gaussian distribution, namely, $\mathscr{N}(T_{\mathbf{\Pi}}, 4\sigma^2 T_{\mathbf{\Pi}})$. Therefore, we obtain

$$\begin{aligned}
&\Pr(\mathscr{E}_1) = \mathbb{E}\mathbb{1}\left(\mathscr{E}_1\right) \\
\overset{(c)}{=}\ & \mathbb{E}_{\mathbf{X}}\left[\mathbb{1}\left(T_{\mathbf{\Pi}} > \frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right)\right. \\
&\left.\times\ \mathbb{E}_{\mathbf{W}}\mathbb{1}\left(\left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2 \leq \frac{2T_{\mathbf{\Pi}}}{3}\right)\right] \\
\overset{(d)}{\leq}\ & \mathbb{E}_{\mathbf{X}}\left[\mathbb{1}\left(T_{\mathbf{\Pi}} > \frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right) \times \exp\left(-\frac{T_{\mathbf{\Pi}}}{72\sigma^2}\right)\right] \\
\leq\ & \exp\left(-\frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{72m\sigma^2}\right) = \exp\left(-\frac{t\times\mathrm{snr}}{72}\right) \qquad (35)
\end{aligned}$$

where (c) results from independence of $\mathbf{X}$ and $\mathbf{W}$, and in (d) we use a standard tail bound for Gaussian random variables.

Next, we bound $\Pr(\mathscr{E}_2)$. We have

$$\begin{aligned}
&\left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2 = \left\|P_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2 - \left\|P_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2 \\
=\ & \left\|P_{\mathbf{\Pi X}\setminus\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2 - \left\|P_{\mathbf{\Pi}^*\mathbf{X}\setminus\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2,
\end{aligned}$$

where $\mathbf{\Pi X} \setminus \mathbf{\Pi}^*\mathbf{X}$ ($\mathbf{\Pi}^*\mathbf{X} \setminus \mathbf{\Pi X}$) is the short-hand for $\mathrm{range}\left(\mathbf{\Pi X}\right) \setminus \mathrm{range}\left(\mathbf{\Pi}^*\mathbf{X}\right)$ ($\mathrm{range}\left(\mathbf{\Pi}^*\mathbf{X}\right) \setminus \mathrm{range}\left(\mathbf{\Pi X}\right)$). Setting $k = p \wedge h$, we have that $\left\|P_{\mathbf{\Pi X}\setminus\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2/\sigma^2$ is χ^2-RV with mk degrees of freedom according to Appendix B.1 in [6]. We conclude that

$$\begin{aligned}
&\Pr\left(\mathscr{E}_2\right) \\
\leq\ & 2\Pr\left(\left|\left\|P_{\mathbf{\Pi X}\setminus\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2 - mk\sigma^2\right| \geq \frac{T_{\mathbf{\Pi}}}{6},\ T_{\mathbf{\Pi}} > \frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right) \\
\overset{(e)}{\leq}\ & 2\exp\left(-\frac{1}{8}\left(\frac{t^2\times\mathrm{snr}^2}{36mk} \wedge \frac{t\times\mathrm{snr}}{6}\right)\right) \\
\leq\ & 2\exp\left(-\frac{1}{8}\left(\frac{t^2\times\mathrm{snr}^2}{36mh} \wedge \frac{t\times\mathrm{snr}}{6}\right)\right), \qquad (36)
\end{aligned}$$

where in (e) we use the concentration inequality for χ^2-RVs given in Appendix H, Lemma 19. Combing Eq. (35) and Eq. (36) will give us

$$\begin{aligned}
&\Pr\left(\mathscr{E}, T_{\mathbf{\Pi}} > \frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right) \\
\leq\ & \exp\left(-\frac{t\times\mathrm{snr}}{72}\right) + 2\exp\left(-\frac{1}{288}\left(\frac{t^2\times\mathrm{snr}^2}{mh} \wedge (t\times\mathrm{snr})\right)\right).
\end{aligned}$$

□

Lemma 13. *Given that* $\mathrm{snr}\times n^{-\frac{2n}{n-p}} \geq 1$ *and* $\log(m\times\mathrm{snr}) \geq 380\left(1+\epsilon+\frac{n\log(n)}{190(n-p)}+\frac{1}{2}\log r(\mathbf{B}^*)\right)$, *where* $\epsilon > 0$ *are some constants, we have*

$$\begin{aligned}
&\exp\left(-\frac{t\times\mathrm{snr}}{72}\right) + 2\exp\left(-\frac{1}{288}\left(\frac{t^2\times\mathrm{snr}^2}{mh} \wedge (t\times\mathrm{snr})\right)\right) \\
&+\ 6r\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}} \\
\leq\ & 9n^{-(1+\epsilon)h}.
\end{aligned}$$

Proof. Here we choose t as $\sqrt{mh}\log\left(\mathrm{snr}\times mn^{-\frac{2n}{n-p}}\right)/\mathrm{snr}$. Easily we can verify that $t < mh$. Next, we separately discuss the following terms

- $\mathscr{T}_1 \triangleq \exp\left(-t\times\mathrm{snr}/72\right)$.
- $\mathscr{T}_2 \triangleq \exp\left(-\left(\frac{t^2\times\mathrm{snr}^2}{mh} \wedge (t\times\mathrm{snr})\right)/288\right)$.
- $\mathscr{T}_3 \triangleq r\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}}$.

Term $\mathscr{T}_1$: We have

$$\begin{aligned}
&\exp\left(-\frac{t\times\mathrm{snr}}{72}\right) = \exp\left(-\frac{\sqrt{mh}}{72}\log\left(\mathrm{snr}\times mn^{-\frac{2n}{n-p}}\right)\right) \\
\leq\ & \exp\left(-\frac{h}{72}\log\left(\mathrm{snr}\times mn^{-\frac{2n}{n-p}}\right)\right). \qquad (37)
\end{aligned}$$

Term $\mathscr{T}_2$: Provided that $\left(t^2\times\mathrm{snr}^2/(mh)\right) \wedge (t\times\mathrm{snr}) = t\times\mathrm{snr}$, the term $\mathscr{T}_2$ is of a similar form as $\mathscr{T}_1$ in Eq. (37). Here we focus on the case in which $\frac{t^2\times\mathrm{snr}^2}{mh} \wedge (t\times\mathrm{snr}) = \frac{t^2\times\mathrm{snr}^2}{mh}$. The right hand side of this equality can be expanded as

$$\frac{t^2\times\mathrm{snr}^2}{mh} = h\log^2\left(\mathrm{snr}\times mn^{-\frac{2n}{n-p}}\right) \overset{(i)}{\geq} h\log\left(\mathrm{snr}\times mn^{-\frac{2n}{n-p}}\right),$$

where in (i) we use the fact $\mathrm{snr}\times mn^{-\frac{2n}{n-p}} \geq 323$, which can be verified by Eq. (13). We then obtain

$$\mathscr{T}_2 \leq \exp\left(-\frac{h}{288}\log\left(\mathrm{snr}\times mn^{-\frac{2n}{n-p}}\right)\right). \qquad (38)$$

Term $\mathscr{T}_3$: We have

$$\begin{aligned}
&r\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}} \\
=\ & r\exp\left[-\frac{h}{10}\left(\log\frac{mh}{tn^{\frac{2n}{n-p}}} + \frac{tn^{\frac{2n}{n-p}}}{mh} - 1\right)\right] \\
\overset{(i)}{=}\ & r\exp\left[-\frac{h}{10}\left(-\frac{1}{2}\log(m) - \log\frac{\log(z)}{z} + \frac{\sqrt{m}\log(z)}{z} - 1\right)\right] \\
\leq\ & r\exp\left[-\frac{h}{10}\left(-\frac{1}{2}\log(m) - \log\frac{\log(z)}{z} + \frac{\log(z)}{z} - 1\right)\right] \\
\overset{(ii)}{\leq}\ & r\exp\left[-\frac{h}{10}\left(\frac{\log(z)}{1.9} - \frac{\log(m)}{2}\right)\right] \\
\overset{(iii)}{\leq}\ & r\exp\left[-\frac{h}{380}\log(\mathrm{snr}\times mn^{-\frac{2n}{n-p}})\right], \qquad (39)
\end{aligned}$$

where in (i) we set $z = \text{snr} \times mn^{-\frac{2n}{n-p}} \geq 323$, in (ii) we use the fact $\frac{\log(z)}{z} - 1 - \log\frac{\log(z)}{z} \geq \frac{\log(z)}{1.9}$ for $z \geq 323$, and in (iii) we use the fact $\text{snr} \times n^{-\frac{2n}{n-p}} \geq 1$.

Combining Eq. (37), (38) and (39), we conclude that

$$\begin{aligned}
&\exp\left(-\frac{t\times\text{snr}}{72}\right) + \exp\left(-\frac{1}{288}\left(\frac{t^2\times\text{snr}^2}{mh} \wedge (t\times\text{snr})\right)\right) \\
&\quad + 6r\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}} \\
&\leq\ 9r\exp\left[-\frac{h}{380}\log\left(\text{snr}\times mn^{-\frac{2n}{n-p}}\right)\right].
\end{aligned}$$

Under the condition specified by Eq. (13), we have

$$\begin{aligned}
\frac{\log\left(\text{snr}\times mn^{-\frac{2n}{n-p}}\right)}{380} &= \frac{\log(m\times\text{snr})}{380} - \frac{n\log(n)}{190(n-p)} \\
&\geq\ (1+\epsilon)\log n + \frac{1}{2}\log r.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&r\exp\left(-\frac{h}{380}\log\left(\text{snr}\times mn^{-\frac{2n}{n-p}}\right)\right) \\
&\leq\ r\exp\left[-h(1+\epsilon)\log n - \frac{h}{2}\log r\right] \overset{(ii)}{\leq} n^{-(1+\epsilon)h},
\end{aligned}$$

where in (ii) we have $r^{1-\frac{h}{2}} \leq 1$ since $h \geq 2$. This completes the proof. □

Appendix G
Proof of Thm. 6

Proof. Here we adopt the same proof strategy as in Thm. 5. For the sake of brevity, we only present the parts that are different compared with the proof of Thm. 5.

- **Stage I:** Given the requirement $d_H(\mathbf{I}, \mathbf{\Pi}^*) \leq h_{\max}$, the triangle inequality implies that

$$d_H(\widehat{\mathbf{\Pi}}, \mathbf{\Pi}^*) \leq d_H(\mathbf{I}, \widehat{\mathbf{\Pi}}) + d_H(\mathbf{I}, \mathbf{\Pi}^*) \leq 2h_{\max}.$$

  Hence, we can confine ourselves to the case in which $d_H(\mathbf{\Pi}, \mathbf{\Pi}^*) \leq 2h_{\max}$.
- **Stage II**: We replace Lemma 12 with Lemma 14.
- **Stage III**: We replace Lemma 13 with Lemma 16.
- **Stage IV**: We use the same argument as Stage IV in proving Thm. 5 and complete the proof.

□

Lemma 14. *Given that $rh \leq n/4$ and $t \leq 0.125h$, we have*

$$\begin{aligned}
&\Pr\left(\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\text{F}}^2 \leq \left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\text{F}}^2,\ d_H(\mathbf{\Pi}, \mathbf{\Pi}^*) = h\right) \\
&\leq\ 6\exp\left(-\frac{c_0 h\varrho(\mathbf{B}^*)}{5}\right) \\
&+\ \exp\left(\frac{rh}{2}\left(\log\left(\frac{t}{h}\right) - \frac{t}{h} + 1\right) + 4.18rh\right) \\
&+\ 2\exp\left(-\frac{mt\times\text{snr}}{288}\left(\frac{t\times\text{snr}}{h}\wedge 1\right)\right) \\
&+\ \exp\left(-\frac{mt\times\text{snr}}{72}\right).
\end{aligned}$$

Proof. Similar to the proof of Lemma 12, we separately bound $\Pr\left(\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\text{F}}^2 \leq \left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\text{F}}^2,\ d_H(\mathbf{\Pi}, \mathbf{\Pi}^*) = h\right)$ based on whether $T_{\mathbf{\Pi}} \leq t\|\mathbf{B}^*\|_{\text{F}}^2$ or not.

Case I: We assume that $T_{\mathbf{\Pi}} \leq t\|\mathbf{B}^*\|_{\text{F}}^2$ and obtain the bound

$$\begin{aligned}
&\Pr\left(\left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\text{F}}^2 \leq \left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\text{F}}^2,\ d_H(\mathbf{\Pi}, \mathbf{\Pi}^*) = h,\ T_{\mathbf{\Pi}} \leq t\|\mathbf{B}^*\|_{\text{F}}^2\right) \\
&\leq \Pr\left(T_{\mathbf{\Pi}} \leq t\|\mathbf{B}^*\|_{\text{F}}^2\right).
\end{aligned}$$

Here, the goal is to prove that, given $rh \leq n/4$, we have

$$\begin{aligned}
\Pr\left(T_{\mathbf{\Pi}} \leq t\|\mathbf{B}^*\|_{\text{F}}^2\right) &\leq 6\exp\left(-\frac{c_0 h\varrho(\mathbf{B}^*)}{5}\right) \\
&\quad + \exp\left(\frac{rh}{2}\left(\log\left(\frac{t}{h}\right) - \frac{t}{h} + 1\right) + 4.18rh\right),
\end{aligned}$$

where $t < 0.125h$, and c_0 is some positive constant.

We observe that $\left\|P^{\perp}_{\mathbf{X}}\mathbf{\Pi}\mathbf{X}\mathbf{B}^*\right\|_{\text{F}}^2 = \left\|P^{\perp}_{\mathbf{X}}(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\right\|_{\text{F}}^2$ and then define two events Ω_1 and Ω_2 as

$$\begin{aligned}
\Omega_1 &\triangleq \left\{\left\|P^{\perp}_{X}\frac{(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*}{\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}}\right\|_{\text{F}}^2 \leq \frac{t\|\mathbf{B}^*\|_{\text{F}}^2}{\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}^2}\right\}, \\
\Omega_2 &\triangleq \left\{0 < \|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}^2 \leq h\|\mathbf{B}^*\|_{\text{F}}^2\right\}.
\end{aligned}$$

Note that

$$\begin{aligned}
\Omega_1 \bigcap \overline{\Omega}_2 &\overset{(i)}{=} \left\{\left\|P^{\perp}_{\mathbf{X}}\frac{(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*}{\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}}\right\|_{\text{F}}^2 \leq \frac{t\|\mathbf{B}^*\|_{\text{F}}^2}{\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}^2},\right. \\
&\qquad \left.\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}^2 > h\|\mathbf{B}^*\|_{\text{F}}^2\right\} \\
&\subseteq \left\{\left\|P^{\perp}_{\mathbf{X}}\frac{(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*}{\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}}\right\|_{\text{F}}^2 < \frac{t}{h}\right\},
\end{aligned}$$

where in (i) we omit the case where $\{\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}} = 0\}$ since it is of measure zero. We obtain

$$\begin{aligned}
&\Pr(\Omega_1) = \Pr\left(\Omega_1\bigcap\Omega_2\right) + \Pr\left(\Omega_1\bigcap\overline{\Omega}_2\right) \\
&\leq\ \Pr(\Omega_2) + \Pr\left(\Omega_1\bigcap\overline{\Omega}_2\right).
\end{aligned}$$

Then we separately bound $\Pr(\Omega_2)$ and $\Pr\left(\Omega_1\bigcap\overline{\Omega}_2\right)$.

Case I.(A) Bounding $\Pr(\Omega_2)$: With $\text{SVD}(\mathbf{B}^*) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$, we can verify that

$$\begin{aligned}
\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}^2 &= \left\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}\right\|_{\text{F}}^2 \\
&=\ \|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{U}\mathbf{\Sigma}\|_{\text{F}}^2 = \left\|(\mathbf{I}-\mathbf{\Pi})\widetilde{\mathbf{X}}\mathbf{\Sigma}\right\|_{\text{F}}^2,
\end{aligned}$$

where $\widetilde{\mathbf{X}} \triangleq \mathbf{X}\mathbf{U}$. Due to the rotational invariance of the Gaussian distribution, $\widetilde{\mathbf{X}}$ has the same distribution $\mathbf{X}$.

When $h = 2$, we assume w.l.o.g. that the first row and second row are permuted. Then we have

$$\begin{aligned}
&\Pr\left(\left\|(\mathbf{I}-\mathbf{\Pi})\widetilde{\mathbf{X}}\mathbf{\Sigma}\right\|_{\text{F}}^2 \leq 2\|\mathbf{B}^*\|_{\text{F}}^2\right) \\
&=\ \Pr\left(2\sum_{i=1}^{r}\beta_i^2\left(\widetilde{X}_{1,i} - \widetilde{X}_{2,i}\right)^2 \leq 2\left(\sum_{i=1}^{r}\beta_i^2\right)\right) \\
&\overset{(ii)}{=}\ \Pr\left(\sum_{i=1}^{r}\beta_i^2\widetilde{z}_{1,i}^2 \leq \frac{\sum_{i=1}^{r}\beta_i^2}{2}\right) \\
&\overset{(iii)}{=}\ \Pr\left(\left\langle\widetilde{\mathbf{z}},\ \mathbf{\Sigma}^2\widetilde{\mathbf{z}}\right\rangle \leq \frac{\sum_{i=1}^{r}\beta_i^2}{2}\right) \overset{(iv)}{\leq} 2\exp\left(-c_0\varrho(\mathbf{B}^*)\right), \quad (40)
\end{aligned}$$

where $\mathbf{\Sigma} = \text{diag}\,(\beta_1, \cdots, \beta_r, 0, \cdots)$, β_i denotes the i-th singular values of $\mathbf{B}^*$, $\widetilde{X}_{i,j}$ denotes the (i,j) element of $\widetilde{\mathbf{X}}$, in (ii) we define $\widetilde{z}_{1,i} = (\widetilde{X}_{1,i} - \widetilde{X}_{2,i})/\sqrt{2}$, in (iii) we define $\widetilde{\mathbf{z}}$ as the vectorized version, and $\mathbb{E}\left\langle \widetilde{\mathbf{z}}, \mathbf{\Sigma}^2 \widetilde{\mathbf{z}} \right\rangle = \sum_{i=1}^r \beta_i^2$, and in (iv) we use Theorem 2.5 in [27] (cf. also Appendix H) and c_0 is the corresponding constant.

Next, we consider the case where $h \geq 3$, by studying the index set $\mathscr{I} \triangleq \{j : \pi(j) \neq j\}$, where $\pi(\cdot)$ is the permutation corresponding to the permutation matrix $\mathbf{\Pi}$. Adopting the same argument as in Lemma 8 in [6], we decompose the index set $\mathscr{I}$ into 3 subsets $\{\mathscr{I}_1, \mathscr{I}_2, \mathscr{I}_3\}$, such that

- $\sum_{i=1}^3 |\mathscr{I}_i| = h$ with $|\mathscr{I}_i| \geq \lfloor h/3 \rfloor, 1 \leq i \leq 3$.
- For arbitrary j, the indices j and $\pi(j)$ will not be in the same index set $\mathscr{I}_i$, $(1 \leq i \leq 3)$ at the same time.

We define a matrix $\mathbf{Z}_i$ which consists of the rows $(\mathbf{I} - \mathbf{\Pi})\widetilde{\mathbf{X}}\mathbf{\Sigma}$ corresponding to indices in $\mathscr{I}_i$. Accordingly, we can verify that $\left\|(\mathbf{I} - \mathbf{\Pi})\widetilde{\mathbf{X}}\mathbf{\Sigma}\right\|_{\text{F}}^2 = \sum_{i=1}^3 \|\mathbf{Z}_i\|_{\text{F}}^2$. Let h_i denote the corresponding cardinality of $|\mathscr{I}_i|$, $i = 1, 2, 3$. We have

$$\Pr\left(\|(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}^2 \leq h\|\mathbf{B}^*\|_{\text{F}}^2\right) \leq \sum_{i=1}^3 \Pr\left(\|\mathbf{Z}_i\|_{\text{F}}^2 \leq h_i\|\mathbf{B}\|_{\text{F}}^2\right).$$

In the sequel, we bound $\Pr\left(\|\mathbf{Z}_1\|_{\text{F}}^2 \leq h_1\|\mathbf{B}\|_{\text{F}}^2\right)$; the other two probabilities can be bounded similarly. Since j and $\pi(j)$ cannot be in $\mathscr{I}_1$ simultaneously, we define $\widetilde{z}_{j,k} = (\widetilde{X}_{j,k} - \widetilde{X}_{\pi(j),k})/\sqrt{2}$, $j \in \mathscr{I}_1, 1 \leq k \leq r$, and can treat the $\{\widetilde{z}_{jk}\}$ as independent $\mathscr{N}(0,1)$-RVs. Similar to the case $h = 2$, we have

$$\begin{aligned}
&\Pr\left(\|\mathbf{Z}_1\|_{\text{F}}^2 \leq h_1\|\mathbf{B}\|_{\text{F}}^2\right) \\
=\ &\Pr\left(\left\langle \widetilde{\mathbf{z}}, \text{diag}(\underbrace{\mathbf{\Sigma}^2, \cdots, \mathbf{\Sigma}^2}_{h_1 \text{ terms}})\widetilde{\mathbf{z}} \right\rangle \overset{(v)}{\leq} \frac{h_1}{2}\left(\sum_{i=1}^r \beta_i^2\right)\right) \\
\overset{(vi)}{\leq}\ &2\exp\left(-c_0 h_1 \varrho(\mathbf{B}^*)\right) \overset{(vii)}{\leq}\ 2\exp\left(-\frac{c_0}{5} h \varrho(\mathbf{B}^*)\right), \quad (41)
\end{aligned}$$

where in (v) we define $\widetilde{\mathbf{z}}$ as the vectorization of $\mathbf{Z}_1$, in (vi) we use Theorem 2.5 in [27] (can also be found in Appendix H), and in (vii) we use the fact $h_i \geq \lfloor h/3 \rfloor$. Combing the above cases in Eq. (40) and Eq. (41), we can bound $\Pr(\Omega_2) \leq 6\exp\left(-c_0 h \varrho(\mathbf{B}^*)/5\right)$.

Case I.(B) Bounding Pr $\left(\Omega_1 \bigcap \overline{\Omega}_2\right)$: For ease of notation, we define $\mathbf{\Theta} = (\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*/\|(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}$. Then the probability of the event $\Omega_1 \bigcap \overline{\Omega}_2$ can be bounded as

$$\begin{aligned}
&\Pr\left(\Omega_1 \bigcap \overline{\Omega}_2\right) \leq \Pr\left(\left\|P_{\mathbf{X}}^{\perp}\mathbf{\Theta}\right\|_{\text{F}}^2 < \frac{t}{h}\right) \\
=\ &\Pr\left(\left\|P_{X}^{\perp}\mathbf{\Theta}\right\|_{\text{F}}^2 < \frac{t}{h}\|\mathbf{\Theta}\|_{\text{F}}^2\right) \\
=\ &\Pr\left(\sum_{i=1}^r \left\|P_{\mathbf{X}}^{\perp}\mathbf{\Theta}_{:,i}\right\|_{\text{F}}^2 \leq \frac{t}{h}\|\mathbf{\Theta}_{:,i}\|_{\text{F}}^2\right) \\
\leq\ &\sum_{i=1}^r \Pr\left(\left\|P_{\mathbf{X}}^{\perp}\mathbf{\Theta}_{:,i}\right\|_{\text{F}}^2 \leq \frac{t}{h}\|\mathbf{\Theta}_{:,i}\|_{\text{F}}^2\right) \\
\overset{(a)}{=}\ &\sum_{i=1}^r \Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_i\right\|_2^2 \leq \frac{t}{h}\right),
\end{aligned}$$

where in (a) we define $\boldsymbol{\theta}_i$ as the normalized version of $\mathbf{\Theta}_{:,i}$, namely, $\mathbf{\Theta}_{:,i}/\|\mathbf{\Theta}_{:,i}\|_2$. Here, we define the set Θ_h by

$$\begin{aligned}
\Theta_h = \ &\big\{\boldsymbol{\theta} \in \mathbb{R}^n \mid \|\boldsymbol{\theta}\|_2 = 1, \\
&\boldsymbol{\theta} \text{ has at most } h \text{ non-zero elements}\big\}.
\end{aligned}$$

We can verify that $\boldsymbol{\theta}_i \in \Theta_h$ for $1 \leq i \leq r$, since $d_H(\mathbf{I}, \mathbf{\Pi}) = h \geq 2$. Before delving into detailed calculations, we first summarize our proof strategy:

- **Step I.(B). [i]**: We cover the set Θ_h with a δ-net $\mathscr{N}_\delta$ such that for arbitrary $\boldsymbol{\theta} \in \Theta_h$, there exists a $\boldsymbol{\theta}_0 \in \mathscr{N}_\delta$ such that $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_2 \leq \delta$.
- **Step I.(B). [ii]**: Define events Ω_Θ and $\Omega_{\mathscr{N}_\delta}$ by
$$\begin{aligned}
\Omega_\Theta &\triangleq \left\{\boldsymbol{\theta} \in \Theta_h \ \text{ s.t. } \ \left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}\right\|_2 < \sqrt{t/h}\right\} \\
\Omega_{\mathscr{N}_\delta} &\triangleq \left\{\boldsymbol{\theta}_0 \in \mathscr{N}_\delta \ \text{ s.t. } \ \left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 < 2\sqrt{t/h}\right\}.
\end{aligned}$$
Setting $\delta = \sqrt{t/h}$, we will prove that
$$\begin{aligned}
&\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\frac{(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*}{\|(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}}\right\|_{\text{F}}^2 < \frac{t}{h}\right) \\
\leq\ &r\Pr\left(\Omega_\Theta\right) \leq r\Pr\left(\Omega_{\mathscr{N}_\delta}\right).
\end{aligned}$$
- **Step I.(B). [iii]**: We consider an arbitrary fixed element $\boldsymbol{\theta}_0 \in \mathscr{N}_\delta$, and study $\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 \leq 2\sqrt{t/h}\right)$. Adopting the union bound
$$\Pr\left(\Omega_{\mathscr{N}_\delta}\right) \leq |\mathscr{N}_\delta| \times \Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 \leq 2\sqrt{t/h}\right),$$
we finish the bound of $\Pr\left(\Omega_1 \bigcap \overline{\Omega}_2\right)$.

The following analysis fills in the details.

Step I.(B). [i]: We cover the set Θ_h with a δ-net $\mathscr{N}_\delta$. Its cardinality can be bounded as

$$|\mathscr{N}_\delta| \overset{(b)}{\leq} \left(1 + \frac{2}{\delta}\right)^h \overset{(c)}{\leq} \left(\frac{3}{\delta}\right)^h,$$

where in (b) we (i) use that elements of Θ_h have at least $(n - h)$ zero elements, and accordingly we cover the sphere $\mathbb{S}^{h-1}$ with a δ-net $\mathscr{N}_\delta$, whose cardinality can be bounded as in [28], and in (c) we assume that $\delta \leq 1$.

Step I.(B). [ii] We will prove the relation

$$\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\frac{(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*}{\|(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\text{F}}}\right\|_{\text{F}}^2 < \frac{t}{h}\right) \overset{(d)}{\leq} r\Pr\left(\Omega_\Theta\right) \overset{(e)}{\leq} r\Pr\left(\Omega_{\mathscr{N}_\delta}\right),$$

when $\delta = \sqrt{t/h}$ and (d) follows from the definition of Ω_Θ. We here focus on proving inequality (e), which is done by

$$\begin{aligned}
&\Pr\left(\Omega_\Theta\right) = \Pr\left(\Omega_\Theta \bigcap \Omega_{\mathscr{N}_\delta}\right) + \Pr\left(\Omega_\Theta \bigcap \overline{\Omega}_{\mathscr{N}_\delta}\right) \\
\leq\ &\Pr\left(\Omega_{\mathscr{N}_\delta}\right) + \Pr\left(\Omega_\Theta \bigcap \overline{\Omega}_{\mathscr{N}_\delta}\right) \overset{(f)}{=} \Pr\left(\Omega_{\mathscr{N}_\delta}\right),
\end{aligned}$$

where (f) is due to the fact $\Pr\left(\Omega_\Theta \bigcap \overline{\Omega}_{\mathscr{N}_\delta}\right) = 0$. Note that, given $\overline{\Omega}_{\mathscr{N}_\delta}$, it holds that for all $\boldsymbol{\theta}_0 \in \mathscr{N}_\delta$, we have $\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 \geq 2\sqrt{t/h}$. Then for arbitrary $\boldsymbol{\theta} \in \Theta_h$, we consider an element $\boldsymbol{\theta}_0 \in \mathscr{N}_\delta$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \delta$ and consequently

$$\begin{aligned}
\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}\right\|_2 &\geq \left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 - \left\|P_{\mathbf{X}}^{\perp}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right\|_2 \\
&\geq\ 2\sqrt{t/h} - \left\|P_{\mathbf{X}}^{\perp}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right\|_2 \\
&\overset{(g)}{\geq}\ 2\sqrt{t/h} - \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \overset{(h)}{\geq} \sqrt{t/h},
\end{aligned}$$

where in (g) we use the contraction property of projections, and in (h) we use the fact $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2 \le \delta = \sqrt{t/h}$.

Step I.(B). [iii] We study the probability $\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2^2 \le \frac{t}{h}\right)$ for fixed $\boldsymbol{\theta}_0 \in \mathscr{N}_\delta$. In virtue of results in [29], we have

$$\begin{aligned}&\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2^2 \le \frac{\alpha(n-p)}{n}\|\boldsymbol{\theta}_0\|_2^2\right)\\ \le\ &\exp\left(\frac{n-p}{2}\left(\log(\alpha)-\alpha+1\right)\right),\quad \alpha\le 1.\end{aligned}$$

We can set $\alpha = 4nt/((n-p)h)$ (<1) and obtain

$$\begin{aligned}&\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2^2 \le \frac{4t}{h}\right)\\ =\ &\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2^2 \le \frac{\alpha(n-p)}{n}\right)\\ \le\ &\exp\left(\frac{n-p}{2}\left(\log\left(\frac{4nt}{(n-p)h}\right)-\frac{4nt}{(n-p)h}+1\right)\right)\\ \overset{(j)}{\le}\ &\exp\left(\frac{n}{4}\left(\log\left(\frac{8t}{h}\right)-\frac{8t}{h}+1\right)\right),\end{aligned}$$

where in (j) we use that (i) $n \ge 2p$, (ii) $\log(x)-x+1$ is increasing in range $(0,1)$, and (iii) $\log(x)+1 \le x$.

In the end, we bound $\Pr\left(\Omega_1 \bigcap \overline{\Omega}_2\right)$ as

$$\begin{aligned}&\Pr\left(\Omega_1 \bigcap \overline{\Omega}_2\right)\\ \le\ &r\left(\frac{3}{\sqrt{t/h}}\right)^h \exp\left(\frac{n}{4}\left(\log\left(\frac{8t}{h}\right)-\frac{8t}{h}+1\right)\right)\\ =\ &\exp\Bigg(h\log(3)-\frac{h}{2}\log\left(\frac{t}{h}\right)+\log(r)\\ &\qquad+\frac{n}{4}\left(\log\left(\frac{8t}{h}\right)-\frac{8t}{h}+1\right)\Bigg)\\ \overset{(k)}{\le}\ &\exp\left(\frac{rh}{2}\left(\log\left(\frac{t}{h}\right)-\frac{16t}{h}+1\right)+3.68rh+\log(r)\right)\\ \overset{(\ell)}{\le}\ &\exp\left(\frac{rh}{2}\left(\log\left(\frac{t}{h}\right)-\frac{t}{h}+1\right)+4.18rh\right)\end{aligned}\tag{42}$$

where in (k) we use the assumption that $n \ge 4rh$, and in (ℓ) we use that $rh \ge 2r \ge 2\log(r)$.

Combining Eq. (40), Eq. (41) and Eq. (42), we finish the proof.

Case II: Here we have $T_{\mathbf{\Pi}} > t\|\mathbf{B}^*\|_{\mathrm{F}}^2$. Using the same argument as in proving Lemma. 12, we can prove that

$$\begin{aligned}&\Pr\left(\left\|P_{\mathbf{\Pi}\mathbf{X}}^{\perp}\mathbf{Y}\right\|_{\mathrm{F}}^2 \le \left\|P_{\mathbf{\Pi}^*\mathbf{X}}^{\perp}\mathbf{Y}\right\|_{\mathrm{F}}^2, T_{\mathbf{\Pi}} > t\|\mathbf{B}^*\|_{\mathrm{F}}^2\right)\\ \le\ &2\exp\left(-\frac{mt\times\mathrm{snr}}{288}\left(\frac{t\times\mathrm{snr}}{h}\wedge 1\right)\right)+\exp\left(-\frac{mt\times\mathrm{snr}}{72}\right).\end{aligned}$$

We complete the proof by combining Case I and Case II. □

Remark 15. *Note that we cannot improve h from $\mathscr{O}\left(\frac{n}{\log(n)}\right)$ to n in general, since there is an inherent problem when dealing with the case $h \to n$. A detailed explanation is given as the following. The key ingredient in bounding* $\Pr\left(\Omega_1 \bigcap \overline{\Omega}_2\right)$ *is based on the step*

$$\begin{aligned}&\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\frac{(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}}{\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}\|_{\mathrm{F}}}\right\|_{\mathrm{F}} \le \sqrt{\frac{t}{h}}\right)\\ \le\ &\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}\right\|_2 \le \sqrt{\frac{t}{h}},\ \ \exists\ \boldsymbol{\theta}\in\Theta_h\right)\\ \le\ &|\mathscr{N}_\delta| \times \Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 \le \sqrt{\frac{t}{h}}+\delta,\ \ \exists\ \boldsymbol{\theta}\in\mathscr{N}_\delta\right) < 1.\end{aligned}$$

For the extreme case when $h = n$, we cannot have $|\mathscr{N}_\delta|\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 \le \sqrt{t/h}+\delta,\ \ \exists\boldsymbol{\theta}\in\mathscr{N}_\delta\right) < 1$ since

$$\begin{aligned}&\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}\right\|_2 \le \sqrt{\frac{t}{h}},\ \ \exists\ \boldsymbol{\theta}\in\Theta_n\right)\\ \ge\ &\Pr\left(\left\|P_{\mathbf{X}}^{\perp}\frac{\mathbf{X}\mathbf{B}^*}{\|\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}}\right\|_{\mathrm{F}} \le \sqrt{\frac{t}{h}}\right) = 1.\end{aligned}$$

The reason behind this is that we lose control of the cardinality $|\mathscr{N}_\delta| \lesssim (C/\delta)^{rh}$ when $h \to n$.

Lemma 16. *Given that* $\mathrm{snr} > 26.2$*, $rh \le n/4$, $t \le 0.125h$, $\varrho(\mathbf{B}^*) \ge 5(1+\epsilon)\log(n)/c_0$, and*

$$\log(\mathrm{snr}) \ge \frac{288(1+\epsilon)\log(n)}{\varrho(\mathbf{B}^*)} + 33.44,$$

we have

$$\begin{aligned}&6\exp\left(-\frac{c_0 h\varrho(\mathbf{B}^*)}{5}\right)\\ +\ &\exp\left(\frac{rh}{2}\left(\log\left(\frac{t}{h}\right)-\frac{t}{h}+1\right)+4.18rh\right)\\ +\ &2\exp\left(-\frac{mt\times\mathrm{snr}}{288}\left(\frac{t\times\mathrm{snr}}{h}\wedge 1\right)\right)\\ +\ &\exp\left(-\frac{mt\times\mathrm{snr}}{72}\right) \le 10n^{-(1+\epsilon)h},\end{aligned}$$

where $c_0, \epsilon > 0$ are positive constants.

Proof. Here we choose $t = h\log(\mathrm{snr})/\mathrm{snr}$. Note that if $\mathrm{snr} > 26.2$, we have $t < 0.125h$. Given Eq. (14), we have

$$\log(\mathrm{snr}) \ge \frac{288(1+\epsilon)\log(n)}{\varrho(\mathbf{B}^*)} \overset{(i)}{\ge} \frac{288(1+\epsilon)\log(n)}{m},\tag{43}$$

where in (i) we use $\varrho^*(\mathbf{B}) \le r(\mathbf{B}^*) \le m$. In the sequel we will separately bound the following terms

- $\mathscr{T}_1 \triangleq \exp\left(-c_0 h\varrho(\mathbf{B}^*)/5\right)$.
- $\mathscr{T}_2 \triangleq \exp\left(\frac{rh}{2}\left(\log\left(\frac{t}{h}\right)-\frac{t}{h}+1\right)+4.18rh\right)$.
- $\mathscr{T}_3 \triangleq \exp\left(-\frac{mt\times\mathrm{snr}}{288}\left(\frac{m\times\mathrm{snr}}{h}\wedge 1\right)\right)$.
- $\mathscr{T}_4 \triangleq \exp\left(-mt\times\mathrm{snr}/72\right)$.

Term $\mathscr{T}_1$: If $\varrho(\mathbf{B}^*)$ satisfies $\varrho(\mathbf{B}^*) \ge 5(1+\epsilon)\log(n)/c_0$, we can easily show that

$$e^{-c_0 h\varrho(\mathbf{B}^*)/5} \le n^{-(1+\epsilon)h}.\tag{44}$$

Term $\mathscr{T}_2$: Then we bound

$$\mathscr{T}_2 \overset{(ii)}{\le} \exp\left(-\frac{rh}{8}\log(\mathrm{snr})+4.18rh\right) \overset{(iii)}{\le} n^{-(1+\epsilon)h},\tag{45}$$

where in (ii) we use $\frac{\log z}{z}-1-\log\frac{\log z}{z} \ge \frac{\log z}{4}$, for $z \ge 1.5$, and in (iii) we use the assumption such that

$$\log(\mathrm{snr}) \ge \frac{8(1+\epsilon)\log(n)}{\varrho(\mathbf{B}^*)} + 33.44.$$

Term $\mathscr{T}_3$: Since we have $\mathrm{snr} \geq 26.2$, we obtain

$$\left(\frac{mt^2 \times \mathrm{snr}^2}{h} \wedge (mt \times \mathrm{snr})\right) \geq mh \log(\mathrm{snr}).$$

We then have

$$\mathscr{T}_3 \leq \exp\left(-\frac{mh}{288} \times \log(\mathrm{snr})\right) \overset{(iv)}{\leq} n^{-(1+\epsilon)h}, \tag{46}$$

where in (iv) we use Eq. (43).

Term $\mathscr{T}_4$: We have

$$\exp\left(-\frac{mt \times \mathrm{snr}}{72}\right) = \exp\left(-\frac{mh}{72}\log(\mathrm{snr})\right) \overset{(v)}{\leq} n^{-(1+\epsilon)h}, \tag{47}$$

where in (v) we use Eq. (43). Combining Eq. (44), Eq. (45), Eq. (46), and Eq. (47), we finish the proof. □

Appendix H
Probability inequalities

This section collects some useful probability inequalities.

Lemma 17 (Thm. 2.5 in [27])**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a non-zero matrix and let $\boldsymbol{\xi} = (\xi_i)_{i=1}^n$ be a random vector with independent sub-Gaussian entries such that (i) $\mathrm{var}(\xi_i) \geq 1$, $1 \leq i \leq n$, and (ii) the sub-Gaussian constant of the $\{\xi_i\}$ is at most β. Then $\forall \mathbf{y} \in \mathbb{R}^n$, there exists a $c_0 > 0$ such that*

$$\Pr\left(\|\mathbf{y} - \mathbf{A}\boldsymbol{\xi}\|_2 \leq \frac{\|\mathbf{A}\|_{\mathrm{F}}}{2}\right) \leq 2\exp\left(-\frac{c_0}{\beta^4}\varrho(\mathbf{A})\right).$$

Lemma 18 (Lemma 2.6 in [27])**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a non-zero matrix and $\mathbf{g}$ be Gaussian $\mathscr{N}(\underline{0}, \mathbf{I}_{n\times n})$. Then we have*

$$\Pr\left(\|\mathbf{y} - \mathbf{A}\mathbf{g}\|_2 \leq \alpha\|\mathbf{A}\|_{\mathrm{F}}\right) \leq \exp\left(\kappa\log(\alpha)\varrho(\mathbf{A})\right),$$

for any $\alpha \in (0, \alpha_0)$, where $\mathbf{y} \in \mathbb{R}^n$ is an arbitrary fixed vector, $\alpha_0 \in (0,1)$ and $\kappa > 0$ are universal constants.

Lemma 19 ([28] (Example 2.11, P29))**.** *For a χ^2-RV Y with ℓ degrees of freedom, we have*

$$\Pr\left(|Y - \ell| \geq t\right) \leq 2\exp\left(-\left(\frac{t^2}{8\ell} \wedge \frac{t}{8}\right)\right), \quad \forall\, t \geq 0.$$

Appendix I
Appendix of Information Theory

To make the paper self-contained, we provide a review of basic tools from information theory used herein [23].

Definition (Discrete entropy (Page 14, [23]))**.** The entropy of a discrete RV X is defined as

$$\begin{aligned} H(X) &= -\sum_x \Pr(X = x)\, \log \Pr(X = x) \\ &= \mathbb{E}_X \phi(X), \quad \text{where } \phi(x) \triangleq -\log(x). \end{aligned}$$

Definition (Conditional entropy (Page 17, [23]))**.** For a pair of discrete RV (X, Y), the conditional entropy $H(Y|X)$ is

$$\begin{aligned} H(Y|X) &= -\sum_{x,y} \Pr(X = x, Y = y) \log \Pr(Y = y | X = x) \\ &= \mathbb{E}_{X,Y}\ \phi(X, Y), \end{aligned}$$

where $\phi(x, y) \triangleq -\log \Pr\left(Y = y \mid X = x\right)$.

One can difficult to check that $H(Y|X)$ can also be defined as $\mathbb{E}_X \psi_Y(X)$, where $\psi_Y(x) \triangleq H(Y|X = x)$ and $H(Y|X = x)$ denotes the discrete entropy of the random variable Y conditional on $\{X = x\}$.

Property 20 (Page 41-43, [23])**.** *Important properties of the entropy are:*

- *We have $H\left(X|Y\right) \leq H(X)$, with equality being achieved when X and Y are independent.*
- *For a sequence of discrete RVs $\{X_i\}_{1\leq i\leq N}$, we have*

$$H\left(X_1, \cdots, X_N\right) = H(X_1) + \sum_{i=2}^N H\left(X_i \mid X_1, \cdots, X_{i-1}\right),$$

which is known as the chain rule *of entropy.*

Definition (Mutual information (Page 20, [23]))**.** For a pair of discrete RV (X, Y), the mutual entropy $I(X;Y)$ is defined as

$$I(X;Y) = H(X) - H(X|Y),$$

where $H(\cdot)$ and $H(\cdot|\cdot)$ represent the entropy and conditional entropy, respectively.

Theorem 21 (Data processing inequality (Theorem 2.5.1, Page 34, [23]))**.** *If X, Y, Z forms a Markov chain such that $X \to Y \to Z$, we have*

$$I(X;Y) \geq I(X;Z).$$

Theorem 22 (Fano's inequality (Theorem 2.10.1, Page 38, [23]))**.** *Consider a discrete RV X with alphabet $\mathscr{X}$, then for any estimator $Y(\cdot)$ such that $\widehat{X} = Y(X)$, we have*

$$H(X|Y) \leq (\log|\mathscr{X}|) \times \Pr\left(\widehat{X} \neq X\right) + 1,$$

where $|\mathscr{X}|$ denotes the cardinality of $\mathscr{X}$.

Next, we consider continuous random variables. Denote $f_X(\cdot)$ as the probability density of X, $f_{X,Y}(\cdot,\cdot)$ as the probability density for (X, Y), and $f_{X|Y}(\cdot|\cdot)$ as the density for the conditional distribution. We have

Definition (Differential entropy (Page 243, [23]))**.** The differential entropy $h(X)$ of a continuous RV X is defined as

$$h(X) = -\int f_X(x)\log f_X(x)dx.$$

Definition (Conditional differential entropy (Page 249, [23]))**.** The conditional entropy $h(X|Y)$ of continuous RVs X, Y can be written as

$$h(X|Y) \triangleq -\int f_{X,Y}(x,y)\log f_{X|Y}(x|y)dxdy.$$

Definition (Mutual information (Page 251, [23]))**.** The mutual entropy $I(X;Y)$ between the continuous RV X, Y is defined as

$$\begin{aligned} I(X;Y) &\triangleq h(X) - h(X|Y) \\ &= \int f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}dxdy. \end{aligned}$$