The Benefits of Diversity: Permutation Recovery in Unlabeled Sensing From Multiple Measurement Vectors

Hang Zhang, Martin Slawski, and Ping Li, *Member, IEEE*

***Abstract*— 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 methods for recovering the unknown permutation. For the "oracle case" with known signal, the maximum likelihood (ML) estimator reduces to a linear assignment problem whose global optimum can be obtained efficiently. If both the signal and the permutation are unknown, the problem becomes a quadratic assignment problem; while such a problem is generally NP-hard and hence poses a significant challenge, we propose to tackle it via projected gradient descent with a non-convex constraint set, and establish a monotonic descent property of this scheme. Numerical experiments based on the proposed computational approach confirm the tightness of our theoretical analysis.**

***Index Terms*— Regression analysis (under statistical analysis), optimal matching (combinatorial mathematics).**

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 m 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 m 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, as can be seen from selected example applications sketched below that motivate the setting (2). It is worth emphasizing that the latter assumes that the permutation is shared across the m sets of measurements, and hence does not apply to situations in which each of those involves its individual permutation.

Header-Free Communication: As discussed, e.g., in [2], [3], in sensor networks with stringent requirements concerning latency and communication footprint, it can be beneficial to omit sensor metadata when transmitting measurements to the fusion center in an effort to minimize latency and communication cost. In this case, signal recovery without metadata such as sensor identifiers involves an unknown permutation.

Post-Linkage Data Analysis: It is often much more cost-efficient to combine data from existing databases rather than collecting new data containing all variables of interest. Due to data formatting and data quality issues, linkage of records pertaining to the same entity can be error-prone. As a result, downstream data analysis such as linear regression or estimation of the cross-covariance between $\mathbf{X}$ and $\mathbf{Y}$ can be affected, and modeling mismatches via a permutation has been studied recently as a mitigation strategy [4].

Data Privacy: In linkage attacks, intruders aim at the disclosure of sensitive data by using external data and record linkage. There is a long history of attacks in which public data was combined with de-identified data to reveal sensitive information [5], [6]. Those examples involve direct comparison of

Manuscript received September 24, 2019; revised March 27, 2021; accepted July 19, 2021. Date of publication November 9, 2021; date of current version March 17, 2022. The work of Martin Slawski was supported in part by NSF Grant CCF1849876. An earlier version of this paper was presented in part at the 2019 IEEE International Symposium on Information Theory. *(Corresponding author: Ping Li.)*

The authors are with the Cognitive Computing Laboratory, Baidu Research, Bellevue, WA 98004 USA (e-mail: pingli98@gmail.com).

Communicated by M. Rodrigues, Associate Editor for Signal Processing.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3127072.

Digital Object Identifier 10.1109/TIT.2021.3127072

two datasets $\mathbf{\Pi}^*\mathbf{X}$ and $\mathbf{Y}$; the regression setup (2) arises as a natural generalization.

Unsupervised Alignment: Aligning two sets of points is a fundamental task with applications in computer vision, curve registration, and natural language processing. A problem of recent interest is the alignment of embeddings of text corpora into the unit sphere $\mathbb{S}^{p-1}$ in $\mathbb{R}^p$ [7], [8]. For example, [7] formulates the automated translation between different versions of medical diagnosis codes used in electronic health systems as a problem of the form (2) given two sets of vectors $\mathbf{X}$ and $\mathbf{Y}$ in $\mathbb{S}^{p-1}$ representing embeddings of two different versions of medical diagnosis codes.

Additional examples can be found among the references provided in the next section.

A. Related Work

The work [9] discusses *signal recovery* under setup (2) dubbed "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 [9] are shown in [10], [11].

A number of recent papers discuss the case $m = 1$ and Gaussian $\mathbf{W}$. The paper [12] establishes the statistical limits of exact and approximate permutation recovery based on the ratio of signal energy and noise variance henceforth referred to as "$\mathbf{snr}$". In [12], it is also demonstrated that the least squares estimation of $\mathbf{\Pi}^*$ is NP-hard in general. In [13], a polynomial-time approximation algorithm is proposed, and lower bounds on the required $\mathbf{snr}$ for approximate signal recovery in the noisy case are shown; related results can be found in [4], [14]. The works [4], [7], [15], [16] 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 [17]. The papers [18], [19] develop the approach in [17] further by combining it with a careful branch-and-bound scheme to solve general unlabeled sensing problems. In [20], sparsity assumption is put on the vector $\mathbf{B}^*$ and the necessary condition $n \geq 2p$ for correct signal recovery is relaxed to $n \ll p$.

Several papers [2], [15], [16], [21], [22] have studied the setting of multiple measurement vectors ($m \geq 2$) and associated potential benefits for permutation recovery. The paper [21] discusses a practical branch-and-bound scheme for permutation recovery but does not provide theoretical insights. The work [2] 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 [15], [16], 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 [15], [16] 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 [15] limits the fraction of permuted rows to a constant multiple of the reciprocal of the signal dimension p, while the method in [16] 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 [15], [16] to a good extent.

B. Summary of Contributions

Results in [12] on the case $m = 1$ indicate that the *maximum likelihood* (ML) estimator in (5) can be regarded as impractical from both statistical and computational viewpoints. On one hand, exact recovery of $\mathbf{\Pi}^*$ requires $\mathbf{snr} = \Omega(n^c)$, 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 (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* ($\mathbf{snr}$)

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

where $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i,j} \mathbf{A}_{ij}^2}$ denotes the Frobenius norm of a matrix $\mathbf{A}$ of arbitrary dimension.

- We improve the requirement $\mathbf{snr} = \Omega(n^c)$ in [4], [12] to roughly $\mathbf{snr} = \Omega(n^{c/\varrho(\mathbf{B}^*)})$ (cf. Theorems 5 and 7), where $\varrho(\mathbf{B}^*) = \frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\|\mathbf{B}^*\|_{\mathrm{OP}}^2}$ is the so-called stable rank of $\mathbf{B}^*$, which is given by the squared ratio of the Frobenius norm and the operator norm $\|\cdot\|_{\mathrm{OP}}$ of a matrix, and constitutes a lower bound on its rank (e.g., Section 2.1.15 in [23]). Once $\varrho(\mathbf{B}^*)$ is of the order $\Omega(\log n)$, we notice that the $\mathbf{snr}$ is only required to be of the order $\Omega(1)$ and hence does no longer need to increase with n. The underlying intuition is that larger values of m lead to relaxed requirements on the $\mathbf{snr}$ since 1) the overall signal energy increases, 2) all MMVs result from the same permutation matrix $\mathbf{\Pi}^*$, which is expected to yield extra information. In our analysis, 1) is reflected by conditions on permutation recovery involving dependence on the overall signal energy, while 2) enters via a dependence on the stable rank $\varrho(\mathbf{B}^*)$ of the signal matrix $\mathbf{B}^*$.
- We verify that the theoretical results can be attained in practice. For this purpose, we develop a practical algorithm for recovery of $\mathbf{\Pi}^*$ and $\mathbf{B}^*$ via least squares fitting. This amounts to solving a quadratic assignment problem which is NP-hard except for the special case with $p = m = 1$. We propose to tackle this problem by means of a projected gradient descent algorithm. The resulting

TABLE I

OVERVIEW ON RESULTS IN RELATED WORK IN COMPARISON TO THOSE SHOWN HEREIN. THE COLUMN "$h_{\max}$" REFERS TO THE MAXIMUM HAMMING DISTANCE BETWEEN $\mathbf{\Pi}^*$ AND THE IDENTITY MATRIX WITH $h_{\max} = n$ REFERRING TO THE FULLY SHUFFLED CASE. "COMPUTABLE" REFERS TO THE AVAILABILITY OF PRACTICAL COMPUTATIONAL SCHEMES THAT ACHIEVE THE THEORETICAL GUARANTEES ESTABLISHED IN EACH WORK

Related work	**snr**	n/p	$h_{\max}$	Minimax Analysis	Computable
[9]–[11]	∞	$\Omega(1)$	n	×	×
[12]	$\Omega(n^c)$	$\Omega(1)$	n	✓	×
[13]	∞	$\Omega(1)$	n	×	✓
[4]	$\Omega(n^c)$	$\Omega(1)$	$O\left[\frac{n-p}{\log(n/h_{\max})}\right]$	×	✓
This work	$\Omega(n^{c/\varrho(\mathbf{B}^*)})$	$\Omega(1)$	$\Omega(\frac{n}{\log n})$	✓	✓
[15]	$\Omega(n^{c/\varrho(\mathbf{B}^*)})$	$\Omega(p \vee h_{\max})$	$\Omega\left[\frac{n}{\log(n/h_{\max})}\right]$	×	✓

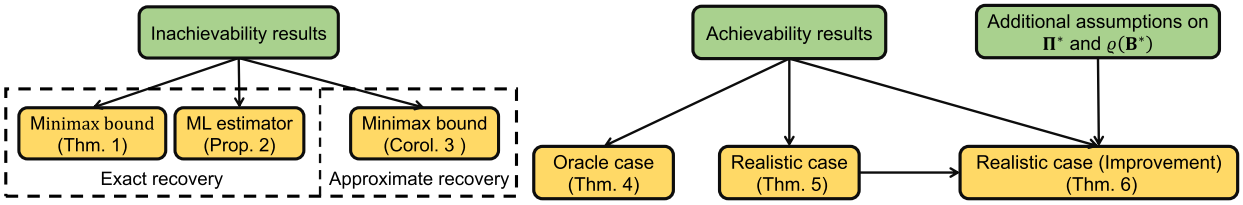

Fig. 1. A roadmap of the main results to be presented in the paper. Left panel: inachievability results (Section III); Right panel: achievability results (Section IV).

scheme is shown to exhibit monotonic descent, i.e., it generates a sequence of iterates with non-increasing objective, despite the non-convexity of the underlying constraint set. Extensive numerical results based on this approach align with our theorems and confirm significant reductions of the **snr** required for recovery of $\mathbf{\Pi}^*$ as the stable rank $\varrho(\mathbf{B}^*)$ of the signal matrix $\mathbf{B}^*$ increases.

We conclude this summary of contributions with an overview presented in Table I that compares the results herein to those obtained in related work.

C. Outline

The rest of the paper is organized as follows. The underlying sensing model is reviewed in Section II. In Section III, we establish conditions that imply failure of recovery (inachievability). This is followed by achievability results presented in Section IV and a discussion of their tightness in relation to the corresponding inachievability results. Our computational scheme based on projected gradient descent is presented in Section V. The empirical evaluation and concluding remarks are provided in Section VI and Section VII, respectively. A graphical representation of the structure of this paper is provided in Figure 1.

II. SYSTEM MODEL

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., $X_{ij} \sim \mathcal{N}(0,1)$, $1 \leq i \leq n,\ 1 \leq j \leq p$. Likewise, we assume that the entries of $\mathbf{W}$ are i.i.d. $\mathcal{N}(0,\sigma^2)$-random variables, where $\sigma^2 > 0$ denotes the noise variance. The ML estimator of $(\mathbf{\Pi}^*, \mathbf{B}^*)$ then results as the least squares solution

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

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

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

where the superscript † denotes the generalized inverse. 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 usual least squares estimator as in the absence of $\mathbf{\Pi}^*$. In the following, we put estimation of $\mathbf{B}^*$ aside and concentrate on analyzing the determining factors for recovery of $\mathbf{\Pi}^*$. Broadly speaking, this task involves two main sources of difficulty.

- *Sensing noise* $\mathbf{W}$. In the *oracle case* in which $\mathbf{B}^*$ is known, computation of the ML estimator of $\mathbf{\Pi}^*$ reduces to the *linear assignment problem* [24]

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

  where $\langle \mathbf{U}, \mathbf{V} \rangle = \operatorname{trace}(\mathbf{U}^{\top}\mathbf{V})$ here refers to the inner product between matrices $\mathbf{U}$ and $\mathbf{V}$ that induces the Frobenius norm. Even though the solution of (7) can be obtained efficiently by solving a linear program, recovery of $\mathbf{\Pi}^*$ is still likely to fail if the noise level σ^2 is large enough.
- *Unknown* $\mathbf{B}^*$. In contrast to the oracle case above, we have no access to $\mathbf{B}^*$ in practice, which suggests that recovery becomes more challenging.

In the sequel, we will show that the sensing noise $\mathbf{W}$ constitutes the major difficulty in recovering $\mathbf{\Pi}^*$ rather than the missing knowledge of $\mathbf{B}^*$. Before delving into our main results, we first 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 $a \asymp b$. For two numbers a and b, we let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we denote $\mathbf{A}_{:,i} \in \mathbb{R}^n$ as the i^{th} column of $\mathbf{A}$ while $\mathbf{A}_{i,:}$ denotes its i^{th} row, viewed as a column vector. 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 [25] (Section 2.3, P. 71). The ratio $\varrho(\cdot) = \|\cdot\|_{\mathrm{F}}^2/\|\cdot\|_{\mathrm{OP}}^2$ represents the stable-rank while $r(\cdot)$ represents the (usual) rank. We denote the *singular value decomposition* (SVD) of the matrix $\mathbf{A}$ as $\mathrm{SVD}(\mathbf{A})$, whose definition can be found in [25] (Section 2.4, P. 76) and is listed in Appendix A as well. We let $\mathscr{P}_n$ denote the set of permutation matrices of size n. Associating each $\mathbf{\Pi} \in \mathscr{P}_n$ with a mapping $\boldsymbol{\pi}$ of $\{1, 2, \ldots, n\}$ which moves index i to $\boldsymbol{\pi}(i)$, $1 \leq i \leq n$, we define the Hamming distance $\mathsf{d_H}(\cdot;\cdot)$ between two permutation matrices as $\mathsf{d_H}(\mathbf{\Pi}_1; \mathbf{\Pi}_2) \triangleq \sum_{i=1}^n \mathbb{1}\left(\boldsymbol{\pi}_1(i) \neq \boldsymbol{\pi}_2(i)\right)$. The *signal-to-noise-ratio* ($\mathbf{snr}$) is defined as $\mathbf{snr} = \|\mathbf{B}^*\|_{\mathrm{F}}^2/(m\sigma^2)$. Additional notations can be found in Appendix A.

III. Inachievability Results

In this section, we present conditions under which exact and approximate recovery of $\mathbf{\Pi}^*$ would *fail* with high probability. To be specific, *exact recovery* refers to the event $\{\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\}$, and *approximate recovery* of $\mathbf{\Pi}^*$ within a Hamming ball of radius $\mathsf{D} \in \{0, 1, \ldots, n\}$ refers to the event $\{\mathsf{d_H}(\mathbf{\Pi}^*; \widehat{\mathbf{\Pi}}) = \sum_{i=1}^n \mathbb{1}\left(\boldsymbol{\pi}^*(i) \neq \widehat{\boldsymbol{\pi}}(i)\right) \leq \mathsf{D}\}$, where $\boldsymbol{\pi}^*$ and $\widehat{\boldsymbol{\pi}}$ denote the mappings associated with $\mathbf{\Pi}^*$ and $\widehat{\mathbf{\Pi}}$, respectively, $1 \leq i \leq n$. The investigation of these cases is intended to provide valuable insights into the fundamental statistical limits. In order to establish inachievability results, it suffices to consider the *oracle case* with $\mathbf{B}^*$ known. The resulting limits apply to the case of unknown $\mathbf{B}^*$ as well, since it is hopeless to recover $\mathbf{\Pi}^*$ even if knowledge of $\mathbf{B}^*$ does not suffice for recovery.

Compared with the case $m = 1$ in which $\mathbf{snr}$ is the only prominent factor in determining the recovery performance [12], our analysis uncovers another crucial factor, namely, the energy distribution over singular values of $\mathbf{B}^*$. Our work shows that a more uniform spread of the signal energy over singular values can greatly facilitate the recovery of $\mathbf{\Pi}^*$.

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

We start by presenting an inachievability result concerning exact recovery.

Theorem 1: Let $\mathscr{H}$ be any subset of $\mathscr{P}_n$. In the oracle case with known $\mathbf{B}^*$, we have

$$\inf_{\widehat{\mathbf{\Pi}}} \sup_{\mathbf{\Pi}^* \in \mathscr{H}} \mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*) \geq \frac{1}{2}$$
$$\text{if} \quad \log\det\left(\mathbf{I} + \frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right) < \frac{\log(|\mathscr{H}|) - 2}{n}, \tag{8}$$

where the probability $\mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\cdot)$ is w.r.t. $\mathbf{X}$ and $\mathbf{W}$, and the infimum is over all estimators $\widehat{\mathbf{\Pi}}$.

Proof outline: Given knowledge of $\mathbf{B}^*$, we view the sensing relation (4) as a process such that 1) $\mathbf{\Pi}^*$ is encoded via the codeword $\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*$ and 2) is passed through a Gaussian channel with additive noise $\mathbf{W}$. We complete the proof based on Fano's inequality following [26] (cf. Section 7.9, P. 206). The key technical contribution is the derivation of a tight upper bound on the conditional mutual information between $\mathbf{\Pi}^*$ and $\mathbf{Y}$ given $\mathbf{X}$ when $\mathbf{\Pi}^*$ is drawn uniformly at random from $\mathscr{H}$. □

Let us point out important implications of Theorem 1. When $\mathscr{H} = \mathscr{P}_n$, we have $\log|\mathscr{H}| = \log n! \approx n\log n$ and the condition in (8) simplifies as $\log\det(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2) \lesssim \log n$. With a smaller set $\mathscr{H}$, the inachievability condition (8) is less likely to be fulfilled. For example, consider the special case in which $\mathscr{H}$ is a Hamming ball around the identity, i.e., $\mathscr{H} = \{\mathbf{\Pi} \in \mathscr{P}_n : \ \mathsf{d_H}(\mathbf{I}; \mathbf{\Pi}) \leq \mathsf{D}\}$ for some fixed non-negative integer D. Then the condition in (8) reduces to $\log\det(\mathbf{I} + \mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2) \lesssim \mathsf{D}/(n - \mathsf{D}) \ll \log n$ when n is sufficiently large.

The second major ingredient in condition (8) is the term $\log\det(\mathbf{I} + \mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2) = \sum_i \log(1 + \lambda_i^2/\sigma^2)$, where λ_i denotes the i^{th} singular value of $\mathbf{B}^*$. Since each singular value λ_i is determined by the matrix $\mathbf{B}^*$ as a 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

$$\log\det\left(\mathbf{I} + \frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right) = \log\left(1 + \frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2}\right). \tag{9}$$

This indicates that in accordance with the intuition of "borrowing strength" across different sets of measurements, performance is expected to improve as the stable rank of $\varrho(\mathbf{B}^*)$ of $\mathbf{B}^*$ increases. To give an illustration of the benefits brought by large stable rank $\varrho(\mathbf{B}^*)$, we numerically evaluate the required $\mathbf{snr} = \|\mathbf{B}^*\|_{\mathrm{F}}^2/(m\sigma^2)$ for the leftmost quantity in (9) to exceed specific thresholds in dependence of selected choices of $\varrho(\mathbf{B}^*)$. The results are listed in Table II.

Example 2: In order to get a better sense of the scaling of $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2\right)}{\log n}$, we consider the case in which the entries of $\mathbf{B}^*$ are sampled i.i.d. from a Gaussian distribution with zero mean and variance p^{-1}, i.e., $B_{ij}^* \overset{i.i.d}{\sim} \mathscr{N}(0, p^{-1})$. We then have

$$\frac{\log\det\left(\mathbf{I} + \mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2\right)}{\log n} \asymp \frac{\varrho(\mathbf{B}^*)\log(1+\mathbf{snr})}{\log n}$$
$$\overset{①}{\approx} \frac{m\log\left(1+\sigma^{-2}\right)}{\left(1+\sqrt{m/p}\right)^2\log n}, \tag{10}$$

TABLE II

THE REQUIRED VALUES OF $\mathbf{snr} = \|\mathbf{B}^*\|_{\mathrm{F}}^2/(m\sigma^2)$ FOR THE CONDITION $\log\det\left(\mathbf{I}+\frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right) > c\cdot\log n$ TO HOLD, $c\in\{1,2,\ldots,6\}$, WHEN $n=1000$, $p=100$, AND $\mathbf{B}^*_{:,i}=\mathbf{e}_i$, WHERE $\mathbf{e}_i$ DENOTES THE i^{th} CANONICAL BASIS VECTOR, $1\le i\le m$, $m\in\{1,10,20,50,100\}$ (LEFTMOST COLUMN)

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

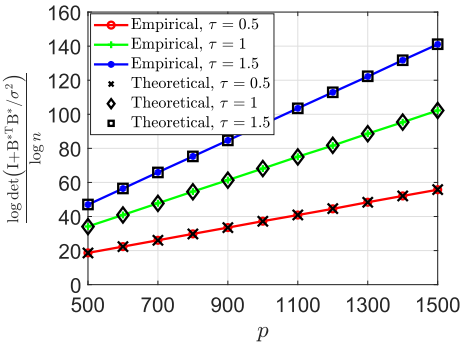

Fig. 2. Comparing the theoretical values (cf. (54)) with the empirical values of $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2\right)}{\log n}$ for $n=5000$, $\sigma^2=1$, and $m,p\to\infty$ with fixed ratio $m/p=\tau$.

where in ① we use the fact $|\|\mathbf{B}^*\|_{\mathrm{F}}-\sqrt{m}|\le\epsilon\sqrt{m}$ with probability exceeding $1-e^{-cm}$, and Bai-Yin's Law (see [27]), namely, $\|\mathbf{B}^*\|_{\mathrm{OP}}\approx 1+\sqrt{m/p}$. The expression on the right hand side of (10) aligns with the requirement $m\gtrsim\log n$ for achieving recovery. A more precise expression that yields the visualization in Figure 2 is deferred to the Appendix, cf. (54).

The statement below provides a condition for failure of recovery of $\mathbf{\Pi}^*$ when using the ML estimator in (5), which is computationally feasible if $\mathbf{B}^*$ is known.

Proposition 3: Let $\mathbf{\Pi}^*$ be an arbitrary element of $\mathscr{P}_n$. The ML estimator $\widehat{\mathbf{\Pi}}$ given in (7) under the oracle case where $\mathbf{B}^*$ is known, satisfies $\Pr_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*)\ge\frac{1}{2}$ for $n\ge 10$ if

$$\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2}\le\frac{2\log n}{4\left(1+c\varrho^{-1/2}(\mathbf{B}^*)\right)^2},\tag{11}$$

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

Proof outline: Without loss of generality, we may work with $\mathbf{\Pi}^*=\mathbf{I}$. We then use a direct argument involving corresponding rows of $\mathbf{Y}$ and $\mathbf{\Pi}^*\mathbf{X}$ and concentration of measure results to show that if (11) holds, $\widehat{\mathbf{\Pi}}$ cannot be $\mathbf{I}$ with the stated probability. □

The proposition states that the total signal energy given by $m\cdot\mathbf{snr}$ should be at least of the order $\log n$ to avoid failure in recovery. This is in agreement with Theorem 1 in the full-rank case.

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

The following corollary of Theorem 1 yields a condition under which even approximate recovery of $\mathbf{\Pi}^*$ within Hamming distance D, i.e., $\mathsf{d_H}(\mathbf{\Pi}^*;\widehat{\mathbf{\Pi}})\le\mathsf{D}$, cannot be guaranteed. Specifically, we state the following corollary of Theorem 1.

Corollary 4: Considering the oracle case with known $\mathbf{B}^*$, we have

$$\inf_{\widehat{\mathbf{\Pi}}}\sup_{\mathbf{\Pi}^*\in\mathscr{P}_n}\Pr_{\mathbf{X},\mathbf{W}}(\mathsf{d_H}(\widehat{\mathbf{\Pi}};\,\mathbf{\Pi}^*)\ge\mathsf{D})\ge\frac{1}{2}\tag{12}$$

$$\text{if}\quad\log\det\left(\mathbf{I}+\frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right)\le\frac{\log(n-\mathsf{D}+1)!-\log 4}{n},$$

where the infimum is over all estimators $\widehat{\mathbf{\Pi}}$.

Comparing the above result with Theorem 1, one can see that the essentially only difference is the replacement of the term $\log|\mathscr{H}|$ by $\log(n-\mathsf{D}+1)!$. An intuitive interpretation is as follows:

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

Consequently, since the cardinality of the covering is $(n-\mathsf{D}+1)!$, we encounter the term $\log(n-\mathsf{D}+1)!$ in place of $\log|\mathscr{H}|\le\log n!$; setting $\mathsf{D}=0$ or 1 gives back Theorem 1. Additionally, we can obtain a lower bound on the minimax risk with respect to $\mathsf{d_H}(\cdot;\cdot)$ effortlessly from the proof of Corollary 4, as

$$\inf_{\widehat{\mathbf{\Pi}}}\sup_{\mathbf{\Pi}^*}\mathbb{E}_{\mathbf{X},\mathbf{W}}\mathsf{d_H}(\widehat{\mathbf{\Pi}};\mathbf{\Pi}^*)\ge\max_{d\in\{0,1,\ldots,n\}}(d+1)\times$$
$$\left(1-\frac{(n/2)\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2\right)+\log 2}{\log(n-d+1)!}\right),\tag{13}$$

where $\mathbb{E}_{\mathbf{X},\mathbf{W}}(\cdot)$ denotes the expectation w.r.t. $\mathbf{X}$ and $\mathbf{W}$.

A unified proof for (12) and (13) can be found in Appendix D. To an extent, (12) strengthens the assertion of Theorem 1 in the sense that if $\log\det(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2)\ll\log n$, $\widehat{\mathbf{\Pi}}$ can be rather far from recovering $\mathbf{\Pi}^*$ in the sense that $\mathsf{d_H}(\widehat{\mathbf{\Pi}};\mathbf{\Pi}^*)=\Omega(n)$.

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.

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 both the oracle case with known $\mathbf{B}^*$ as well as the "realistic case" with unknown $\mathbf{B}^*$. For the conciseness of presentation, we hide explicit values for numerical constants in most cases and provide them in the appendix for interested readers. We believe that those values can be improved further since no specific effort was made to obtain optimal constants.

A. Oracle Case: Known $\mathbf{B}^*$

As previously mentioned, in this case the ML estimator in (5) is given by (7). The condition on the $\mathbf{snr}$ 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 5: In the oracle case with known $\mathbf{B}^*$, if

$$\log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2}\right) \geq \frac{8\log n}{\kappa\varrho(\mathbf{B}^*)} + \log\left(\kappa\varrho(\mathbf{B}^*) \vee \alpha_1 \log n\right) + \alpha_2, \tag{14}$$

then the ML estimator in (5) satisfies

$$\mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*) \leq \frac{2\alpha_0^{2\kappa\varrho(\mathbf{B}^*)}}{n^2} \overset{(\alpha_0<1)}{<} \frac{2}{n^2},$$

where $0 < \alpha_0 < 1$, $\kappa > 0$ are universal constants, $\alpha_1 = 2/\log(\alpha_0^{-1})$, and $\alpha_2 = \log(64\alpha_0^{-4}\log\alpha_0^{-1})$.

Proof outline: We show that each row of $\mathbf{Y}$ is closest in Euclidean distance to its matching row in $\mathbf{XB}^*$ with the stated probability, which implies the desired event $\{\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*\}$. The key ingredient is a careful probabilistic lower bound on the minimum distance between any pairs of rows in $\mathbf{XB}^*$ based on a small ball probability result in high-dimensional geometry due to Latala et al. [28]. □

To illustrate the tightness of conditions in (14), we would like to consider two special cases for $\mathbf{B}^*$, namely, the full-rank case and the rank-one case, and compare it with the condition for failure of recovery in Theorem 1. First, we consider the full-rank case with constant singular values, i.e., $\mathbf{B}^{*\top}\mathbf{B}^* = \gamma\mathbf{I}$, where $\gamma > 0$ is a positive constant; in particular, $\varrho(\mathbf{B}^*) = m$. Then a simple term re-arrangement of (14) suggests that having

$$\begin{aligned}\log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\varrho(\mathbf{B}^*)\sigma^2}\right) &= \log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m\sigma^2}\right) \\ &= \log\left(\frac{\gamma}{\sigma^2}\right) \gtrsim \frac{\log n}{\varrho(\mathbf{B}^*)}\end{aligned} \tag{15}$$

ensures success, while Theorem 1 suggests that

$$\log\left(1 + \frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m\sigma^2}\right) = \log\left(1 + \frac{\gamma}{\sigma^2}\right) \lesssim \frac{\log n}{\varrho(\mathbf{B}^*)} \tag{16}$$

implies failure. Conditions (15) and (16) thus match up to multiplicative factors.

Next, we consider the rank-one case. Without loss of generality, we set $\mathbf{B}^* = \mathbf{B}^*_{:,1} = \gamma\mathbf{e}_1$. Theorem 5 (cf. (14)) suggests that $\log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2}\right) = \log\left(\frac{\gamma}{\sigma^2}\right) \gtrsim \log n$ ensures success, while Theorem 1 suggests that $\log\left(1 + \frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2}\right) = \log(1 + \frac{\gamma}{\sigma^2}) \lesssim \log n$ leads to failure. Putting things together, we conclude tightness for this case. For a more clear view of (14), we omit the non-dominant terms, and state the implications in the following remark.

Remark: Consider the oracle case with known $\mathbf{B}^*$. The ML estimator of $\mathbf{\Pi}^*$ achieves permutation recovery with high probability (w.h.p.) in the following situations:

$\{\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\}$ holds w.h.p.

$$\text{if } \begin{cases}\kappa\varrho(\mathbf{B}^*) < \alpha_1 \log n \text{ and } \log(\mathbf{snr}) \geq \frac{(8+\kappa)\log n}{\kappa\varrho(\mathbf{B}^*)} + c_0, \\ \kappa\varrho(\mathbf{B}^*) \geq \alpha_1 \log n \text{ and } \mathbf{snr} \geq c_1.\end{cases}$$

In summary, Theorem 5 yields a more relaxed requirement on the $\mathbf{snr}$ needed to recover $\mathbf{\Pi}^*$ as the stable rank $\varrho(\mathbf{B}^*)$ exceeds a certain threshold. More specifically, the requirement becomes $\mathbf{snr} \geq c_1$ for some positive constant c_1 (in particular, the right hand side does not grow with n) once $\kappa\rho(\mathbf{B}^*) \gtrsim \log n$.

B. Realistic Case: Unknown $\mathbf{B}^*$

For this case with $\mathbf{B}^*$ unknown, we first present a basic result in Theorem 6 that will be improved upon later under additional assumptions.

Theorem 6: Let $\epsilon > 0$ be arbitrary, and suppose that $n > N_1(\epsilon)$, where $N_1(\epsilon) > 0$ is a positive constant depending only on ϵ. Suppose the following conditions hold: (i) $\mathbf{snr}\cdot n^{-\frac{2n}{n-p}} \geq 1$; and (ii)

$$\log(m \cdot \mathbf{snr}) \geq (c_0 + c_1\epsilon)\log n, \tag{17}$$

then the ML estimator in (5) equals $\mathbf{\Pi}^*$ with probability exceeding $1 - c_3 n^{-\epsilon}\left[(n^\epsilon - 1)^{-1} \vee 1\right]$, where $c_0, c_1, c_2, c_3 > 0$ are fixed positive constants.

Proof outline: The proof extends the proof strategy employed in Pananjady et al. [12] for $m = 1$ to arbitrary m, which amounts to showing that $\left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}} < \left\|P^{\perp}_{\mathbf{\Pi}\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}$ holds true for all $\mathbf{\Pi} \neq \mathbf{\Pi}^*$, where for $\mathbf{\Pi} \in \mathscr{P}_n$, $P^{\perp}_{\mathbf{\Pi}\mathbf{X}}$ denotes the projection on the orthogonal complement of the range of $\mathbf{\Pi X}$. Several critical steps in [12] do no longer apply for the case of multiple m considered herein, and prompt new technical challenges to be overcome. More details are available in Appendix F, including specific values of the constants c_0 and c_1. □

Theorem 6 states that exact recovery of $\mathbf{\Pi}^*$ can be achieved with high probability if $\log(m\cdot\mathbf{snr}) \gtrsim \log n$. For the rank-one case, we can see that this result is tight up to multiplicative constants in light of Theorem 1, which implies failure of recovery with high probability provided that $\log(1+m\mathbf{snr}) \lesssim \log n$. However, Theorem 6 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-one case. Unlike Theorem 5 concerning the oracle case, Theorem 6 thus fails to capture potential improvement brought by higher measurement diversity as quantified by the stable rank $\varrho(\mathbf{B}^*)$. To address this limitation, we present a refined result that comes at the expense of additional assumptions on $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}^*)$ and $\varrho(\mathbf{B}^*)$.

Theorem 7: Suppose that $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}^*) \leq h_{\max}$ with $h_{\max}$ satisfying the relation $h_{\max}r(\mathbf{B}^*) \leq n/8$. Let further $\epsilon > 0$ be arbitrary, and suppose that $n > N_2(\epsilon)$, where $N_2(\epsilon) > 0$ is a positive constant depending only on ϵ. In addition, suppose that the following conditions hold:

$$\begin{aligned}&(i)\ \mathbf{snr} > c_0, \quad (ii)\ \varrho(\mathbf{B}^*) \geq c_1(1+\epsilon)\log n, \\ &(iii)\ \log(\mathbf{snr}) \geq \frac{c_3(1+\epsilon)\log n}{\varrho(\mathbf{B}^*)} + c_4.\end{aligned} \tag{18}$$

Then the ML estimator (5) subject to the constraint $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}) \leq h_{\mathsf{max}}$ equals $\mathbf{\Pi}^*$ with probability at least $1 - 10n^{-\epsilon}\left[(n^{\epsilon}-1)^{-1} \vee 1\right]$, where $c_0, \ldots, c_4 > 0$ are some positive constants.

Remark 8: The $\mathbf{snr}$ requirement in (18) matches the minimax bound in Theorem 1 up to a logarithmic factor, when setting $\mathscr{H}$ as $\{\mathbf{\Pi} \in \mathscr{P}_n : \mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}) \leq h_{\mathsf{max}}\}$ and $h_{\mathsf{max}} \asymp \frac{n}{\log n}$.

In contrast to Theorem 6, the above theorem uncovers the benefits brought by larger stable rank $\varrho(\mathbf{B}^*)$. The outline of the proof strategy is given as follows with the technical details being placed in Section G.

Proof outline: The proof is analogous to that of Theorem 6. The key result is an improved upper bound on the probability $\Pr\left(\left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}} \leq c\|\mathbf{B}^*\|^2_{\mathrm{F}}\right)$, where $P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}$ denotes the projection on the orthogonal complement of the range of $\mathbf{\Pi}^*\mathbf{X}$. The above probability is bounded based on an ε-covering of the set of sparse unit vectors whose relevance is a consequence of the constraint $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}^*) \leq h_{\max}$. □

Let us comment on the additional constraint (i) $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}^*) \leq h_{\max}$ in Theorem 7. 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 $\mathbf{snr} \geq C$ for some constant $C > 0$, which then also matches the assertion in Theorem 5. At the same time, h_{max} is required to be of the order $h_{\mathsf{max}} \lesssim \frac{n}{\log n}$, which is only slightly sub-optimal compared to $h_{\max}$ being a linear fraction of n. We hypothesize that the constraint on $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}^*)$ can be eliminated, either with the help of more advanced proof techniques or by imposing more stringent constraints involving the ratio n/p.

Since the order for the required $\mathbf{snr}$ to achieve correct recovery remains the same as in Theorem 5, 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.

V. Computational Approach

In this section, we focus on computational aspects of the problem. Recall that for the oracle case, the ML estimator in (5) reduces to the linear assignment problem (7), and can be solved efficiently by the Hungarian algorithm [29] or the auction algorithm [30]. The emphasis in this section hence concerns the realistic case with $\mathbf{B}^*$ unknown. As proved in [12], computation of the ML estimator in this case is NP-hard except for the special case $m = p = 1$. In light of (6), we have the following

$$\min_{\mathbf{\Pi},\mathbf{B}} \|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}\|^2_{\mathrm{F}} = \min_{\mathbf{\Pi}} \left\|P^{\perp}_{\mathbf{X}}\mathbf{\Pi}^{\top}\mathbf{Y}\right\|^2_{\mathrm{F}}, \tag{19}$$

where $P^{\perp}_{\mathbf{X}}$ denotes the projection onto the orthogonal complement of the space spanned by the columns of $\mathbf{X}$. Problem (19) can be expressed as a *quadratic assignment problem* (QAP), cf. [31], [32]. This class is known to be challenging, and there are generally no algorithms that can provably deliver a global optimum. Approaches employed in practice are thus often based on various heuristics. We here adopt a similar strategy; specifically, we propose to tackle (19) via the projected gradient method as detailed in Algorithm 1. Given an initial iterate, the proposed approach is shown to either reduce the objective, or terminates (cf. Theorem 9 below).

Numerical experiments in Section VI show that Algorithm 1 performs well with a suitable initialization, and confirm the scaling law predicted by Theorem 7.

Algorithm 1 Projected Gradient Descent for the Recovery of $\mathbf{\Pi}^*$

- **Input**: $(\mathbf{X}, \mathbf{Y})$, initial permutation matrix $\mathbf{\Pi}^{(0)}$, step size $\alpha > 0$, convergence tolerance $\epsilon > 0$, and iteration limit T_{max}.
- **For** t **from** 0 **to** $(T_{\mathsf{max}} - 1)$**:** Update $\mathbf{\Pi}^{(t+1)}$ as

$$\mathbf{D}^{(t+1)} = \mathbf{\Pi}^{(t)} + 2\alpha\mathbf{Y}\mathbf{Y}^{\top}\mathbf{\Pi}^{(t)}P_{\mathbf{X}}, \tag{20}$$

$$\mathbf{\Pi}^{(t+1)} = \mathrm{argmax}_{\mathbf{\Pi}} \left\langle \mathbf{\Pi}, \mathbf{D}^{(t+1)} \right\rangle, \tag{21}$$

  where $P_{\mathbf{X}} \triangleq \mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}$ is the projection onto the column space of $\mathbf{X}$.
- **Termination**: Stop the algorithm once $\left\|\mathbf{\Pi}^{(t+1)} - \mathbf{\Pi}^{(t)}\right\|_{\mathrm{F}} \leq \epsilon$; or $t = T_{\mathsf{max}}$.

A. Convergence Analysis

The following Theorem states that the sequence $\left\{\mathbf{\Pi}^{(t)}\right\}$ generated by Algorithm 1 yields non-increasing objective values. Its proof is deferred to Section H.

Theorem 9: Consider the objective in (19). For any $\alpha > 0$, the iterates generated by Algorithm 1 satisfy

$$\left\|P^{\perp}_{\mathbf{X}}\mathbf{\Pi}^{(t+1)}\mathbf{Y}\right\|_{\mathrm{F}} \leq \left\|P^{\perp}_{\mathbf{X}}\mathbf{\Pi}^{(t)}\mathbf{Y}\right\|_{\mathrm{F}},$$

where the inequality is achieved when $\mathbf{\Pi}^{(t+1)} = \mathbf{\Pi}^{(t)}$.

We emphasize that the above result is non-trivial since the underlying constraint, the set of permutation matrices of size n, is not convex. Although the proposed algorithm converges, it may not necessarily converge to the global optimum. The choice of the initialization tends to be critical in this regard. According to our numerical experiments, we have found that when the Hamming distance between $\mathbf{\Pi}^*$ and $\mathbf{I}$ is not too large, we can simply initialize $\mathbf{\Pi}^{(0)} = \mathbf{I}$.

B. Gradient Computation

Apart from the initialization, it turns out that the specific way of computing the gradient in (20) also plays an important role. Instead of direct evaluation, which yields $\nabla\left\|P^{\perp}_{\mathbf{X}}\mathbf{\Pi}^{\top}\mathbf{Y}\right\|^2_{\mathrm{F}} = 2\mathbf{Y}\mathbf{Y}^{\top}\mathbf{\Pi}P^{\perp}_{\mathbf{X}}$, we use that for any permutation matrix $\mathbf{\Pi}$, it holds that $\left\|P^{\perp}_{\mathbf{X}}\mathbf{\Pi}^{\top}\mathbf{Y}\right\|^2_{\mathrm{F}} = \|\mathbf{Y}\|^2_{\mathrm{F}} - \left\|P_{\mathbf{X}}\mathbf{\Pi}^{\top}\mathbf{Y}\right\|^2_{\mathrm{F}}$ and hence

$$\begin{aligned}\nabla\left\|P^{\perp}_{\mathbf{X}}\mathbf{\Pi}^{\top}\mathbf{Y}\right\|^2_{\mathrm{F}} &= \nabla\left[\|\mathbf{Y}\|^2_{\mathrm{F}} - \left\|P_{\mathbf{X}}\mathbf{\Pi}^{\top}\mathbf{Y}\right\|^2_{\mathrm{F}}\right] \\ &= -2\mathbf{Y}\mathbf{Y}^{\top}\mathbf{\Pi}P_{\mathbf{X}},\end{aligned}$$

which proves to be a better search direction according to our numerical experiments. One intuitive explanation is as follows. In the regime of large $\mathbf{snr}$, namely, $\mathbf{snr} \gg 1$, we have the approximation

$$\mathbf{Y}\mathbf{Y}^{\top}\mathbf{\Pi}P_{\mathbf{X}} \approx \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\mathbf{B}^{*\top}\mathbf{X}^{\top}\mathbf{\Pi}^{*\top}\mathbf{\Pi}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top},$$

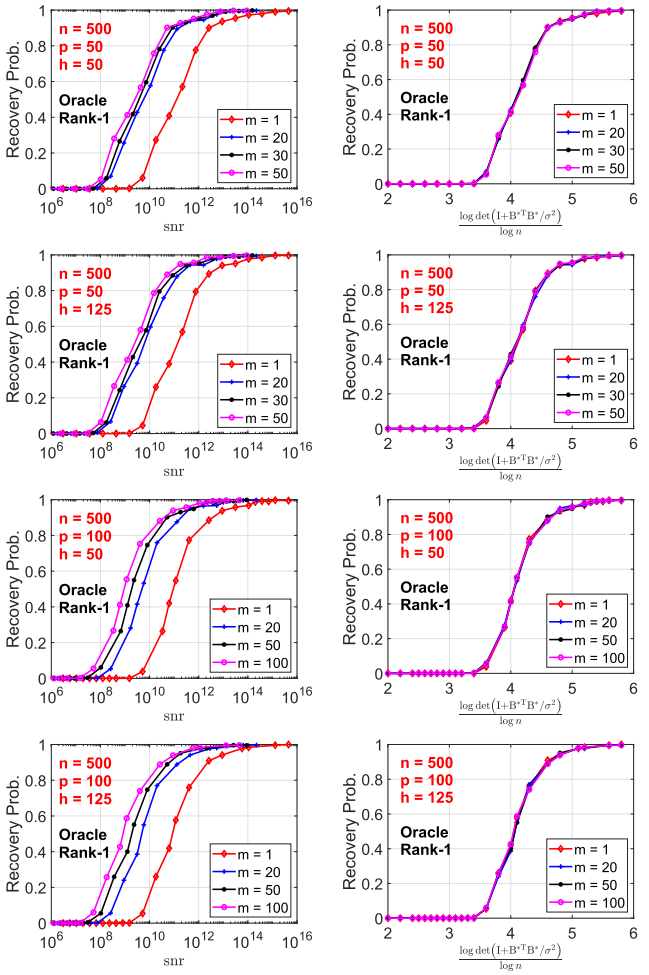

Fig. 3. Oracle case (Rank-1): Correct recovery probability $\mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*)$ versus $\mathbf{snr}$ (left panels) or $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^{*}/\sigma^2\right)}{\log n} = \frac{\log(1+m\cdot\mathbf{snr})}{\log n}$ (right panels).

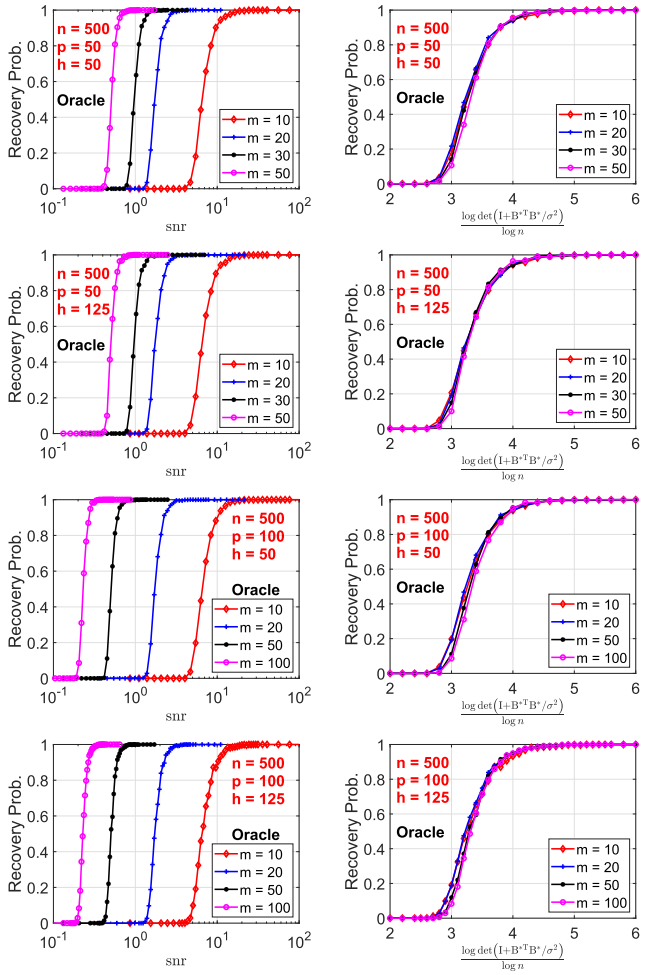

Fig. 4. Oracle case (Full-rank): Correct recovery probability $\mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*)$ versus $\mathbf{snr}$ (left panels) or $\frac{\log(1+m\cdot\mathbf{snr})}{\log n}$ (right panels).

where $\approx$ reflects the omission of the noise $\mathbf{W}$. Since $n \gg p$ and $X_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, the matrix $\mathbf{X}^\top\mathbf{X}/n$ is close to the identity matrix, and furthermore $\mathbf{X}^\top\mathbf{\Pi}^*\mathbf{\Pi}\mathbf{X} \approx \mathbb{E}\mathbf{X}^\top\mathbf{\Pi}^*\mathbf{\Pi}\mathbf{X} \propto \mathbf{I}$ if $\mathsf{d}_\mathsf{H}(\mathbf{\Pi}^*,\mathbf{\Pi})/n \ll 1$. In summary, this yields the approximation

$$\begin{aligned}\mathbf{Y}\mathbf{Y}^\top\mathbf{\Pi}P_{\mathbf{X}} &\approx \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\mathbf{B}^{*\top}\mathbb{E}\left(\mathbf{X}^\top\mathbf{\Pi}^{*\top}\mathbf{\Pi}\mathbf{X}\right)\mathbf{X}^\top \\ &\propto \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\mathbf{B}^{*\top}\mathbf{X}^\top.\end{aligned}$$

This implies that given sufficient proximity of $\mathbf{\Pi}$ and $\mathbf{\Pi}^*$ with respect to the Hamming distance, $\mathbf{Y}\mathbf{Y}^\top\mathbf{\Pi}P_{\mathbf{X}}$ will be roughly aligned with the direction $\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\mathbf{B}^{*\top}\mathbf{X}^\top$. The latter generates $\mathbf{\Pi}^*$ after the projection step in (21) since $\mathrm{argmax}_{\mathbf{\Pi}}\left\langle\mathbf{\Pi}, \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\mathbf{B}^{*\top}\mathbf{X}^\top\right\rangle = \mathrm{argmax}_{\mathbf{\Pi}}\left\langle\mathbf{\Pi}\mathbf{X}\mathbf{B}^*, \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\right\rangle = \mathbf{\Pi}^*$.

VI. Numerical Results

In this section, we present simulation results and investigate the relation between the correct recovery rate $\mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*)$ and the signal energy. The experiments are divided into two parts: 1) the oracle case (known $\mathbf{B}^*$) and 2) the realistic case with $\mathbf{B}^*$ being unknown.

A. Oracle Case

In this subsection, we study the relation between the correct probability $\mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*)$ and the $\mathbf{snr}$ in the oracle case with known $\mathbf{B}^*$. As mentioned previously, the ML estimator is obtained as the solution of the linear assignment problem (7). The latter is here solved by the auction algorithm [30]. The simulation results confirm our theoretical results in Theorem 1 and Proposition 3. In virtue of Theorem 1, we 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. We also use $\mathbf{snr}$ on the horizontal axis to illustrate the energy savings brought by multiple measurement vectors.

1) Rank-One Case: We use $\mathbf{B}^*$ such that all $\{\mathbf{B}^*_{:,i}\}_{i=1}^m$ are identical. The simulation results are displayed in Figure 3. The left panels use $\mathbf{snr}$ for the horizontal axis while the right panels show the corresponding values of $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^{*}/\sigma^2\right)}{\log n}$, which can be rewritten as $\frac{\log(1+m\mathbf{snr})}{\log n}$ in this case. Observing that the curves coincide in the right panels, we conclude that increasing values of m are irrelevant in this case given a fixed ratio of the total signal energy to the noise variance $\|\mathbf{B}^*\|_{\mathrm{F}}^2/\sigma^2$.

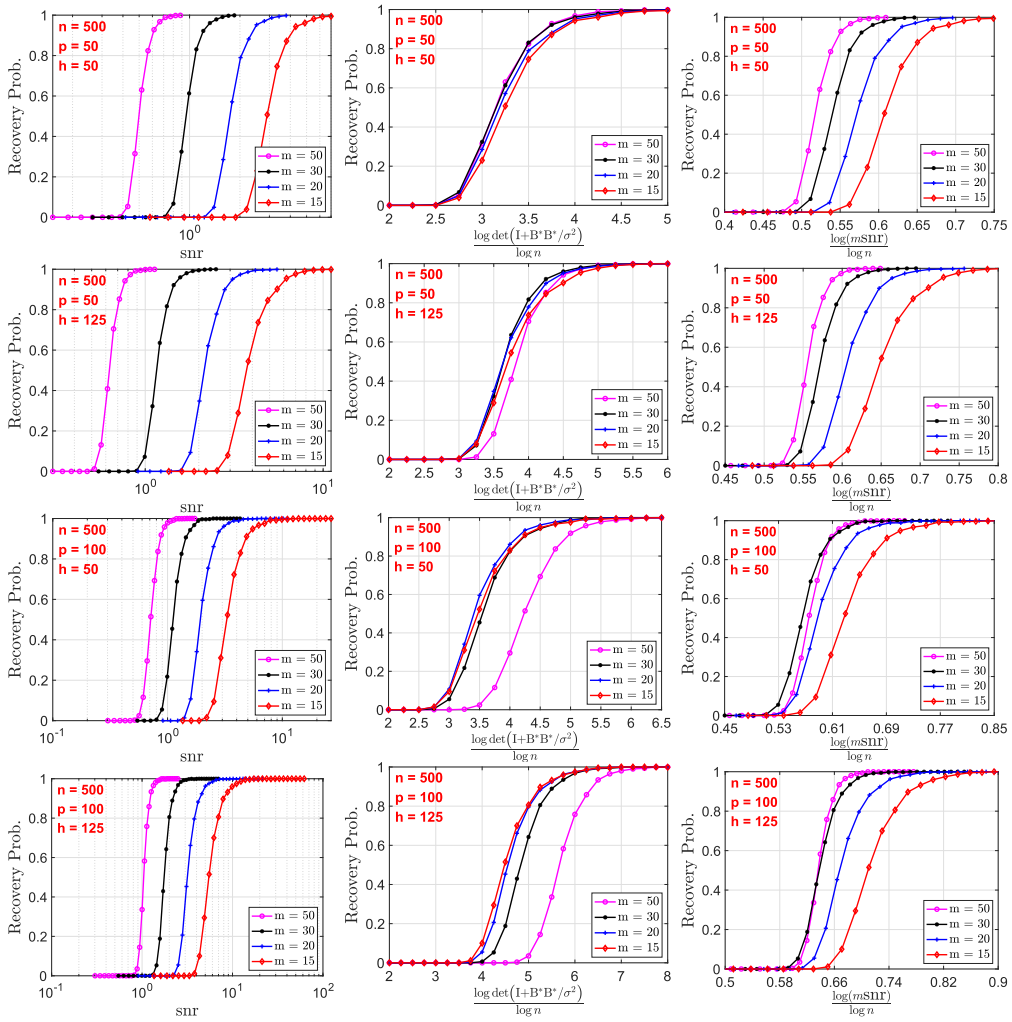

Fig. 5. Realistic case: Correct recovery probability $\Pr_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*)$ versus $\mathbf{snr}$ (left panels) or $\frac{\log\det(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^{*}/\sigma^{2})}{\log n}$ (middle panels), or $\frac{\log(m\mathbf{snr})}{\log n}$ (right panels).

2) Full-Rank Case: We consider the case in which the columns of $\mathbf{B}^*$ are orthogonal to each other, i.e., $\mathbf{B}^*_{:,i} \perp \mathbf{B}^*_{:,j}$, $1 \leq i \neq j \leq m$. For simplicity, we set $\mathbf{B}^*_{:,i} \parallel \mathbf{e}_i$, where $\{\mathbf{e}_i\}$ denotes the canonical basis. The simulation results for this setting are shown in Figure 4. As for the rank-one case, we observe that the curves displaying the correct recovery rate $\Pr_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*)$ for different values of m almost coincide when using the quantity $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^{*}/\sigma^{2}\right)}{\log n}$ for the horizontal axis. The latter is thus confirmed to be the central determining factor in predicting whether $\mathbf{\Pi}^*$ can be successfully recovered or not. However, different from the rank-one case, we witness a significant decrease regarding the required $\mathbf{snr}$ needed for high recovery rates. For example, $\mathbf{snr} \approx 10^{14}$ is required in the rank-one case, while in the full-rank case, the required value of $\mathbf{snr}$ is less than 10. As predicted by Theorem 1 and Theorem 5, this reduction is a consequence of an increased stable rank $\varrho(\mathbf{B}^*)$.

B. Realistic Case

This subsection is concerned with the realistic case in which $\mathbf{B}^*$ is not known. We fix $n = 500$ and consider $p = \{50, 100\}$ as well as $h = \{50, 125\}$, where $h = \mathsf{d}_{\mathsf{H}}(\mathbf{I}; \mathbf{\Pi}^*)$. The estimator of $\mathbf{\Pi}^*$ is obtained by applying Algorithm 1. The results are shown in Figure 5. Given the excessive requirements regarding $\mathbf{snr}$ in the rank-one case even in the oracle case, we here focus on the full-rank case. Apart from using $\mathbf{snr}$ and $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^{*}/\sigma^{2}\right)}{\log n}$ for the horizontal axis, we additionally consider $\frac{\log(m\mathbf{snr})}{\log n}$ in virtue of Theorem 6.

Inspection of the left panels of Figure 5 indicates a similar phenomenon as observed in the oracle case, namely, a significant reduction of the required $\mathbf{snr}$ with large stable rank $\varrho(\mathbf{B}^*) = m$. When $m = 15$, the $\mathbf{snr}$ required to achieve recovery falls in the range $[1,\ 10]$. When m is increased to 50, the required $\mathbf{snr}$ drops below 1 in alignment with the implications of Theorem 1 and Theorem 7.

However, different from the oracle case where the ratio $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^{*}/\sigma^{2}\right)}{\log n}$ required for permutation recovery is almost independent of the triple (n, p, h), we now observe variation across different settings. When $n = 500, p = 100, h = 50$, permutation recovery is achieved if the quantity $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^{*}/\sigma^{2}\right)}{\log n}$ exceeds 5, which is almost identical to

the oracle case shown in Figure 4. However, the associated recovery threshold is contained in $[5,\ 6]$ when $(n,p,h) = (500,100,50)$, and further increases to 7 when $(n,p,h) = (500,100,125)$. Generally speaking, the larger the ratio n/p and the smaller the Hamming distance h, the lower the required value of the ratio $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^{*}/\sigma^2\right)}{\log n}$.

VII. Conclusion

In this paper, we have studied the unlabeled sensing problem given multiple measurement vectors. First, we establish the statistical limits in terms of conditions on the $\mathbf{snr}$ implying failure of recovery with high probability, namely, $\varrho(\mathbf{B}^*)\log(\mathbf{snr}) \lesssim \log n$. 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 need $\log(m\cdot\mathbf{snr}) \gtrsim \log n$ for correct recovery, which matches the lower bound for the oracle case with $\varrho(\mathbf{B}^*)=1$. By imposing the additional assumption $\mathsf{d_H}(\mathbf{I};\mathbf{\Pi}^*) \le h_{\mathsf{max}} \lesssim n/\log n$, it can be proved that $\varrho(\mathbf{B}^*)\log(\mathbf{snr}) \gtrsim \log n$ is also sufficient for correct recovery, which matches the corresponding minimax lower bound up to a logarithmic factor. On the computational side, the underlying problem is cast as a quadratic assignment problem, and a practical optimization scheme based on the projected gradient method is proposed to tackle the computational challenge associated with the ML estimator. The results of our simulations based on the proposed numerical scheme largely corroborate our theoretical findings.

Appendix A
Notations

We begin the appendix with a restatement of the notations we use. 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 $(i,j)^{\mathsf{th}}$ 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 $\mathbf{A}$ [25] (Section 2.4, P. 76) 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 $\|\cdot\|_{\mathsf{OP}}$ is used for the operator norm, whose definitions can be found in [25] (Section 2.3, P. 71). The ratio $\varrho(\cdot) = \|\cdot\|_{\mathrm{F}}^2/\|\cdot\|_{\mathsf{OP}}^2$ represents the stable rank while $r(\cdot)$ represents the usual rank of a matrix.

We write $\boldsymbol{\pi}(\cdot)$ for a permutation of $\{1,2,\cdots,n\}$ that moves index i to $\boldsymbol{\pi}(i)$, $1\le i\le n$. The corresponding permutation matrix is denoted by $\mathbf{\Pi}$. We use $\mathsf{d_H}(\cdot;\cdot)$ to denote the Hamming distance between two permutation matrices, *i.e.*, $\mathsf{d_H}(\mathbf{\Pi}_1;\mathbf{\Pi}_2) = \sum_{i=1}^n \mathbb{1}\left(\boldsymbol{\pi}_1(i)\neq\boldsymbol{\pi}_2(i)\right)$. Viewing $\mathbf{\Pi}$ as a RV distributed among set $\mathscr{H}$, we denote its entropy as $\mathsf{H}(\mathbf{\Pi})$. The differential entropy is denoted as $\mathsf{h}(\cdot)$ and the mutual information is denoted as $\mathrm{I}(\cdot;\cdot)$.

For an event $\mathscr{E}$, we denote its complement by $\overline{\mathscr{E}}$, and use $\Psi(\mathscr{E})$ to denote $\mathbb{E}\mathbb{1}(\mathscr{E})$. 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 Theorem 1

Proof: The proof of Theorem 1 heavily relies on Lemma 10. We put a uniform prior on $\mathbf{\Pi}^*$ over the support $\mathscr{H}$, which maximizes the entropy $\mathsf{H}(\mathbf{\Pi}^*) = \log|\mathscr{H}|$, and exploit the inequality

$$\sup_{\mathbf{\Pi}} \mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}) \ge \mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}\left(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*\right), \tag{22}$$

where the probability measure $\mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\cdot)$ is w.r.t. $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{\Pi}^*$. Since Lemma 10 holds for arbitrary estimator $\widehat{\mathbf{\Pi}}$, we can safely add $\inf_{\widehat{\mathbf{\Pi}}}$ to the left-hand side in (22) and complete the proof. □

Lemma 10: Viewing $\mathbf{\Pi}^*$ as a RV distributed among the set $\mathscr{H}$, we have

$$\begin{aligned}&\mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*)\\ \ge\ &\frac{\mathsf{H}(\mathbf{\Pi}^*)-1-(n/2)\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2\right)}{\log\left(|\mathscr{H}|\right)},\end{aligned}$$

for an arbitrary estimator $\widehat{\mathbf{\Pi}}$, where $\mathsf{H}(\cdot)$ is the entropy of $\mathbf{\Pi}^*$, and the probability measure $\mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\cdot)$ is w.r.t. $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{\Pi}^*$.

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 (4), i.e., $\mathbf{Y} = \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^* + \mathbf{W}$, as the following transmission process

$$\mathbf{\Pi}^* \overset{①}{\to} \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^* \overset{②}{\to} \mathbf{\Pi}^*\mathbf{X}\mathbf{B}^* + \mathbf{W}, \tag{23}$$

where in ① the signal $\mathbf{\Pi}^*$ is encoded to the codeword $\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*_{:,i}$, and in ② 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 [26] (cf. Section 7.9, P. 206), we have

$$\begin{aligned}\mathsf{H}(\mathbf{\Pi}^*) &\overset{③}{=} \mathsf{H}(\mathbf{\Pi}^*\mid\mathbf{X}) \overset{④}{=} \mathsf{H}(\mathbf{\Pi}^*\mid\widehat{\mathbf{\Pi}},\mathbf{X}) + \mathrm{I}(\mathbf{\Pi}^*;\ \widehat{\mathbf{\Pi}}\mid\mathbf{X})\\ &\overset{⑤}{\le} \mathsf{H}(\mathbf{\Pi}^*\mid\widehat{\mathbf{\Pi}}) + \mathrm{I}(\mathbf{\Pi}^*;\ \widehat{\mathbf{\Pi}}\mid\mathbf{X})\\ &\overset{⑥}{\le} 1+\log\left(|\mathscr{H}|\right)\mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*) + \mathrm{I}(\mathbf{\Pi}^*;\ \widehat{\mathbf{\Pi}}\mid\mathbf{X})\\ &\overset{⑦}{\le} 1+\log\left(|\mathscr{H}|\right)\mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*) + \mathrm{I}(\mathbf{\Pi}^*;\ \mathbf{Y}\mid\mathbf{X})\\ &\overset{⑧}{\le} 1+\log\left(|\mathscr{H}|\right)\mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\widehat{\mathbf{\Pi}}\neq\mathbf{\Pi}^*)\\ &\quad+\frac{n}{2}\log\det\left(\mathbf{I}+\frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right),\end{aligned}$$

where λ_i denote the i^{th} singular value of $\mathbf{B}^*$, in ③ we use the fact that $\mathbf{X}$ and $\mathbf{\Pi}^*$ are independent, in ④ we use the definition of the conditional mutual information $\mathrm{I}(\mathbf{\Pi}^*;\widehat{\mathbf{\Pi}}\mid\mathbf{X})$, in ⑤ we use $\mathsf{H}(\mathbf{\Pi}^*\mid\widehat{\mathbf{\Pi}},\mathbf{X}) \le \mathsf{H}(\mathbf{\Pi}^*\mid\widehat{\mathbf{\Pi}})$, in ⑥ we use Fano's inequality in Theorem 2.10.1 in [26], in ⑦ we use the data-processing inequality, noting that $\mathbf{\Pi}^*\to\mathbf{Y}\to\widehat{\mathbf{\Pi}}$ forms

a Markov chain [26], and in ⑧ we use Lemma 11 to upper bound the conditional mutual information $\mathrm{I}(\mathbf{\Pi}^*;\ \mathbf{Y} \mid \mathbf{X})$.

We thus obtain the following lower bound on $\mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\widehat{\mathbf{\Pi}} \neq \mathbf{\Pi}^*)$ reading as

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

which is bounded below by $1/2$ provided that

$$\mathsf{H}(\mathbf{\Pi}^*) > 1 + \frac{n}{2}\log\det\left(\mathbf{I} + \frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right) + \frac{\log\left(|\mathscr{H}|\right)}{2},$$

and complete the proof. □

Lemma 11: For the channel described in (23), we have

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

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

Proof: Let $\mathrm{vec}\,(\mathbf{Y})$ ($\mathrm{vec}\,(\mathbf{W})$) be the vector 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 $\mathrm{I}\left(\mathbf{\Pi}^*;\ \mathrm{vec}\,(\mathbf{Y})\,|\mathbf{X}\right)$ as a shortcut for $\mathrm{I}\left(\mathbf{\Pi}^*;\ \mathbf{Y}_{:,1}, \mathbf{Y}_{:,2}, \ldots, \mathbf{Y}_{:,m} |\ \mathbf{X}\right)$. We then calculate the conditional mutual information $\mathrm{I}\left(\mathbf{\Pi}^*;\ \mathrm{vec}\,(\mathbf{Y})\,|\mathbf{X}\right)$ as

$$\begin{aligned}
&\mathrm{I}\left(\mathbf{\Pi}^*; \mathrm{vec}\,(\mathbf{Y})\,|\mathbf{X}\right)\\
&\overset{①}{=} \mathsf{h}(\mathrm{vec}(\mathbf{Y})|\mathbf{X}) - \mathsf{h}(\mathrm{vec}(\mathbf{Y})|\mathbf{X}, \mathbf{\Pi}^*)\\
&\overset{②}{=} \mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}\mathsf{h}\left(\mathrm{vec}(\mathbf{Y})|\mathbf{X}=\mathbf{x}\right) - \mathsf{h}(\mathrm{vec}(\mathbf{W}))\\
&\overset{③}{=} \mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}\mathsf{h}\left(\mathrm{vec}(\mathbf{Y})|\mathbf{X}=\mathbf{x}\right) - \frac{mn}{2}\log\sigma^2\\
&\overset{④}{\leq} \mathbb{E}_{\mathbf{X}}\frac{1}{2}\log\det\left(\mathbb{E}_{\mathbf{\Pi}^*,\mathbf{W}|\mathbf{X}=\mathbf{x}}\mathrm{vec}\,(\mathbf{Y})\,\mathrm{vec}\,(\mathbf{Y})^\top\right) - \frac{mn}{2}\log\sigma^2,\\
&\leq \frac{1}{2}\log\det\mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}\mathrm{vec}\,(\mathbf{Y})\,\mathrm{vec}\,(\mathbf{Y})^\top - \frac{mn}{2}\log\sigma^2, \qquad (24)
\end{aligned}$$

where in ① we use the definition of the conditional mutual information $\mathrm{I}(\mathbf{\Pi}^*;\ \mathrm{vec}\,(\mathbf{Y})\mid\mathbf{X})$, in ② we have used that

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

in ③ 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 ④ we use a result in [26] (Theorem 8.6.5, P. 254) which yields

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

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 $\mathbb{E}_{\mathbf{\Pi}^*,\mathbf{X},\mathbf{W}}\mathrm{vec}\,(\mathbf{Y})\,\mathrm{vec}\,(\mathbf{Y})^\top$. For simplicity of notation, the latter matrix will henceforth be denoted by $\mathbf{\Sigma}$. First note that $\mathrm{vec}\,(\mathbf{Y})$ 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 $(j_1,j_2)^{\mathsf{th}}$ 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 written as

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

where $\boldsymbol{\pi}^*$ is the permutation corresponding to the permutation matrix $\mathbf{\Pi}^*$ as defined in Appendix A.

We then split the calculation into three sub-cases:

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

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

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

where $\otimes$ denotes the Kronecker product [25] (Section 1.3.6, P. 27). According to [25] (Section 12.3.1, P. 709), we have

$$\begin{aligned}
\det\left(\mathbf{\Sigma}\right) &= \left(\det\left(\mathbf{\Sigma}_1\right)\right)^n\ \left(\det\left(\mathbf{I}_{n\times n}\right)\right)^m\\
&\overset{⑤}{=} \sigma^{2nm}\left(\det\left(\mathbf{I} + \frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right)\right)^n, \qquad (25)
\end{aligned}$$

where in ⑤ we have calculated $\det(\mathbf{\Sigma}_1)$ as

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

Combining (24) and (25), yields the upper bound

$$\begin{aligned}
\mathrm{I}\left(\mathbf{\Pi}^*;\ \mathrm{vec}\,(\mathbf{Y})\,|\mathbf{X}\right) &\leq \frac{n}{2}\log\det\left(\mathbf{I} + \frac{\mathbf{B}^{*\top}\mathbf{B}^*}{\sigma^2}\right)\\
&\overset{⑥}{=} \sum_i \log\left(1 + \frac{\lambda_i^2}{\sigma^2}\right),
\end{aligned}$$

where ⑥ can be verified via the singular value decomposition SVD $(\mathbf{B}^*) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ as introduced in Appendix A) and by using basic properties of the matrix determinant [33] (Section 0.3, P. 8). □

APPENDIX C
PROOF OF PROPOSITION 3

A. Roadmap

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 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. The proof is then completed with the following three stages.

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

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

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

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

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

Stage III: Provided Condition (11) holds, we are allowed to set $\rho_0 = 2\sqrt{2\sigma^2 \log n}$ without violating the requirement of Lemma 13. We thereby conclude the proof by setting $\rho_0 = 2\sqrt{2\sigma^2 \log n}$ and invoking Lemma 12 and Lemma 13.

B. Proof Details

Proof: 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 $\Pr_{\mathbf{X},\mathbf{W}}\left(\widehat{\mathbf{\Pi}} \ne \mathbf{I}\right)$ as

$$\begin{aligned} &\Pr_{\mathbf{X},\mathbf{W}}\left(\widehat{\mathbf{\Pi}} \ne \mathbf{I}\right) \\ \ge\ &\Pr_{\mathbf{X},\mathbf{W}}\left(\exists\ (i,j), \text{s.t. } \widetilde{W}_{i,j} \ge \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2\right) \\ \overset{①}{\ge}\ &\Pr_{\mathbf{X},\mathbf{W}}\left(\widetilde{W}_{1,j_0} \ge \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{1,:} - \mathbf{X}_{j_0,:}\right)\right\|_2\right) \\ \ge\ &\Pr_{\mathbf{X},\mathbf{W}}\left(\widetilde{W}_{1,j_0} \ge \rho_0 \,\middle|\, \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{1,:} - \mathbf{X}_{j_0,:}\right)\right\|_2 \le \rho_0\right) \\ &\times \Pr_{\mathbf{X}}\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{1,:} - \mathbf{X}_{j_0,:}\right)\right\|_2 \le \rho_0\right) \\ \overset{②}{=}\ &\Pr_{\mathbf{W}}\left(\widetilde{W}_{1,j_0} \ge \rho_0\right) \Pr_{\mathbf{X}}\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{1,:} - \mathbf{X}_{j_0,:}\right)\right\|_2 \le \rho_0\right), \end{aligned}$$

where in ① we pick j_0 as $\operatorname{argmax}_j \widetilde{W}_{1,j}$ and in ② we use the independence between $\widetilde{W}_{i,j}$ and $\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2$. □

C. Supporting Lemmas

Lemma 12: When n is large ($n \ge 10$), we have

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

Proof: This result is quite standard and can be easily proved by combining Section 2.5 (P. 31) and Theorem 5.6 (P. 126) in [34]. We omit the details for the sake of brevity. □

Lemma 13: 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

$$\Pr_{\mathbf{X}}\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2 \le \rho_0\right) \ge \frac{5}{9},$$

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

Proof: We begin the analysis by defining the following notations,

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

respectively, where $\mathbf{x} \in \mathbb{R}^p$ is a Gaussian RV satisfying $\mathscr{N}(\mathbf{0}, \mathbf{I}_{p\times p})$, and ρ_0 is some positive parameter awaiting to be set. First we prove the inequality $\Pr_{\mathbf{X}}(A_{\rho_0}^{(i,j)}) \ge \zeta^2$. Provided that $\mathbf{X}_{i,:} \in \mathbb{B}, \mathbf{X}_{j,:} \in \mathbb{B}$, we have $\mathscr{A}_{\rho_0}^{(i,j)}$ be true because

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

Then we conclude

$$\Pr_{\mathbf{X}}\left(A_{\rho_0}^{(i,j)}\right) \ge \mathbb{E}_{\mathbf{X}}\mathbb{1}\left(\mathbf{X}_{i,:} \in \mathbb{B}, \mathbf{X}_{j,:} \in \mathbb{B}\right) \overset{①}{=} \zeta^2, \tag{26}$$

where ① is because of the independence between $\mathbf{X}_{i,:}$ and $\mathbf{X}_{j,:}$. It thus remains to lower bound ζ, which is accomplished by setting ρ_0 to be $\rho_0 \ge 2(1+t)\|\mathbf{B}^*\|_{\mathrm{F}}$ and invoking Theorem 2.1 in [35]

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

for $t \ge 0$. Setting $t = 1.4356/\sqrt{c_1 \varrho(\mathbf{B}^*)}$, we have $\zeta \ge \sqrt{5}/3$, which implies
$\Pr\left(\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:} - \mathbf{X}_{j,:}\right)\right\|_2 \le \rho_0\right) \ge \zeta^2 \ge 5/9$ in view of (26) and completes the proof. □

Appendix D
Proof of Corollary 4

Proof: First we define $\mathscr{E} \triangleq \mathbb{1}\left\{\mathsf{d}_{\mathsf{H}}(\widehat{\mathbf{\Pi}}; \mathbf{\Pi}^*) \ge \mathsf{D}\right\}$, 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. Using the same logic as in Section 1, we conclude

$$\inf_{\widehat{\mathbf{\Pi}}} \sup_{\mathbf{\Pi}^*} \Pr_{\mathbf{X},\mathbf{W}}(\mathsf{d}_{\mathsf{H}}(\widehat{\mathbf{\Pi}};\ \mathbf{\Pi}^*) \ge \mathsf{D}) \ge \Pr_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\mathscr{E} = 1).$$

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

$$\begin{aligned}
&\mathsf{H}(\mathscr{E}, \mathbf{\Pi}^* | \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \\
&= \mathsf{H}(\mathbf{\Pi}^* \mid \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) + \mathsf{H}(\mathscr{E} \mid \mathbf{\Pi}^*, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \\
&\overset{①}{=} \mathsf{H}(\mathbf{\Pi}^* \mid \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \overset{②}{=} \mathsf{H}(\mathbf{\Pi}^* \mid \mathbf{Y}, \mathbf{X}),
\end{aligned} \tag{27}$$

where in ① we have used that $\mathsf{H}\left(\mathscr{E} \mid \mathbf{\Pi}^*, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}\right) = 0$ since $\mathscr{E}$ is deterministic once $\mathbf{\Pi}^*, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}$ are given, and in ② we use the fact $\mathrm{I}(\widehat{\mathbf{\Pi}}; \mathbf{\Pi}^* \mid \mathbf{Y}, \mathbf{X}) = 0$ since $\widehat{\mathbf{\Pi}}$ and $\mathbf{\Pi}^*$ are independent given $\mathbf{X}$ and $\mathbf{Y}$. At the same time, we have

$$\begin{aligned}
&\mathsf{H}\left(\mathscr{E}, \mathbf{\Pi}^* | \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}\right) \\
&= \mathsf{H}\left(\mathscr{E} | \widehat{\mathbf{\Pi}}\right) + \mathsf{H}\left(\mathbf{\Pi}^* \mid \mathscr{E}, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}\right) \\
&\overset{③}{\leq} \log 2 + \mathsf{H}\left(\mathbf{\Pi}^* \mid \mathscr{E}, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}\right) \\
&\leq \log 2 + \mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\mathscr{E}=1)\mathsf{H}(\mathbf{\Pi}^* | \mathscr{E}=1, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \\
&+ \mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\mathscr{E}=0)\mathsf{H}(\mathbf{\Pi}^* | \mathscr{E}=0, \widehat{\mathbf{\Pi}}, \mathbf{Y}, \mathbf{X}) \\
&\leq \log 2 + \mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\mathscr{E}=1)\,\mathsf{H}(\mathbf{\Pi}^* | \mathscr{E}=1,\ \widehat{\mathbf{\Pi}}) \\
&+ \mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\mathscr{E}=0)\,\mathsf{H}(\mathbf{\Pi}^* | \mathscr{E}=0,\ \widehat{\mathbf{\Pi}}) \\
&\leq \log 2 + (1 - \mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\mathscr{E}=0))\,\mathsf{H}(\mathbf{\Pi}^*) \\
&+ \mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\mathscr{E}=0) \log\left(\frac{n!}{(n-\mathsf{D}+1)!}\right) \\
&\overset{④}{=} \log 2 + \mathsf{H}(\mathbf{\Pi}^*) - \mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\mathscr{E}=0)\log(n-\mathsf{D}+1)!,
\end{aligned} \tag{28}$$

where in ③ we use the fact that $\mathscr{E}$ is binary and hence $\mathsf{H}(\mathscr{E}|\widehat{\mathbf{\Pi}}) \leq \log 2$, and in ④ we use the fact that $\mathsf{H}(\mathbf{\Pi}^*) = \log(n!)$. Combing (27) and (28) yields

$$\begin{aligned}
\mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}(\mathscr{E}=0) &\leq \frac{\mathrm{I}(\mathbf{\Pi}^*;\ \mathbf{Y},\ \mathbf{X}) + \log 2}{\log(n-\mathsf{D}+1)!} \\
&\overset{⑤}{=} \frac{\mathrm{I}(\mathbf{\Pi}^*; \mathbf{X}) + \mathrm{I}(\mathbf{\Pi}^*; \mathbf{Y}|\mathbf{X}) + \log 2}{\log(n-\mathsf{D}+1)!} \\
&\overset{⑥}{=} \frac{\mathrm{I}(\mathbf{\Pi}^*; \mathbf{Y}|\mathbf{X}) + \log 2}{\log(n-\mathsf{D}+1)!} \\
&\overset{⑦}{\leq} \frac{(n/2)\sum_i \log\left(1+\lambda_i^2/\sigma^2\right) + \log 2}{\log(n-\mathsf{D}+1)!},
\end{aligned} \tag{29}$$

which completes the proof of Corollary 4, where ⑤ is because of the chain rule of $\mathrm{I}(\mathbf{\Pi}^*; \mathbf{Y}, \mathbf{X})$, ⑥ is because $\mathbf{\Pi}^*$ and $\mathbf{X}$ are independent and hence $\mathrm{I}(\mathbf{\Pi}^*; \mathbf{X}) = 0$, and ⑦ is because of Lemma 11.

In the end, we present the proof for (13), which proceeds as

$$\begin{aligned}
&\inf_{\widehat{\mathbf{\Pi}}} \sup_{\mathbf{\Pi}^*} \mathbb{E}_{\mathbf{X},\mathbf{W}} \mathsf{d}_{\mathsf{H}}(\widehat{\mathbf{\Pi}}; \mathbf{\Pi}^*) \\
\geq\ & (d+1)\mathrm{Pr}_{\mathbf{X},\mathbf{W},\mathbf{\Pi}^*}\left[\mathsf{d}_{\mathsf{H}}\left(\widehat{\mathbf{\Pi}}; \mathbf{\Pi}^*\right) \geq d+1\right],
\end{aligned}$$

where $\mathbb{E}_{\mathbf{X},\mathbf{W}}$ is the expectation taken w.r.t $\mathbf{X}$ and $\mathbf{W}$, and d is an arbitrary integer between 0 and n. Replacing D with d, we finish the proof with (29). □

APPENDIX E
PROOF OF THEOREM 5

A. Notations

We first define the events $\mathscr{E}_0, \mathscr{E}_1(\delta), \mathscr{E}_2(\delta)$ as

$$\begin{aligned}
\mathscr{E}_0 &\triangleq \bigcap_{i=1}^n \left\{ \left\| \mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{i,:} \right\|_2^2 < \min_{j\neq i} \left\| \mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{j,:} \right\|_2^2 \right\}, \\
\mathscr{E}_1(\delta) &\triangleq \bigcup_{i=1}^n \bigcup_{j\neq i} \left\{ 2\left\langle \mathbf{W}_{i,:},\ \frac{\mathbf{B}^{*\top}(\mathbf{X}_{j,:} - \mathbf{X}_{i,:})}{\|\mathbf{B}^{*\top}(\mathbf{X}_{j,:} - \mathbf{X}_{i,:})\|_2} \right\rangle \geq \delta \right\}, \\
\mathscr{E}_2(\delta) &\triangleq \bigcup_{i=1}^n \bigcup_{j\neq i} \left\{ \left\| \mathbf{B}^{*\top}(\mathbf{X}_{j,:} - \mathbf{X}_{i,:}) \right\|_2 \leq \delta \right\},
\end{aligned}$$

where $\delta > 0$ is an arbitrary positive number. In addition, we define probabilities $\mathscr{P}_1$ and $\mathscr{P}_2$ as

$$\begin{aligned}
\zeta_1 &\triangleq \sum_{i=1}^n \sum_{j\neq i} \mathrm{Pr}_{\mathbf{X},\mathbf{W}} \left(2\left\langle \mathbf{W}_{i,:}, \frac{\mathbf{B}^{*\top}(\mathbf{X}_{j,:} - \mathbf{X}_{i,:})}{\|\mathbf{B}^{*\top}(\mathbf{X}_{j,:} - \mathbf{X}_{i,:})\|_2} \right\rangle \geq \delta \right), \\
\zeta_2 &\triangleq \sum_{i=1}^n \sum_{j\neq i} \mathrm{Pr}_{\mathbf{X}} \left(\left\| \mathbf{B}^{*\top}(\mathbf{X}_{j,:} - \mathbf{X}_{i,:}) \right\|_2 \leq \delta \right).
\end{aligned}$$

B. Roadmap

We start the proof by first outlining the proof strategy.

Stage I: We first show that $\left\{\widehat{\mathbf{\Pi}} \neq \mathbf{I}\right\} \subseteq \overline{\mathscr{E}}_0$.

Stage II: We would like to upper bound the probability of error $\mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}} \neq \mathbf{I})$ by $\Psi(\overline{\mathscr{E}}_0)$. By re-arranging terms, we show $\overline{\mathscr{E}}_0 \subseteq \mathscr{E}_1(\delta) \bigcup \mathscr{E}_2(\delta)$, and separately upper bound $\Psi\left(\mathscr{E}_1(\delta)\right)$ and $\Psi\left(\mathscr{E}_2(\delta)\right)$.

Stage III: Treating the above upper bounds as functions of δ, we complete the proof by choosing δ appropriately and invoking the Condition (14).

C. Proof Details

Proof: Following a similar argument as in Appendix C, we assume that $\mathbf{\Pi}^* = \mathbf{I}$ w.l.o.g. and consider correct recovery $\{\widehat{\mathbf{\Pi}} = \mathbf{I}\}$.

Stage I: We first establish that $\left\{\widehat{\mathbf{\Pi}} \neq \mathbf{I}\right\} \subseteq \overline{\mathscr{E}}_0$ by showing that $\mathscr{E}_0 \subseteq \left\{\widehat{\mathbf{\Pi}} = \mathbf{I}\right\}$. Notice that $\mathscr{E}_0$ can be rewritten as

$$\mathscr{E}_0 = \bigcap_{i=1}^n \bigcap_{j\neq i} \left\{ \left\| \mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{i,:} \right\|_2^2 < \left\| \mathbf{Y}_{i,:} - \mathbf{B}^{*\top}\mathbf{X}_{j,:} \right\|_2^2 \right\}.$$

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

$$\left\| \mathbf{Y} - \widehat{\mathbf{\Pi}}\mathbf{X}\mathbf{B}^* \right\|_2^2 \leq \|\mathbf{Y} - \mathbf{X}\mathbf{B}^*\|_2^2, \tag{30}$$

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 (30). This proves that $\mathscr{E}_0 \subseteq \left\{\widehat{\mathbf{\Pi}}=\mathbf{I}\right\}$.

Stage II: In this stage, we will prove that $\overline{\mathscr{E}}_0 \subseteq \mathscr{E}_1(\delta)\bigcup\mathscr{E}_2(\delta)$. First, we expand $\overline{\mathscr{E}}_0$ as

$$\overline{\mathscr{E}}_0 = \bigcup_{i=1}^{n}\bigcup_{j\neq i}\left\{\left\|\mathbf{Y}_{i,:}-\mathbf{B}^{*\top}\mathbf{X}_{i,:}\right\|_2^2 \geq \left\|\mathbf{Y}_{i,:}-\mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2^2\right\}.$$

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

$$\begin{aligned}&\left\|\mathbf{Y}_{i,:}-\mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2^2 = \left\|\mathbf{B}^{*\top}\mathbf{X}_{i,:}+\mathbf{W}_{i,:}-\mathbf{B}^{*\top}\mathbf{X}_{j,:}\right\|_2^2\\ =\ & \|\mathbf{W}_{i,:}\|_2^2+\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{i,:}-\mathbf{X}_{j,:}\right)\right\|_2^2\\ &+2\left\langle\mathbf{W}_{i,:},\ \mathbf{B}^{*\top}\left(\mathbf{X}_{i,:}-\mathbf{X}_{j,:}\right)\right\rangle. \end{aligned}\tag{31}$$

Hence, the event $\overline{\mathscr{E}}_0$ is equivalent to

$$\begin{aligned}\overline{\mathscr{E}}_0 = \bigcup_{i=1}^{n}\bigcup_{j\neq i}\Bigg\{&2\left\langle\mathbf{W}_{i,:},\frac{\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)}{\left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)\right\|_2}\right\rangle\\ &\geq \left\|\mathbf{B}^{*\top}\left(\mathbf{X}_{j,:}-\mathbf{X}_{i,:}\right)\right\|_2\Bigg\}\subseteq\mathscr{E}_1(\delta)\bigcup\mathscr{E}_2(\delta),\end{aligned}$$

since otherwise we will have the inequality reversed. Hence, we can upper bound $\Psi(\overline{\mathscr{E}}_0)$ as

$$\Psi\left(\overline{\mathscr{E}}_0\right) \leq\ \Psi(\mathscr{E}_1(\delta))+\Psi(\mathscr{E}_2(\delta)) \leq \zeta_1+\zeta_2,$$

where $\Psi(\cdot)$ denotes $\mathbb{E}_{\mathbf{X},\mathbf{W}}\mathbb{1}(\cdot)$, and the terms ζ_1 and ζ_2 can be bounded by Lemma 14 and Lemma 15 (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 ζ_1 as

$$\zeta_1 \leq n^2\ \exp\left(-\frac{16\sigma^2}{8\sigma^2}\log\frac{n}{\epsilon_0}\right)=\epsilon_0^2.\tag{32}$$

At the same time, we can show that ζ_2 is no greater than ϵ_0^2. To invoke Lemma 15, first we need to verify the condition $\delta^2<\alpha_0^2\|\mathbf{B}^*\|_{\mathrm{F}}^2/2$. This is proved by

$$\begin{aligned}\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2} &\overset{①}{\geq} 32\log\left(\frac{n}{\epsilon_0}\right)\left(\frac{n}{\epsilon_0}\right)^{4/\kappa\varrho(\mathbf{B}^*)}\\ &\overset{②}{=} 32\log\left(\frac{n}{\epsilon_0}\right)\left(\frac{n^2}{\alpha_0^{\kappa\varrho(\mathbf{B}^*)}}\right)^{4/\kappa\varrho(\mathbf{B}^*)}\\ &\overset{③}{\geq}\frac{32}{\alpha_0^2}\log\left(\frac{n}{\epsilon_0}\right),\end{aligned}\tag{33}$$

where in ① we use condition (14), in ② we use the definition of $\epsilon_0=\alpha_0^{\kappa\varrho(\mathbf{B}^*)}/n$, and in ③ we use $\alpha_0\in(0,1)$ and $n\geq 1$.

We can then invoke Lemma 15 and bound ζ_2 as

$$\begin{aligned}\zeta_2 &\leq n^2\left(\frac{2\delta^2}{\|\mathbf{B}^*\|_{\mathrm{F}}^2}\right)^{\kappa\varrho(\mathbf{B}^*)/2}\\ &\overset{④}{=} n^2\exp\left[-\frac{\kappa\varrho(\mathbf{B}^*)}{2}\left(\log\left(\frac{\|\mathbf{B}^*\|_{\mathrm{F}}^2}{\sigma^2}\right)-\log\left(32\log\left(\frac{n}{\epsilon_0}\right)\right)\right)\right]\\ &\overset{⑤}{\leq} n^2\exp\left[-\frac{\kappa\varrho(\mathbf{B}^*)}{2}\left(\frac{4}{\kappa\varrho(\mathbf{B}^*)}\log\frac{n}{\epsilon_0}\right)\right]=\epsilon_0^2,\end{aligned}\tag{34}$$

where in ④ we plug in the definition $\delta^2=16\sigma^2\log(n/\epsilon_0)$, and in ⑤ we use Condition (13). Combining the bounds for ζ_1 in (32) and ζ_2 in (34) will complete the proof. □

D. Supporting Lemmas

Lemma 14: It holds that

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

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 we obtain

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

where in ① 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 15: Given that $\|\mathbf{B}^*\|_{\mathrm{F}}^2>2\delta^2/\alpha_0^2$, we have

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

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

Proof: We consider an arbitrary term $\|\mathbf{B}^{*\top}(\mathbf{X}_{i,:}-\mathbf{X}_{j,:})\|_2$ and define $\mathbf{Z}=(\mathbf{X}_{i,:}-\mathbf{X}_{j,:})/\sqrt{2}$ $(i<j)$. 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

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

According to Lemma 2.6 in [28] (which is re-stated in Appendix I herein), this probability can be bounded as

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

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 THEOREM 6

A. Notations

We define the events $\mathscr{E}_i(\cdot)$ $(3\leq i\leq 6)$ as

$$\mathscr{E}_3(h)\triangleq\left\{\left\|P_{\mathbf{\Pi}\mathbf{X}}^{\perp}\mathbf{Y}\right\|_{\mathrm{F}}^2\leq\left\|P_{\mathbf{\Pi}^*\mathbf{X}}^{\perp}\mathbf{Y}\right\|_{\mathrm{F}}^2,\ \mathsf{d}_{\mathsf{H}}(\mathbf{\Pi};\mathbf{\Pi}^*)=h\right\},$$

$$\mathscr{E}_4(t,h)\triangleq\left\{T_{\mathbf{\Pi}}\leq\frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m},\mathsf{d}_{\mathsf{H}}(\mathbf{\Pi};\mathbf{\Pi}^*)=h\right\},$$

$$\mathscr{E}_5(t,h)\triangleq\Big\{\Big\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{Y}\Big\|_{\mathrm{F}}^2-\Big\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\Big\|_{\mathrm{F}}^2\le\frac{2T_{\mathbf{\Pi}}}{3},\ \mathsf{d_H}(\mathbf{\Pi};\mathbf{\Pi}^*)=h\Big\},$$
$$\mathscr{E}_6(t,h)\triangleq\Big\{\Big|\Big\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}\Big\|_{\mathrm{F}}^2-\Big\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\Big\|_{\mathrm{F}}^2\Big|\ge\frac{T_{\mathbf{\Pi}}}{3},\ \mathsf{d_H}(\mathbf{\Pi};\mathbf{\Pi}^*)=h\Big\},\tag{35}$$

where $T_{\mathbf{\Pi}}\triangleq\left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}^2$. Additionally we define $T_i(t,h)$ $(1\le i\le 3)$ as

$$T_1(t,h)\triangleq\exp\left(-t\times\mathbf{snr}/72\right),$$
$$T_2(t,h)\triangleq 2\exp\left(-\left(\frac{t^2\times\mathbf{snr}^2}{mh}\wedge(t\times\mathbf{snr})\right)/288\right),$$
$$T_3(t,h)\triangleq 6r\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}},\tag{36}$$

respectively, where $t<mh$ and r is the rank of $\mathbf{B}^*$.

B. Roadmap

We first restate Theorem 6 with the specific values of c_0,c_1 and c_2.

Theorem: Fix $\epsilon>0$ and let $n>C(\epsilon)$, where $C(\epsilon)>0$ is a positive constant depending only on ϵ. Provided the following conditions hold: (i) $\mathbf{snr}\cdot n^{-\frac{2n}{n-p}}\ge 1$; and (ii)

$$\frac{\log(m\cdot\mathbf{snr})}{380}\ge\left(1+\epsilon+\frac{n}{190(n-p)}\right)\log n+\frac{1}{2}\log r(\mathbf{B}^*),\tag{37}$$

then the ML estimator in (5) gives the ground-truth permutation matrix $\mathbf{\Pi}^*$ with probability exceeding $1-c_2n^{-\epsilon}\left[(n^{\epsilon}-1)^{-1}\vee 1\right]$.

With the requirement $n\ge 2p$ and $r(\mathbf{B}^*)\le(m\wedge p)\le n/2$, we can further relax (37) to

$$\log(m\cdot\mathbf{snr})\ge\ (571+380\epsilon)\log n,$$

which reduces to the form given in Theorem 6. Before proceeding, we give an outline of our proof.

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

$$\left\{\widehat{\mathbf{\Pi}}\ne\mathbf{\Pi}^*\right\}=\bigcup_{\mathbf{\Pi}\ne\mathbf{\Pi}^*}\left\{\left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{Y}\right\|_{\mathrm{F}}^2\le\left\|P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y}\right\|_{\mathrm{F}}^2\right\}=\bigcup_{h\ge 2}\mathscr{E}_3(h),\tag{38}$$

and bound the probability of each individual event in (38).

Stage II: For fixed Hamming distance $\mathsf{d_H}(\mathbf{\Pi};\mathbf{\Pi}^*)=h$, we will prove $\Psi(\mathscr{E}_3(h))\le\sum_{i=1}^3T_i(t,h)+r\exp\left(-n\log\frac{n}{2}\right)$, where r denotes the rank of $\mathbf{B}^*$, and $t>0$ is an arbitrary positive number.

Stage III: Under the condition specified by (17) and $\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\ge 323$, we set t as $\sqrt{m}h\log\left(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\right)/\mathbf{snr}$ and show that

$$\Psi(\mathscr{E}_3(h))\le 9n^{-(1+\epsilon)h}+r\exp\left(-n\log\frac{n}{2}\right).\tag{39}$$

Stage IV: We prove that

$$\mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}}\ne\mathbf{\Pi}^*)\le 10.36\left(\frac{1}{n^{\epsilon}\left(n^{\epsilon}-1\right)}\vee\frac{1}{n^{\epsilon}}\right),$$

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

C. Proof Details

Proof: 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 16 and Lemma 17, respectively.

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

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

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

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

which proves the claim.

Stage II and Stage III: As stated above, the detailed proof can be found in Lemma 16 and Lemma 17.

Stage IV: We have

$$\begin{aligned}
&\mathrm{Pr}_{\mathbf{X},\mathbf{W}}(\widehat{\mathbf{\Pi}}\ne\mathbf{\Pi}^*)\le\sum_{h\ge 2}\binom{n}{h}h!\Psi(\mathscr{E}_3(h))\\
&\overset{①}{\le}\sum_{h\ge 2}\binom{n}{h}h!\left(9n^{-(1+\epsilon)h}+r\exp\left(-n\log\frac{n}{2}\right)\right)\\
&\overset{②}{\le}9\sum_{h\ge 2}n^hn^{-(1+\epsilon)h}+r\sum_{h\ge 2}n!\exp\left(-n\log\frac{n}{2}\right)\\
&\overset{③}{\le}9\sum_{h\ge 2}n^{-\epsilon h}+\ r\sum_{h\ge 2}e\sqrt{n}\ \exp\left(n\log n-n\log\left(\frac{n}{2}\right)-n\right)\\
&\le\frac{9}{n^{\epsilon}\left(n^{\epsilon}-1\right)}+e\sum_{h}rn^{\frac{1}{2}}\exp\left(-n\log\left(\frac{e}{2}\right)\right)\\
&\overset{④}{\le}\frac{9}{n^{\epsilon}\left(n^{\epsilon}-1\right)}+\frac{e}{2}\sum_{h}n^{\frac{3}{2}}\exp\left(-n\log\left(\frac{e}{2}\right)\right)\\
&\le\frac{9}{n^{\epsilon}\left(n^{\epsilon}-1\right)}+\frac{e}{2}n^{\frac{5}{2}}\exp\left(-n\log\left(\frac{e}{2}\right)\right)\\
&\overset{⑤}{\le}\frac{9}{n^{\epsilon}\left(n^{\epsilon}-1\right)}+\frac{e}{2}\exp\left(-\epsilon\log n\right)\\
&\le 10.36\left(\frac{1}{n^{\epsilon}\left(n^{\epsilon}-1\right)}\vee\frac{1}{n^{\epsilon}}\right),
\end{aligned}$$

where in ① we use (39), in ② we use $\frac{n!}{(n-h)!}\le n^h$ and $\frac{n!}{(n-h)!}\le n!$, in ③ we use *Stirling's approximation* in the form $n!\le en^{n+0.5}e^{-n}$, in ④ we use $r\le\min(m,p)$ and $p\le\frac{n}{2}$ (according to our assumption in Section II), and in ⑤, 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 $n\ge 36$; when $\epsilon=1$, we require $n\ge 44$). The proof is hence complete. □

D. Supporting Lemmas

Lemma 16: We have

$$\Psi(\mathscr{E}_3(h)) \leq \sum_{i=1}^{3} T_i(t,h) + r\left(\frac{2}{n}\right)^n,$$

where $\Psi(\cdot)$ denotes $\mathbb{E}_{\mathbf{X},\mathbf{W}}\mathbb{1}(\cdot)$, and $t < mh$ is an arbitrary positive number.

Proof: The proof is completed by the following decomposition, which reads

$$\begin{aligned}\Psi(\mathscr{E}_3(h)) \leq\ & \Psi(\mathscr{E}_4(t,h)) + \Psi\left(\mathscr{E}_3(h)\bigcap\overline{\mathscr{E}}_4(t,h)\right)\\ \overset{①}{\leq}\ & \Psi(\mathscr{E}_4(t,h)) + \Psi\left(\overline{\mathscr{E}}_4(t,h)\bigcap\mathscr{E}_5(t,h)\right)\\ &+ \Psi\left(\overline{\mathscr{E}}_4(t,h)\bigcap\mathscr{E}_6(t,h)\right),\end{aligned}$$

where ① is due to the relation $\mathscr{E}_3(h) \subseteq \mathscr{E}_5(t,h)\bigcup\mathscr{E}_6(t,h)$. A detailed explanation is given as follows. Conditional on $\overline{\mathscr{E}_5}(t,h)\bigcap\overline{\mathscr{E}_6}(t,h)$, we have

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

where in ② we use the fact $P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{Y} = P^{\perp}_{\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}$, and in ③ we use the definitions of $\overline{\mathscr{E}_5}(t,h)$ and $\overline{\mathscr{E}_6}(t,h)$. This suggests that $\overline{\mathscr{E}_5}(t,h)\bigcap\overline{\mathscr{E}_6}(t,h) \subseteq \overline{\mathscr{E}}_3(h)$, which is equivalent to $\mathscr{E}_3(h) \subseteq \mathscr{E}_5(t,h)\bigcup\mathscr{E}_6(t,h)$. We then separately bound the above terms.

Term $\Psi(\mathscr{E}_4(t,h))$**.** We first perform $\mathrm{SVD}\,(\mathbf{B}^*) = \mathbf{U\Sigma V}^{\top}$ (as defined in Appendix A), such that $\mathbf{\Sigma} = \mathrm{diag}\,(\beta_1,\ \beta_2,\ \cdots,\beta_r,0,\cdots)$, 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 $\|P^{\perp}_{\mathbf{X}}\mathbf{\Pi X\Sigma}\|_F^2$. Therefore, we have

$$\begin{aligned}\Psi(\mathscr{E}_4(t,h)) &\leq \sum_{i=1}^{r}\mathrm{Pr}_{\mathbf{X}}\left(\left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{\Pi}^*\mathbf{X}\beta_i\mathbf{e}_i\right\|_{\mathrm{F}}^2 \leq \frac{t\beta_i^2}{m}\right)\\ &\overset{④}{\leq} r\left[\left(\frac{2}{n}\right)^n + 6\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}}\right],\end{aligned} \tag{40}$$

where ④ follows from Lemma 5 in [12].

Term $\Psi\left(\overline{\mathscr{E}}_4(t,h)\bigcap\mathscr{E}_5(t,h)\right)$**.** We 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, $\mathscr{N}(T_{\mathbf{\Pi}}, 4\sigma^2T_{\mathbf{\Pi}})$. Therefore, we obtain

$$\begin{aligned}&\Psi\left(\mathscr{E}_5(t,h)\right)\\ \overset{⑤}{=}\ & \mathbb{E}_{\mathbf{X}}\left[\mathbb{1}\left(T_{\mathbf{\Pi}} > \frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right)\right.\\ &\left.\times\ \mathbb{E}_{\mathbf{W}}\mathbb{1}\left(\left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{Y}\right\|_{\mathrm{F}}^2 - \left\|P^{\perp}_{\mathbf{\Pi X}}\mathbf{W}\right\|_{\mathrm{F}}^2 \leq \frac{2T_{\mathbf{\Pi}}}{3}\right)\right]\\ \overset{⑥}{\leq}\ & \mathbb{E}_{\mathbf{X}}\left[\mathbb{1}\left(T_{\mathbf{\Pi}} > \frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right)\times\exp\left(-\frac{T_{\mathbf{\Pi}}}{72\sigma^2}\right)\right]\\ \leq\ & \exp\left(-\frac{t\times\mathbf{snr}}{72}\right),\end{aligned} \tag{41}$$

where ⑤ results from the independence between $\mathbf{X}$ and $\mathbf{W}$, and in ⑥ we use a standard tail bound for Gaussian random variables.

Term $\Psi\left(\overline{\mathscr{E}}_4(t,h)\bigcap\mathscr{E}_6(t,h)\right)$**.** We have

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

where $\mathbf{\Pi X}\setminus\mathbf{\Pi}^*\mathbf{X}$ ($\mathbf{\Pi}^*\mathbf{X}\setminus\mathbf{\Pi X}$) is the short-hand for range $(\mathbf{\Pi X})\setminus$ range $(\mathbf{\Pi}^*\mathbf{X})$ (range $(\mathbf{\Pi}^*\mathbf{X})\setminus$ range $(\mathbf{\Pi X})$). 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 [12].

We conclude that

$$\begin{aligned}&\Psi\left(\mathscr{E}_6(t,h)\right)\\ \leq\ & 2\mathrm{Pr}_{\mathbf{X},\mathbf{W}}\left\{\left|\left\|P_{\mathbf{\Pi X}\setminus\mathbf{\Pi}^*\mathbf{X}}\mathbf{W}\right\|_{\mathrm{F}}^2 - mk\sigma^2\right| \geq \frac{T_{\mathbf{\Pi}}}{6},\right.\\ &\qquad\left. T_{\mathbf{\Pi}} > \frac{t\|\mathbf{B}^*\|_{\mathrm{F}}^2}{m}\right\}\\ \overset{⑦}{\leq}\ & 2\exp\left(-\frac{1}{8}\left(\frac{t^2\times\mathbf{snr}^2}{36mk}\wedge\frac{t\times\mathbf{snr}}{6}\right)\right)\\ \leq\ & 2\exp\left(-\frac{1}{8}\left(\frac{t^2\times\mathbf{snr}^2}{36mh}\wedge\frac{t\times\mathbf{snr}}{6}\right)\right),\end{aligned} \tag{42}$$

where in ⑦ we use the concentration inequality for χ^2-RVs given in Appendix I, Lemma 26. We complete the proof by combing (40), (41) and (42). □

Lemma 17: Given that $\mathbf{snr}\cdot n^{-\frac{2n}{n-p}} \geq 1$ and $\log(m\cdot\mathbf{snr}) \geq 380\left(1+\epsilon+\frac{n\log n}{190(n-p)}+\frac{1}{2}\log r(\mathbf{B}^*)\right)$, where $\epsilon > 0$ is a constant, we have one positive $0 < t < mh$ such that $\sum_{i=1}^{3}T_i(t,h) \leq 9n^{-(1+\epsilon)h}$.

Proof: We complete the proof by choosing t as $\sqrt{m}h\log\left(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\right)/\mathbf{snr}$ and separately bounding $T_i(t,h)$, $1 \leq i \leq 3$. Before proceed, we first check that $t < mh$, which can be easily verified.

Term $T_1(t,h)$**:** We have

$$\begin{aligned}&\exp\left(-\frac{t\times\mathbf{snr}}{72}\right)\\ =\ & \exp\left(-\frac{\sqrt{m}h}{72}\log\left(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\right)\right)\\ \leq\ & \exp\left(-\frac{h}{72}\log\left(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\right)\right).\end{aligned} \tag{43}$$

Term $T_2(t,h)$**:** Provided that $\left(t^2\times\mathbf{snr}^2/(mh)\right)\wedge(t\times\mathbf{snr}) = t\times\mathbf{snr}$, the term $T_2(t,h)$ is of a similar form as $T_1(t,h)$ in (43). Here we focus on the case in which $\frac{t^2\times\mathbf{snr}^2}{mh}\wedge(t\times\mathbf{snr}) = \frac{t^2\times\mathbf{snr}^2}{mh}$. The right hand side of this equality can

be expanded as

$$\begin{aligned}\frac{t^2\times\mathbf{snr}^2}{mh} &= h\log^2\left(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\right)\\ &\overset{①}{\geq} h\log\left(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\right),\end{aligned}$$

where in ① we use the fact $\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\geq 323$, which can be verified by (17). We then obtain

$$T_2(t,h)\leq 2\exp\left[-\frac{h}{288}\log\left(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\right)\right]. \tag{44}$$

Term $T_3(t,h)$**:** We have

$$\begin{aligned}&r\left[\frac{tn^{\frac{2n}{n-p}}}{mh}\exp\left(1-\frac{tn^{\frac{2n}{n-p}}}{mh}\right)\right]^{\frac{h}{10}}\\ &= r\exp\left[-\frac{h}{10}\left(\log\frac{mh}{tn^{\frac{2n}{n-p}}}+\frac{tn^{\frac{2n}{n-p}}}{mh}-1\right)\right]\\ &\overset{②}{=} r\exp\left[-\frac{h}{10}\left(-\frac{1}{2}\log m-\log\frac{\log z}{z}+\frac{\sqrt{m}\log z}{z}-1\right)\right]\\ &\leq r\exp\left[-\frac{h}{10}\left(-\frac{1}{2}\log m-\log\frac{\log z}{z}+\frac{\log z}{z}-1\right)\right]\\ &\overset{③}{\leq} r\exp\left[-\frac{h}{10}\left(\frac{\log z}{1.9}-\frac{\log m}{2}\right)\right]\\ &\overset{④}{\leq} r\exp\left[-\frac{h}{380}\log(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}})\right],\end{aligned} \tag{45}$$

where in ② we set $z=\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\geq 323$, in ③ 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 ④ we use the fact $\mathbf{snr}\cdot n^{-\frac{2n}{n-p}}\geq 1$.

Combining (43), (44) and (45), we conclude that $\sum_{i=1}^3 T_i(t,h)\leq 9r\exp\left[-\frac{h}{380}\log\left(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\right)\right]$. Under the condition specified by (17), we have

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

Hence, we have

$$\begin{aligned}&r\exp\left[-\frac{h}{380}\log\left(\mathbf{snr}\cdot mn^{-\frac{2n}{n-p}}\right)\right]\\ &\leq r\exp\left[-h(1+\epsilon)\log n-\frac{h}{2}\log r\right]\overset{⑤}{\leq} n^{-(1+\epsilon)h},\end{aligned}$$

where in ⑤ we have $r^{1-\frac{h}{2}}\leq 1$ since $h\geq 2$. This completes the proof. □

APPENDIX G
PROOF OF THEOREM 7

A. Notations

We define events $\mathscr{E}_7(h),\mathscr{E}_8(t,h)$ as

$$\mathscr{E}_7(h)\triangleq\left\{0<\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}^2\leq h\|\mathbf{B}^*\|_{\mathrm{F}}^2,\ \mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi})=h\right\},$$

$$\mathscr{E}_8(t,h)\triangleq\left\{\left\|P_{\mathbf{X}}^{\perp}\frac{(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*}{\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}}\right\|_{\mathrm{F}}^2<\frac{t}{h},\ \mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi})=h\right\}.$$

Furthermore, we define the terms T_4 as

$$T_4(t,h)\triangleq\exp\left(\frac{rh}{2}\left(\log\left(\frac{t}{h}\right)-\frac{t}{h}+1\right)+4.18rh\right),$$

where r denotes the rank of $\mathbf{B}^*$.

B. Proof Outline

We first restate Theorem 7 as the following, where the specific values of c_i $(0\leq i\leq 5)$ are given. Notice that our proof focuses on the order and hence these values are not sharpened to their limits.

Theorem: Suppose that $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}^*)\leq h_{\max}$ with $h_{\max}$ satisfying the relation $h_{\max}r(\mathbf{B}^*)\leq n/8$. Let further $\epsilon>0$ be arbitrary, and suppose that $n>N_2(\epsilon)$, where $N_2(\epsilon)>0$ is a positive constant depending only on ϵ. In addition, suppose that the following conditions hold:

$$(i)\ \mathbf{snr}>26.2,\quad (ii)\ \varrho(\mathbf{B}^*)\geq 5(1+\epsilon)\log n/c_0,$$
$$(iii)\ \log(\mathbf{snr})\geq\frac{288(1+\epsilon)\log n}{\varrho(\mathbf{B}^*)}+33.44.$$

Then the ML estimator (5) subject to the constraint $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi})\leq h_{\max}$ equals $\mathbf{\Pi}^*$ with probability at least $1-10n^{-\epsilon}\left[(n^{\epsilon}-1)^{-1}\vee 1\right]$, where $c_0,\ldots,c_4>0$ are some positive constants.

Here we adopt the same proof strategy as in Theorem 6. For the sake of brevity, we only present the parts that are different compared with the proof of Theorem 6.

Stage I: Given the requirement $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}^*)\leq h_{\max}$, the triangle inequality implies that

$$\mathsf{d}_{\mathsf{H}}(\widehat{\mathbf{\Pi}};\mathbf{\Pi}^*)\leq\mathsf{d}_{\mathsf{H}}(\mathbf{I};\widehat{\mathbf{\Pi}})+\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}^*)\leq 2h_{\max}.$$

Hence, we can confine ourselves to the case in which $\mathsf{d}_{\mathsf{H}}(\mathbf{\Pi};\mathbf{\Pi}^*)\leq 2h_{\max}$.

Stage II: We replace Lemma 16 with Lemma 18.

Stage III: We replace Lemma 17 with Lemma 19.

Stage IV: We use the same argument as Stage IV in proving Theorem 6 and complete the proof.

C. Supporting Lemmas

Lemma 18: Given that $rh\leq n/4$ and $t\leq 0.125h$, we have $\Psi(\mathscr{E}_3(h))\leq T_1(mt,h)+T_2(mt,h)+T_4(t,h)+6\exp\left(-\frac{c_0h\varrho(\mathbf{B}^*)}{5}\right)$, where $h\geq 2$, $\mathscr{E}_3$ is defined in (35), and $T_1(\cdot,\cdot),T_2(\cdot,\cdot)$ are defined in (36).

Proof: Similar to the proof of Lemma 16, we bound $\Psi(\mathscr{E}_3(h))$ by decomposing it as

$$\begin{aligned}\Psi(\mathscr{E}_3(h)) &\leq \Psi(\mathscr{E}_3(h)\bigcap\overline{\mathscr{E}}_4(mt,h))+\Psi(\mathscr{E}_4(mt,h))\\ &\overset{①}{\leq} \Psi(\mathscr{E}_3(h)\bigcap\overline{\mathscr{E}}_4(mt,h))+\Psi(\mathscr{E}_7(h))+\Psi(\mathscr{E}_8(t,h)),\end{aligned}$$

where $\mathscr{E}_3(h),\mathscr{E}_4(mt,h)$ are defined in (35), ① is due to

$$\mathscr{E}_4(mt)=\left\{\left\|P_{\mathbf{X}}^{\perp}\mathbf{\Pi}\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}^2\leq t\|\mathbf{B}^*\|_{\mathrm{F}}^2,\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi})=h\right\},$$

event $\{\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}=0\}$ being with measure zero, and the relation $\left\|P_{\mathbf{X}}^{\perp}\mathbf{\Pi}\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}=\left\|P_{\mathbf{X}}^{\perp}(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\right\|_{\mathrm{F}}$.

Using the same argument as in proving Lemma. 16, we can prove $\Psi(\mathscr{E}_3(h)\bigcap\overline{\mathscr{E}}_4(mt,h)) \le T_1(mt,h)+T_2(mt,h)$. For the clarify of presentation, we leave the proof for $\Psi(\mathscr{E}_7(h))$ and $\Psi(\mathscr{E}_8(t,h))$ to Lemma 20 and Lemma 21, respectively. The proof is completed by summarizing the upper bounds. □

Lemma 19: Given that $\mathbf{snr} > 26.2$, $rh \le n/4$, $t \le 0.125h$, $\varrho(\mathbf{B}^*) \ge 5(1+\epsilon)\log n/c_0$, and

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

we have one positive number $0 < t < 0.125h$ such that

$$\begin{aligned} &T_1(mt,h)+T_2(mt,h)+T_4(t,h)+6\exp\left(-\frac{c_0h\varrho(\mathbf{B}^*)}{5}\right)\\ \le\ & 10n^{-(1+\epsilon)h}, \end{aligned}$$

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

Proof: We complete the proof by choosing $t = h\log(\mathbf{snr})/\mathbf{snr}$. Note that if $\mathbf{snr} > 26.2$, we have $t < 0.125h$. Given (18), we have

$$\log(\mathbf{snr}) \ge \frac{288(1+\epsilon)\log n}{\varrho(\mathbf{B}^*)} \overset{①}{\ge} \frac{288(1+\epsilon)\log n}{m}, \quad (46)$$

where in ① we use $\varrho^*(\mathbf{B}) \le r(\mathbf{B}^*) \le m$. First we verify

$$e^{-c_0h\varrho(\mathbf{B}^*)/5} \le n^{-(1+\epsilon)h}, \quad (47)$$

if $\varrho(\mathbf{B}^*)$ satisfies $\varrho(\mathbf{B}^*) \ge 5(1+\epsilon)\log n/c_0$. In the sequel we will separately bound the terms.

Term $T_1(mt,h)$**:** We have

$$\begin{aligned} \exp\left(-\frac{mt\times\mathbf{snr}}{72}\right) =\ & \exp\left(-\frac{mh}{72}\log(\mathbf{snr})\right)\\ \overset{②}{\le}\ & n^{-(1+\epsilon)h}, \end{aligned} \quad (48)$$

where in ② we use (46).

Term $T_2(mt,h)$**:** Since we have $\mathbf{snr} \ge 26.2$, we obtain $\left(\frac{mt^2\times\mathbf{snr}^2}{h} \wedge (mt\times\mathbf{snr})\right) \ge mh\log(\mathbf{snr})$ and have

$$\begin{aligned} T_2(mt,h) \le\ & 2\exp\left(-\frac{mh}{288}\times\log(\mathbf{snr})\right)\\ \overset{③}{\le}\ & 2n^{-(1+\epsilon)h}, \end{aligned} \quad (49)$$

where in ③ we use (46).

Term $T_4(t,h)$**:** We have

$$\begin{aligned} T_4(t,h) \overset{④}{\le}\ & \exp\left(-\frac{rh}{8}\log(\mathbf{snr}) + 4.18rh\right)\\ \overset{⑤}{\le}\ & n^{-(1+\epsilon)h}, \end{aligned} \quad (50)$$

where in ④ we use $\frac{\log z}{z} - 1 - \log\frac{\log z}{z} \ge \frac{\log z}{4}$, for $z \ge 1.5$, and in ⑤ we use the assumption such that

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

We finish the proof by combining (47), (48), (49), and (50). □

Lemma 20: We bound $\Psi(\mathscr{E}_7(h)) \le 6\exp\left(-\frac{c_0h\varrho(\mathbf{B}^*)}{5}\right)$ for $2 \le h \le n$.

Proof: With $\mathrm{SVD}(\mathbf{B}^*) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we first verify $\|(\mathbf{I}-\mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}} = \left\|(\mathbf{I}-\mathbf{\Pi})\widetilde{\mathbf{X}}\mathbf{\Sigma}\right\|_{\mathrm{F}}$, 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 $\mathbf{X}$. We separately consider the cases where $h = 2$ and $h \ge 3$.

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

$$\begin{aligned} &\mathrm{Pr}_{\mathbf{X}}\left(\|(\mathbf{I}-\mathbf{\Pi})\widetilde{\mathbf{X}}\mathbf{\Sigma}\|_{\mathrm{F}}^2 \le 2\|\mathbf{B}^*\|_{\mathrm{F}}^2\right)\\ =\ & \mathrm{Pr}_{\mathbf{X}}\left[2\sum_{i=1}^r\beta_i^2\left(\widetilde{X}_{1,i}-\widetilde{X}_{2,i}\right)^2 \le 2\left(\sum_{i=1}^r\beta_i^2\right)\right]\\ \overset{①}{=}\ & \mathrm{Pr}_{\mathbf{X}}\left[\sum_{i=1}^r\beta_i^2\widetilde{z}_{1,i}^2 \le \frac{\sum_{i=1}^r\beta_i^2}{2}\right]\\ \overset{②}{=}\ & \mathrm{Pr}_{\mathbf{X}}\left[\left\langle\widetilde{\mathbf{z}},\ \mathbf{\Sigma}^2\widetilde{\mathbf{z}}\right\rangle \le \frac{\sum_{i=1}^r\beta_i^2}{2}\right]\\ \overset{③}{\le}\ & 2\exp\left(-c_0\varrho(\mathbf{B}^*)\right), \end{aligned} \quad (51)$$

where $\mathbf{\Sigma} = \mathrm{diag}(\beta_1,\cdots,\beta_r,0,\cdots)$, 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 ① we define $\widetilde{z}_{1,i} = (\widetilde{X}_{1,i}-\widetilde{X}_{2,i})/\sqrt{2}$, in ② 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 ③ we use Theorem 2.5 in [28] (c.f. also Appendix I) and c_0 is the corresponding constant.

Then we consider the case where $h \ge 3$, by studying the index set $I \triangleq \{j : \boldsymbol{\pi}(j) \ne j\}$, where $\boldsymbol{\pi}(\cdot)$ is the permutation corresponding to the permutation matrix $\mathbf{\Pi}$. Adopting the same argument as in Lemma 8 in [12], we decompose the index set I into 3 subsets $\{I_1, I_2, I_3\}$, such that

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

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

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

In the sequel, we bound $\mathrm{Pr}_{\mathbf{X}}\left(\|\widetilde{\mathbf{Z}}_1\|_{\mathrm{F}}^2 \le h_1\|\mathbf{B}\|_{\mathrm{F}}^2\right)$; the other two probabilities can be bounded similarly. Since j and $\boldsymbol{\pi}(j)$ cannot be in I_1 simultaneously, we define $\widetilde{z}_{j,k} = (\widetilde{X}_{j,k} - \widetilde{X}_{\boldsymbol{\pi}(j),k})/\sqrt{2}$, $j \in I_1, 1 \le k \le r$, and can treat the $\{\widetilde{z}_{jk}\}$ as independent $\mathscr{N}(0,1)$-RVs. Similar to the case $h = 2$, we have

$$\begin{aligned} &\mathrm{Pr}_{\mathbf{X}}\left(\|\widetilde{\mathbf{Z}}_1\|_{\mathrm{F}}^2 \le h_1\|\mathbf{B}\|_{\mathrm{F}}^2\right)\\ \overset{④}{=}\ & \mathrm{Pr}_{\mathbf{X}}\left[\left\langle\widetilde{\mathbf{z}},\mathrm{diag}(\mathbf{\Sigma}^2,\cdots,\mathbf{\Sigma}^2)\widetilde{\mathbf{z}}\right\rangle \overset{⑤}{\le} \frac{h_1(\sum_{i=1}^r\beta_i^2)}{2}\right]\\ \overset{⑥}{\le}\ & 2\exp\left(-c_0h_1\varrho(\mathbf{B}^*)\right) \overset{⑦}{\le} 2\exp\left(-\frac{c_0h\varrho(\mathbf{B}^*)}{5}\right), \end{aligned} \quad (52)$$

where the diagonal matrix $\mathrm{diag}(\mathbf{\Sigma}^2,\cdots,\mathbf{\Sigma}^2)$ in ④ consists of h_1 terms, in ⑤ we define $\widetilde{\mathbf{z}}$ as the vectorization of $\widetilde{\mathbf{Z}}_1$, in ⑥ we use Theorem 2.5 in [28] (also listed in Appendix I), and

in ⑦ we use the fact $h_i \geq \lfloor h/3 \rfloor$. We hence bound $\Psi(\mathscr{E}_7(h))$ by combing the above cases in (51) and (52). □

Lemma 21: We bound

$$\Psi(\mathscr{E}_8(t,h)) \leq \exp\left(\frac{rh}{2}\left(\log\left(\frac{t}{h}\right) - \frac{t}{h} + 1\right) + 4.18rh\right)$$

for $rh \leq \frac{n}{4}$ and $t \leq 0.125h$.

Proof: For ease of notation, we define $\mathbf{\Theta} = (\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*/\|(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}$. Then the probability of the event $\mathscr{E}_8$ can be bounded as

$$\begin{aligned}
\Psi(\mathscr{E}_8) &= \mathrm{Pr}_{\mathbf{X}}\left(\left\|P_X^{\perp}\mathbf{\Theta}\right\|_{\mathrm{F}}^2 < \frac{t}{h}\|\mathbf{\Theta}\|_{\mathrm{F}}^2\right) \\
&= \mathrm{Pr}_{\mathbf{X}}\left(\sum_{i=1}^{r}\left\|P_{\mathbf{X}}^{\perp}\mathbf{\Theta}_{:,i}\right\|_{\mathrm{F}}^2 \leq \frac{t}{h}\|\mathbf{\Theta}_{:,i}\|_{\mathrm{F}}^2\right) \\
&\leq \sum_{i=1}^{r}\mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\mathbf{\Theta}_{:,i}\right\|_{\mathrm{F}}^2 \leq \frac{t}{h}\|\mathbf{\Theta}_{:,i}\|_{\mathrm{F}}^2\right) \\
&\overset{①}{=} \sum_{i=1}^{r}\mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_i\right\|_2^2 \leq \frac{t}{h}\right),
\end{aligned}$$

where in ① 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 $\mathsf{d}_{\mathsf{H}}(\mathbf{I};\mathbf{\Pi}) = h \geq 2$. Before delving into detailed calculations, we first summarize our proof strategy:

- **Step I**: We cover the set Θ_h with an ε-net $\mathscr{N}_\varepsilon$ such that for arbitrary $\boldsymbol{\theta} \in \Theta_h$, there exists a $\boldsymbol{\theta}_0 \in \mathscr{N}_\varepsilon$ such that $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_2 \leq \varepsilon$.
- **Step II**: Setting $\varepsilon = \sqrt{t/h}$, we define events $\mathscr{E}_\Theta$ and $\mathscr{E}_{\mathscr{N}_\varepsilon}$ by

  $$\begin{aligned}
  \mathscr{E}_\Theta &\triangleq \left\{\exists\ \boldsymbol{\theta} \in \Theta_h \ \text{ s.t. } \ \left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}\right\|_2 < \varepsilon = \sqrt{t/h}\right\}, \\
  \mathscr{E}_{\mathscr{N}_\varepsilon} &\triangleq \left\{\exists\ \boldsymbol{\theta}_0 \in \mathscr{N}_\varepsilon \ \text{ s.t. } \ \left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 < 2\varepsilon = 2\sqrt{t/h}\right\}.
  \end{aligned}$$

  Then we will prove

  $$\begin{aligned}
  &\mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\frac{(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*}{\|(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}}\right\|_{\mathrm{F}}^2 < \frac{t}{h}\right) \\
  \leq\ & r\mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_\Theta\right) \leq r\mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_{\mathscr{N}_\varepsilon}\right).
  \end{aligned}$$

- **Step III**: We consider an arbitrary fixed element $\boldsymbol{\theta}_0 \in \mathscr{N}_\varepsilon$, and study $\mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 \leq 2\varepsilon\right)$. Adopting the union bound

  $$\mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_{\mathscr{N}_\varepsilon}\right) \leq |\mathscr{N}_\varepsilon| \times \mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 \leq 2\varepsilon\right),$$

  we finish the bound of $\Psi(\mathscr{E}_8)$.

The following analysis fills in the details.

Stage I: We cover the set Θ_h with an ε-net $\mathscr{N}_\varepsilon$. Its cardinality can be bounded as

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

where in ② we use that elements of Θ_h have at least $(n-h)$ zero elements, and accordingly we cover the sphere $\mathbb{S}^{h-1}$ with an ε-net $\mathscr{N}_\varepsilon$, whose cardinality can be bounded as in [36], and in ③ we assume that $\varepsilon \leq 1$.

Stage II: We will prove the relation

$$\begin{aligned}
&\mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\frac{(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*}{\|(\mathbf{I} - \mathbf{\Pi})\mathbf{X}\mathbf{B}^*\|_{\mathrm{F}}}\right\|_{\mathrm{F}}^2 < \frac{t}{h}\right) \\
\overset{④}{\leq}\ & r\mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_\Theta\right) \overset{⑤}{\leq} r\mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_{\mathscr{N}_\varepsilon}\right),
\end{aligned}$$

when $\varepsilon = \sqrt{t/h}$ and ④ follows from the definition of $\mathscr{E}_\Theta$. We here focus on proving inequality ⑤, which is done by

$$\begin{aligned}
&\mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_\Theta\right) = \mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_\Theta \bigcap \mathscr{E}_{\mathscr{N}_\varepsilon}\right) + \mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_\Theta \bigcap \overline{\mathscr{E}}_{\mathscr{N}_\varepsilon}\right) \\
\leq\ & \mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_{\mathscr{N}_\varepsilon}\right) + \mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_\Theta \bigcap \overline{\mathscr{E}}_{\mathscr{N}_\varepsilon}\right) \overset{⑥}{=} \mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_{\mathscr{N}_\varepsilon}\right),
\end{aligned}$$

where ⑥ is due to the fact $\mathrm{Pr}_{\mathbf{X}}\left(\mathscr{E}_\Theta \bigcap \overline{\mathscr{E}}_{\mathscr{N}_\varepsilon}\right) = 0$. A detailed explanation is given as follows. Note that, given $\overline{\mathscr{E}}_{\mathscr{N}_\varepsilon}$, it holds that for all $\boldsymbol{\theta}_0 \in \mathscr{N}_\varepsilon$, we have $\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 \geq 2\varepsilon$. Then for arbitrary $\boldsymbol{\theta} \in \Theta_h$, we consider an element $\boldsymbol{\theta}_0 \in \mathscr{N}_\varepsilon$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \varepsilon$ and consequently

$$\begin{aligned}
\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}\right\|_2 &\geq \left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 - \left\|P_{\mathbf{X}}^{\perp}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right\|_2 \\
&\geq 2\varepsilon - \left\|P_{\mathbf{X}}^{\perp}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right\|_2 \\
&\overset{⑦}{\geq} 2\varepsilon - \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \overset{⑧}{\geq} \sqrt{\frac{t}{h}},
\end{aligned}$$

where in ⑦ we use the contraction property of projections, and in ⑧ the fact $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \varepsilon = \sqrt{t/h}$.

Stage III: We study the probability $\mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2^2 \leq \frac{4t}{h}\right)$ for fixed $\boldsymbol{\theta}_0 \in \mathscr{N}_\varepsilon$. In virtue of results in [37], we have

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

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

$$\begin{aligned}
&\mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2^2 \leq \frac{4t}{h}\right) \\
=\ & \mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2^2 \leq \frac{\alpha(n-p)}{n}\right) \\
\leq\ & \exp\left(\frac{n-p}{2}\left(\log\left(\frac{4nt}{(n-p)h}\right) - \frac{4nt}{(n-p)h} + 1\right)\right) \\
\overset{⑨}{\leq}\ & \exp\left(\frac{n}{4}\left(\log\left(\frac{8t}{h}\right) - \frac{8t}{h} + 1\right)\right),
\end{aligned}$$

where in ⑨ we use that (a) $n \geq 2p$, (b) $\log x - x + 1$ is increasing in range $(0,1)$, and (c) $\log x + 1 \leq x$.

In the end, we bound $\Psi(\mathscr{E}_8)$ as

$$\begin{aligned}
&\Psi(\mathscr{E}_8) \\
\leq\ & r\left(\frac{3}{\sqrt{t/h}}\right)^h \exp\left(\frac{n}{4}\left(\log\left(\frac{8t}{h}\right) - \frac{8t}{h} + 1\right)\right) \\
=\ & \exp\Bigg[h\log(3) - \frac{h}{2}\log\left(\frac{t}{h}\right) + \log r \\
&\quad + \frac{n}{4}\left(\log\left(\frac{8t}{h}\right) - \frac{8t}{h} + 1\right)\Bigg]
\end{aligned}$$

$$\begin{aligned}&\overset{\text{Ⓐ}}{\leq}\ \exp\left[\frac{rh}{2}\left(\log\left(\frac{t}{h}\right)-\frac{16t}{h}+1\right)+3.68rh+\log r\right]\\&\overset{\text{Ⓑ}}{\leq}\ \exp\left[\frac{rh}{2}\left(\log\left(\frac{t}{h}\right)-\frac{t}{h}+1\right)+4.18rh\right],\end{aligned}\tag{53}$$

where in Ⓐ we use the assumption that $n \geq 4rh$, and in Ⓑ we use that $rh \geq 2r \geq 2\log(r)$. Combining (51), (52) and (53), we finish the proof. □

Remark 22: Note that we cannot improve h from $\Omega\left(\frac{n}{\log n}\right)$ 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 $\Psi(\mathscr{E}_8)$ is based on the step

$$\begin{aligned}&\Psi(\mathscr{E}_8) \leq \mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}\right\|_2 \leq \sqrt{\frac{t}{h}},\ \ \exists\ \boldsymbol{\theta} \in \Theta_h\right)\\&\leq\ |\mathscr{N}_\varepsilon| \cdot \mathrm{Pr}_{\mathbf{X}}\left(\left\|P_{\mathbf{X}}^{\perp}\boldsymbol{\theta}_0\right\|_2 \leq \sqrt{\frac{t}{h}}+\varepsilon,\ \exists\ \boldsymbol{\theta} \in \mathscr{N}_\varepsilon\right) < 1.\end{aligned}$$

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

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

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

Appendix H
Proof of Theorem 9

Proof: Define the function $f(\mathbf{\Pi}) \triangleq \|P_{\mathbf{X}}^{\perp}\mathbf{\Pi}^{\top}\mathbf{Y}\|_{\mathrm{F}}^2$. We would like to show that the sequence of function values generated by Algorithm 1 is non-increasing, i.e., $f(\mathbf{\Pi}^{(t+1)}) \leq f(\mathbf{\Pi}^{(t)})$. Since the function $f(\cdot)$ is quadratic w.r.t. $\mathbf{\Pi}$, we conclude

$$\begin{aligned}&f\left(\mathbf{\Pi}^{(t+1)}\right)\\&=\ f\left(\mathbf{\Pi}^{(t)}\right)+\left\langle -2\mathbf{Y}\mathbf{Y}^{\top}\mathbf{\Pi}^{(t)}P_{\mathbf{X}}, \mathbf{\Pi}^{(t+1)}-\mathbf{\Pi}^{(t)}\right\rangle\\&\qquad -\left\|P_{\mathbf{X}}\left(\mathbf{\Pi}^{(t)}-\mathbf{\Pi}^{(t+1)}\right)^{\top}\mathbf{Y}\right\|_{\mathrm{F}}^2\\&\leq\ f\left(\mathbf{\Pi}^{(t)}\right)+\left\langle -2\mathbf{Y}\mathbf{Y}^{\top}\mathbf{\Pi}^{(t)}P_{\mathbf{X}}, \mathbf{\Pi}^{(t+1)}-\mathbf{\Pi}^{(t)}\right\rangle.\end{aligned}$$

Recalling the definition of $\mathbf{D}^{(t+1)}$ in (20), we can rewrite $-2\mathbf{Y}\mathbf{Y}^{\top}\mathbf{\Pi}^{(t)}P_{\mathbf{X}}$ as $\alpha^{-1}\left(\mathbf{\Pi}^{(t)}-\mathbf{D}^{(t+1)}\right)$, and obtain

$$\begin{aligned}f\left(\mathbf{\Pi}^{(t+1)}\right)-f\left(\mathbf{\Pi}^{(t)}\right) &\leq\ \frac{1}{\alpha}\left\langle \mathbf{\Pi}^{(t)}-\mathbf{D}^{(t+1)}, \mathbf{\Pi}^{(t+1)}-\mathbf{\Pi}^{(t)}\right\rangle\\&\overset{①}{\leq}\ \frac{1}{\alpha}\left\langle \mathbf{\Pi}^{(t)}, \mathbf{\Pi}^{(t+1)}-\mathbf{\Pi}^{(t)}\right\rangle \overset{②}{\leq} 0,\end{aligned}$$

which completes the proof. In ①, we use the property of the projection step (21), namely, $\left\langle \mathbf{D}^{(t+1)}, \mathbf{\Pi}^{(t+1)}\right\rangle \geq \left\langle \mathbf{D}^{(t+1)}, \mathbf{\Pi}^{(t)}\right\rangle$; in ②, we use the Cauchy-Schwarz inequality such that $\left\langle \mathbf{\Pi}^{(t+1)}, \mathbf{\Pi}^{(t)}\right\rangle \leq \left\|\mathbf{\Pi}^{(t)}\right\|_{\mathrm{F}}\left\|\mathbf{\Pi}^{(t+1)}\right\|_{\mathrm{F}} = \left\|\mathbf{\Pi}^{(t)}\right\|_{\mathrm{F}}^2$. □

Appendix I
Useful Facts

Theorem 23 (Theorem 2.35 in [38]): For a $p \times m$ matrix $\mathbf{B}$ whose entries are independent zero-mean real RVs with variance p^{-1} and fourth moment of order $O(p^{-2})$, we have the empirical distribution of the eigenvalues of $\mathbf{B}^{\top}\mathbf{B}$ converge to the distribution with density

$$\begin{aligned}f_\tau(x) =\ &\left[0 \vee \left(1-\tau^{-1}\right)\right]\delta(x)\\&+\frac{\sqrt{\left[\left((1+\sqrt{\tau})^2-x\right)\vee 0\right]\times\left[\left(x-(1-\sqrt{\tau})^2\right)\vee 0\right]}}{2\pi\tau x},\end{aligned}$$

as $m, p \to \infty$ with $m/p \to \tau$.

Define $\vartheta(\tau,\sigma)$ as

$$\vartheta(\tau,\sigma) \triangleq \left(\sqrt{\left(\sqrt{\tau}+1\right)^2+\sigma^2}-\sqrt{\left(1-\sqrt{\tau}\right)^2+\sigma^2}\right)^2.$$

With the above theorem, we can obtain a more accurate formula of $\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2\right)}{\log n}$ when $B_{ij} \overset{i.i.d}{\sim} \mathscr{N}(0,p^{-1})$, which reads

$$\begin{aligned}&\frac{\log\det\left(\mathbf{I}+\mathbf{B}^{*\top}\mathbf{B}^*/\sigma^2\right)}{\log n}\\&\to\ \frac{p\tau}{\log n}\int_{\left(1-\sqrt{\tau}\right)^2}^{\left(1+\sqrt{\tau}\right)^2} f_\tau(x)\log\left(1+\frac{x}{\sigma^2}\right)dx\\&=\ -\frac{p\vartheta(\tau,\sigma)}{4\log n}+\frac{p}{\log n}\log\left(1+\frac{\tau}{\sigma^2}-\frac{\vartheta(\tau,\sigma)}{4\sigma^2}\right)\\&\quad+\frac{p\tau}{\log n}\log\left(1+\frac{1}{\sigma^2}-\frac{\vartheta(\tau,\sigma)}{4\sigma^2}\right),\end{aligned}\tag{54}$$

when $p, m \to \infty$ with $m/p \to \tau$ ($\tau > 0$ is some constant). The comparison with the theoretical values of (54) and the numerical values are plotted in Figure 2, which are almost identical. For more details on the calculation above, readers are referred to P. 8, (1.14) in [38].

Lemma 24 (Theorem 2.5 in [28]): 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

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

Lemma 25 (Lemma 2.6 in [28]): 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

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

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 26 ([36] (Example 2.11, P. 29)): For a χ^2-RV Y with ℓ degrees of freedom, we have

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

REFERENCES

[1] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. New York, NY, USA: Academic, 1979.

[2] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Denoising linear models with permuted data," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 446–450.

[3] Y. Shi and Y. Shi, "Learning to branch-and-bound for header-free communications," in *Proc. IEEE Globecom Workshops*, Waikoloa, HI, USA, Jun. 2019, pp. 1–6.

[4] M. Slawski and E. Ben-David, "Linear regression with sparsely permuted data," *Electron. J. Statist.*, vol. 13, no. 1, pp. 1–36, Jan. 2019.

[5] L. Sweeney, "Computational disclosure control: A primer on data privacy protection," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, 2001.

[6] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy*, Oakland, CA, USA, May 2008, pp. 111–125.

[7] X. Shi, X. Li, and T. Cai, "Spherical regression under mismatch corruption with application to automated knowledge translation," *J. Amer. Stat. Assoc.*, vol. 4, pp. 1–12, May 2020.

[8] E. Grave, A. Joulin, and Q. Berthet, "Unsupervised alignment of embeddings with wasserstein procrustes," in *Proc. 22nd Int. Conf. Artif. Intell. Statist. (AISTATS)*, Okinawa, Japan, 2019, pp. 1880–1890.

[9] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, "Unlabeled sensing: Solving a linear system with unordered measurements," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, 2015, pp. 786–793.

[10] I. Dokmanic, "Permutations unlabeled beyond sampling unknown," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 823–827, Jun. 2019.

[11] M. C. Tsakiris, "Determinantal conditions for homomorphic sensing," 2018, *arXiv:1812.07966*.

[12] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Linear regression with shuffled data: Statistical and computational limits of permutation recovery," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3286–3300, May 2018.

[13] D. J. Hsu, K. Shi, and X. Sun, "Linear regression without correspondence," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 1531–1540.

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

[15] M. Slawski, E. Ben-David, and P. Li, "A two-stage approach to multivariate linear regression with sparsely mismatched data," 2019, *arXiv:1907.07148*.

[16] M. Slawski, M. Rahmani, and P. Li, "A sparse representation-based approach to linear regression with partially shuffled labels," in *Proc. 25th Conf. Uncertainty in Artif. Intell. (UAI)*, Tel Aviv, Israel, 2019, p. 7.

[17] S. Haghighatshoar and G. Caire, "Signal recovery from unlabeled samples," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1242–1257, Mar. 2018.

[18] M. C. Tsakiris and L. Peng, "Homomorphic sensing," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, pp. 6335–6344.

[19] L. Peng and M. C. Tsakiris, "Linear regression without correspondences via concave minimization," 2020, *arXiv:2003.07706*.

[20] H. Zhang and P. Li, "Sparse recovery with shuffled labels: Statistical limits and practical estimators," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, VIC, Australia, Jul. 2021, pp. 1760–1765.

[21] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval, "Compressed sensing with unknown sensor permutation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 1040–1044.

[22] H. Zhang and P. Li, "Optimal estimator for unlabeled linear regression," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, Virtual Event, 2020, pp. 11153–11162.

[23] J. A. Tropp, "An introduction to matrix concentration inequalities," *Found. Trends Mach. Learn.*, vol. 8, pp. 1–230, May 2015.

[24] R. Burkard, M. Dell'Amico, and S. Martello, *Assignment Problems, Revised Reprint*, vol. 106. Philadelphia, PA, USA: SIAM, 2012.

[25] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 4th ed. Baltimore, MD, USA: Johns Hopkins Univ. Press, 2013.

[26] T. M. Cover and J. A. Thomas, *Elements Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2012.

[27] Z. D. Bai and Y. Q. Yin, "Limit of the smallest eigenvalue of a large dimensional sample covariance matrix," *Ann. Probab.*, vol. 21, no. 3, pp. 1275–1294, 1993. [Online]. Available: http://www.jstor.org/stable/2244575

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

[29] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[30] D. P. Bertsekas and D. A. Castañon, "A forward/reverse auction algorithm for asymmetric assignment problems," *Comput. Optim. Appl.*, vol. 1, no. 3, pp. 277–297, Dec. 1992.

[31] T. C. Koopmans and M. Beckmann, "Assignment problems and the location of economic activities," *Econ., J. Econ. Soc.*, vol. 4, pp. 53–76, Jan. 1957.

[32] S. Umeyama, "An eigendecomposition approach to weighted graph matching problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMT-10, no. 5, pp. 695–703, Sep. 1988.

[33] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[34] S. Boucheron and M. Thomas, "Concentration inequalities for order statistics," *Electron. Commun. Probab.*, vol. 17, pp. 1–12, Jan. 2012.

[35] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-Gaussian concentration," *Electron. Commun. Probab.*, vol. 18, pp. 1–9, Jan. 2013.

[36] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge, U.K.: Cambridge Univ. Press, 2019.

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

[38] A. M. Tulino, S. Verdú, and S. Verdu, *Random Matrix Theory Wireless Communication*. New York, NY, USA: Now, 2004.

Hang Zhang is currently a postdoctoral researcher at Baidu Research. His research interests are in the areas of high-dimensional statistics, cost-accuracy trade-offs, robust and efficient algorithms. He was a research intern with Baidu Research in the summers of 2018, 2019, and 2020, while pursuing his PhD degree in the ECE Department at the Georgia Institute of Technology.

Martin Slawski received degrees in Statistics from Ludwig-Maximilans-University Munich, Germany, and the Ph.D. degree in Computer Science from Saarland University, Germany. He is an assistant professor in the Department of Statistics, George Mason University. His research interests include Record Linkage, Data Compression, High-Dimensional Data, and the Interface between Statistics and Optimization. He was a consulting researcher for Baidu Research.

Ping Li (Member, IEEE) received his Ph.D. in Statistics, from Stanford University, where he also earned two master's degrees in Computer Science and in Electrical Engineering. Prior to Stanford, Ping Li gradated two master's degrees from University of Washington, in Earthquake Geotechnical Engineering and in Applied Mathematics. Prior to graduate schools, he studied several majors including Computer Science, Architectural/Structural Engineering, Engineering Geology & Hydrogeology, etc. Before moving to Baidu Research - Bellevue WA, Ping Li taught at Rutgers University and Cornell University. Together with co-authors, Ping Li won best papers, best paper honorable mentions, or best student papers, from KDD, ASONAM, SIGIR, NIPS, etc. Ping Li was a recipient of the Young Investigator Award from the Air Force Office of Scientific Research (AFOSR-YIP), and also a recipient of the Young Instigator Award from the Office of Naval Research (ONR-YIP).