A simple algorithm for r-local and k-sparse unlabeled sensing

Ahmed Ali Abbasi, Abiy Tasissa, Shuchin Aeron

Abstract—The unlabeled sensing problem is to recover an unknown signal from permuted linear measurements. We propose an alternating minimization algorithm with a suitable initialization for the widely considered k-sparse permutation model. Assuming either a Gaussian measurement matrix or a sub-Gaussian signal, we upper bound the initialization error for the r-local and k-sparse permutation models in terms of the block size r and the number of shuffles k, respectively. Our algorithm is computationally scalable and, compared to baseline methods, achieves superior performance on real and synthetic datasets.

Index Terms—unlabeled sensing, linear regression without correspondence, r-local, k-sparse

I. INTRODUCTION

THE standard least square problem is $\min_{\mathbf{X}} \quad \|\mathbf{Y} - \mathbf{BX}\|^2$, where $\mathbf{Y} \in \mathbb{R}^{n \times m}$ denotes the measurements of unknown $\mathbf{X}^* \in \mathbb{R}^{d \times m}$, observed via the linear sensing operator $\mathbf{B}$. In the unlabeled sensing problem, the measurements $\mathbf{Y}$ are scrambled by an unknown permutation $\mathbf{P}^*$

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

where $\mathbf{P}^* \in \mathbb{R}^{n \times n}$ is an unknown permutation matrix and $\mathbf{W} \in \mathbb{R}^{n \times m}$ is additive Gaussian noise with $\mathbf{W}_{ij} = \mathcal{N}(0, \sigma^2)$. Given scrambled measurements $\mathbf{Y}$ and the sensing matrix $\mathbf{B}$, the problem is to estimate the signal $\mathbf{X}^*$. The problem is referred to as single-view (multi-view) unlabeled sensing when $m = 1 (m > 1)$.

The work in [1] was the first to consider the unlabeled sensing problem and establish that $n = 2d$ measurements are necessary and sufficient for recovery of $\mathbf{x}^*$. Subsequent works [2], [3] have generalized this result and developed information theoretic inachievability analysis. Because estimating the unknown permutation matrix $\mathbf{P}^*$ is a challenging problem, several works [4]–[9] assume a k-sparse, or partially shuffled, permutation structure. [10], [11] consider an r-local, or block diagonal, permutation structure. A permutation matrix $\mathbf{P}^*_k$ is k-sparse if it has k off-diagonal elements, i.e., $\langle \mathbf{I}, \mathbf{P}^*_k \rangle = n - k$. A permutation matrix $\mathbf{P}^*_r$ is r-local with s blocks if $\mathbf{P}^*_r = \text{blkidag}(P_1, \cdots, P_s)$. Fig. 1 illustrates the two models. The two models are compared by information-theoretic inachiveability in [11].

For single-view unlabeled sensing, algorithms based on branch and bound and expectation maximization are proposed in [12]–[14]. These algorithms are applicable to small problem sizes. A stochastic alternating minimization approach is proposed in [15]. For multi-view unlabeled sensing, the Levsort subspace matching algorithm was proposed in [16]. A subspace clustering approach was proposed in [7]. The works

[4], [5] propose methods based on bi-convex optimization and robust regression, respectively. A spectral initialization based method was proposed recently in [17].

A. Contributions

We extend our alternating minimization algorithm for the r-local unlabeled sensing model [10], [11] to the widely considered k-sparse model, Section II. We propose and analyze the initialization to our algorithm, Section III. Specifically, assuming random $\mathbf{B}$ or $\mathbf{X}^*$, we derive upper bounds on the initialization error for both r-local and k-sparse models. Numerical results and comparisons to baseline methods showing superior performance are given in Section IV.

The main part of the analysis is relating the error of a linear least-squares system to a linear system obtained by summing a subset of the rows. This is a problem studied under the topic of sketching linear systems [18]. In contrast to the standard setting for sketching, our original linear system contains an unknown permutation $\mathbf{P}^*$ and our sampling is deterministic resulting from the permutation model. Our analysis uses tools from high-dimensional probability and recent results for random quadratic forms of a positive semi-definite matrix. We remark that the initialization analysis is insightful because the formulated optimization problem is non-convex. We think that the presented analysis is general and might be applicable to other class of under-determined linear inverse problems.

B. Applications of unlabeled sensing

The linked linear regression problem [19]–[21], or linear regression under blocking, is to fit a linear model by minimizing over $\mathbf{x}$ and $\mathbf{P}$ the objective $\|\mathbf{y} - \mathbf{PBx}\|_2^2$, where $\mathbf{P}$ is an unknown r-local permutation. For example, consider a simple statistical model where the weight of an individual depends linearly on age and height. The vector $\mathbf{y} \in \mathbb{R}^n$ contains the weight of $n = 10$ individuals, the matrix $\mathbf{B} \in \mathbb{R}^{n \times d}$, $d = 2$ contains the values of the age, height of the individuals. The 10 individuals comprise 5 males, 5 females, and each record (weight, age, height) is known to belong to a male or a female individual. Letting the first (second) block of 5 rows of $\mathbf{y}, \mathbf{B}$ contain the records of male (female) individuals, the unknown r-local permutation, $r = 5$, assigns each weight value (row of $\mathbf{y}$) to its corresponding age, height (row of $\mathbf{B}$).

For the pose and correspondence problem [22], the rows of $\mathbf{B} \in \mathbb{R}^{n \times d}$ contain the d-dimensional coordinates of the landmark points of an object. The object is transformed by an unknown linear transformation $\mathbf{X}^*$ and the rows of $\mathbf{Y} = \mathbf{P}^* \mathbf{B} \mathbf{X}^*$ contain the coordinates of the points of the
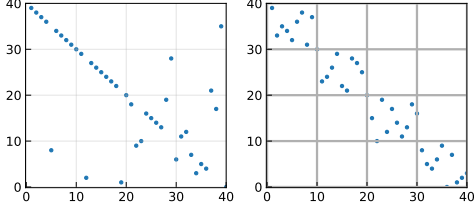
Fig. 1: Left. Sparse (or partially shuffled) permutation considered in [4]–[7], with number of shuffles $k = 10$. Right. The r-local permutation structure considered in [10], [11], with block size $r = 10$. In this paper, we propose and analyze a general algorithm that is applicable to both permutation models.

transformed object. The pose and correspondence problem is to assign a point in $\mathbf{B}$ to the corresponding point in $\mathbf{Y}$ and also recover the unknown transformation $\mathbf{X}^*$. Another application of unlabeled sensing is signal amplitude estimation from binary quantized unlabeled samples [23]. The unknown scalar quantity x is the signal amplitude and the sensing matrix $\mathbf{B}$ is replaced by a threshold function that models bit quantization. The image correspondence problem [24] and the 1-d unassigned distance geometry problem (uDGP) were recently formulated as unlabeled sensing problems in [11].

C. Notation and definitions

$\mathbf{P}_r^*$ denotes r-local permutations with block size r. $\mathbf{P}_k^*$ denotes k-sparse permutations with k off-diagonal elements. $\|X\|_{\varphi_1}$ denotes the sub-exponential norm of the sub-exponential random variable X. $\|X\|_{\varphi_2}$ denotes the sub-Gaussian norm of the sub-Gaussian random variable X. $c' \in \mathbb{R}$ denotes an absolute constant. Π_n denotes the set of $n \times n$ permutations. The Hamming distortion d_H between two permutation matrices is the number of mismatches $d_H = \Sigma_i \mathbb{1}(\widehat{\mathbf{P}}(i) \neq \mathbf{P}(i))$.

II. ALGORITHM

We propose to estimate the signal $\mathbf{X}^*$ and the permutation $\mathbf{P}^*$ by minimizing the forward error in the model (1)

$$\underset{\mathbf{X}, \mathbf{P} \in \Pi_n}{\operatorname{argmin}} \ F(\mathbf{X}, \mathbf{P}) = \|\mathbf{Y} - \mathbf{PBX}\|_F^2. \tag{2}$$

Alternating minimization (alt-min) updates to minimize the objective in (2) for $\mathbf{P}, \mathbf{X}$ are

$$\mathbf{P}^{(t)} = \underset{\mathbf{P} \in \Pi_n}{\operatorname{argmin}} \ \langle -\mathbf{Y}\mathbf{X}^{(t)\mathsf{T}}\mathbf{B}^{\mathsf{T}}, \mathbf{P}\rangle, \tag{3}$$

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \ F(\mathbf{X}, \mathbf{P}^{(t)}) = (\mathbf{P}^{(t)}\mathbf{B})^{\dagger}\mathbf{Y}. \tag{4}$$

When the unknown permutation $\mathbf{P}^*$ is r-local, the $\mathbf{P}$ update in (3) decouples along the blocks of $\mathbf{P}^*$ and reduces to the simpler updates in line 9 of Algorithm 1.

The optimization problem in (2) is non-convex as the set of permutations is a discrete (non-convex) set. Therefore, the initialization to the updates (3), (4) has to be chosen carefully. We propose and analyze initializations for both r-local and k-sparse models in the following sections. A key contribution of our algorithm is that the model assumption is incorporated in the initialization.

Algorithm 1 Alt-min

Input: mode $\in \{r\text{-local}, k\text{-sparse}\}$, convergence threshold ϵ
1: **if** mode is r-local **then**
2: $\quad \widehat{\mathbf{X}} \leftarrow$ collapsed initialization in (7)
3: $\quad \widehat{\mathbf{Y}} \leftarrow \mathbf{B}\widehat{\mathbf{X}}$
4: **else**
5: $\quad \widehat{\mathbf{Y}} \leftarrow \mathbf{Y}$
6: **while** relative change $> \epsilon$ **do**
7: $\quad$ **if** mode is r-local **then**
8: $\quad\quad$ **for** $i \in 1 \cdots s$ **do** $\quad\quad$ //s is the number of blocks
9: $\quad\quad\quad \widehat{\mathbf{P}}_i \leftarrow \operatorname{argmin}_{\mathbf{P}_i \in \Pi_r} - \langle \mathbf{Y}_i\widehat{\mathbf{Y}}_i^{\mathsf{T}}, \mathbf{P}_i\rangle$
10: $\quad\quad \widehat{\mathbf{P}} \leftarrow \operatorname{diag}(\widehat{\mathbf{P}}_1, \cdots, \widehat{\mathbf{P}}_s)$
11: $\quad$ **else**
12: $\quad\quad \widehat{\mathbf{P}} = \operatorname{argmin}_{\mathbf{P} \in \Pi_n} - \langle \mathbf{Y}\widehat{\mathbf{Y}}^{\mathsf{T}}, \mathbf{P}\rangle$
13: $\quad \widehat{\mathbf{X}} \leftarrow \mathbf{B}^{\dagger}\widehat{\mathbf{P}}^{\mathsf{T}}\mathbf{Y}$
14: $\quad \widehat{\mathbf{Y}} \leftarrow \mathbf{B}\widehat{\mathbf{X}}$
15: **Return** $\widehat{\mathbf{P}}, \widehat{\mathbf{X}}$

A. Initialization for the r-local model

For r-local permutation $\mathbf{P}_r^*$, we propose to initialize $\widehat{\mathbf{X}}$ by the *collapsed* initialization (7). Let P_i denote each $r \times r$ block of $\mathbf{P}_r^* = \operatorname{diag}(P_1, \cdots, P_s)$. Let $B_i \in \mathbb{R}^{r \times d}$ denote blocks of the sensing matrix $\mathbf{B} = [B_1; \cdots ; B_s]$. The r permuted measurements in each block $[Y_i : P_iB_i]$ can be summed to extract one labelled measurement on $\mathbf{X}^* \in \mathbb{R}^{d \times m}$

$$\sum_{j=1}^{j=r}[P_iB_i(j)]^{\mathsf{T}}\mathbf{X}^* = \sum_{j=1}^{j=r}Y_i(j) \ \forall i \in [s], \tag{5}$$

where $B_i(j)$ denotes row j of block B_i. Assuming all blocks have the same size r, the number of blocks is $s = n/r$. The s labelled measurements in (5) are represented in the $s \times d$ *collapsed* linear system of equations

$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{B}}\mathbf{X}^*, \tag{6}$$

where each row $\tilde{\mathbf{b}}_i = \sum_{j=1}^{j=r} B_i(j) \ \forall i \in [s]$ is the sum of the r rows in block B_i of the collapsed matrix $\widetilde{\mathbf{B}} \in \mathbb{R}^{s \times d}$. The proposed initialization $\widehat{\mathbf{X}}$ is the minimum-norm solution to (6)

$$\widehat{\mathbf{X}} = \widetilde{\mathbf{B}}^{\dagger}\widetilde{\mathbf{Y}}. \tag{7}$$

Substituting the SVD form of $\widetilde{\mathbf{B}}^{\dagger} = \widetilde{\mathbf{V}}\mathbf{S}^{-1}\widetilde{\mathbf{U}}^{\mathsf{T}}$, where $\widetilde{\mathbf{B}} = \widetilde{\mathbf{U}}\mathbf{S}\widetilde{\mathbf{V}}^{\mathsf{T}}$, shows that $\widehat{\mathbf{X}}$ is the projection of $\mathbf{X}^*$ onto the the row space of $\widetilde{\mathbf{B}}$. Formally,

$$\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_2 = \|(\mathbf{I} - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^{\mathsf{T}})\mathbf{X}^*\|_2. \tag{8}$$

For $s \geq d$, $\widehat{\mathbf{X}} = \mathbf{X}^*$. In section III-A, we upper bound the error in initialization $\widehat{\mathbf{X}}$ for the under-determined $s < d$ case.

B. Initialization for the k-sparse model

For the case of k-sparse permutation $\mathbf{P}_k^*$, we propose to initialize $\widehat{\mathbf{P}}^{(0)} = \mathbf{I}$, which sets $\widehat{\mathbf{Y}}^{(0)} = \mathbf{Y} = \mathbf{P}_k^*\mathbf{B}\mathbf{X}^*$. Since the permutation matrix $\mathbf{P}_k^*$ has k off-diagonal elements, the initialization $\widehat{\mathbf{P}}^{(0)} = \mathbf{I}$ is close to the true solution $\mathbf{P}_k^*$ as $d_H(\widehat{\mathbf{P}}^{(0)}, \mathbf{P}_k^*) = k$. The initialization error is upper bounded in section III-B.

III. Initialization analysis

A. r-local model

Assuming random models on either $\mathbf{B}$ or $\mathbf{X}^*$, the results in this section upper bound the error in the proposed initialization (7). Without any assumptions, the worst case error

$$\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 \leq \frac{\|\mathbf{y}\|_2^2 - \sigma_{\min}^2(\mathbf{B})\|\hat{\mathbf{x}}\|_2^2}{\sigma_{\min}^2(\mathbf{B})} \qquad (9)$$

is achieved for $\mathbf{x}^*$ aligned with the direction of the singular vector corresponding to the smallest singular value $\sigma_{\min}(\mathbf{B})$. Assuming entries of $\mathbf{x}^*$ are zero-mean, independent, sub-Gaussian random variables, Theorem 2.1 in [25] shows that the error $\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2$ is sub-Gaussian. In contrast, Theorem 1 in this work shows a sub-Gaussian error distribution for fixed $\mathbf{x}^*$ and Gaussian $\mathbf{B}$. Theorem 2 shows that the error is sub-Gaussian, assuming fixed $\mathbf{B}$ and sub-Gaussian $\mathbf{x}^*$. Particularly, our result in Theorem 2 does not assume that the entries of $\mathbf{x}^*$ are independently drawn.

Lemma 1. *Let $\hat{\mathbf{x}}$ be as defined in (7). Let $\mathbf{x}^* \in \mathbb{R}^d$ be the fixed unknown vector, and let $\mathbf{B}$ be a Gaussian sensing matrix. For $t > 0$,*

$$\Pr_{\mathbf{B}}\left[\frac{\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2}{\|\mathbf{x}^*\|_2} \geq (1+t)\sqrt{\frac{d-s}{d}}\right] \leq 2\exp(-c't^2(d-s)). \qquad (10)$$

The result in (10) confirms that the relative error decreases with increasing number of measurements s in the under-determined system. For example, let $s = 3d/4$, the probability that the relative error exceeds $1/2$ decays with a sub-Gaussian tail. The bound in (10) also sharp because it is the upper tail of the following two-sided inequality

$$\Pr_{\mathbf{B}}\left[(1-t)\sqrt{\frac{d-s}{d}} \leq \frac{\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2}{\|\mathbf{x}^*\|_2} \leq (1+t)\sqrt{\frac{d-s}{d}}\right] \geq 1 - 2\exp(-c't^2(d-s)).$$

Proof of Lemma 1. Since $\mathbf{B} \in \mathbb{R}^{n \times d}$ is assumed Gaussian, the scaled collapsed matrix $\widetilde{\mathbf{B}}/r$ (6) is also a Gaussian matrix. The singular vectors of a rectangular Gaussian matrix $\mathbf{M} \in \mathbb{R}^{p \times q}$, $p > q$, span a q-dimensional uniformly random subspace of $\mathbb{R}^p$, see Section 5.2.6 in [26]. Therefore, $\hat{\mathbf{x}} = (\mathbf{I} - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\intercal)\mathbf{x}^*$ is the projection of $\mathbf{x}^* \in \mathbb{R}^d$ onto a uniformly random $(d-s)$-dimensional subspace. The result in eq. (10) follows from invoking the Johonson-Lindenstrauss (JL) Lemma (65) with $p = d$ and $q = d - s$.

Theorem 1. *Let $\widehat{\mathbf{X}}$ be as defined in (7). Let $\mathbf{X}^* \in \mathbb{R}^{d \times m}$ be the fixed unknown matrix, and let $\mathbf{B}$ be a Gaussian sensing matrix. For $\log m \leq c't^2(d-s)/2$,*

$$\Pr_{\mathbf{B}}\left[\frac{\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F}{\|\mathbf{X}^*\|_F} \geq (1+t)\sqrt{\frac{d-s}{d}}\right] \leq 2\exp(-c't^2(d-s)/2). \qquad (11)$$

Proof of Theorem 1. Let event $\mathcal{E}$ be defined as the union of events $\mathcal{E}_i$

$$\mathcal{E} = \bigcup_{i=1}^{i=m} \mathcal{E}_i, \text{ where}$$

$$\mathcal{E}_i = \{\mathbf{x}^* \mid \|\mathbf{x}_i^* - \hat{\mathbf{x}}_i\|_2^2 \geq \frac{(1+t)^2(d-s)}{d}\|\mathbf{x}_i^*\|_2^2\}. \qquad (12)$$

Then

$$= \Pr[\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F^2 \leq \frac{(1+t)^2(d-s)}{d}\|\mathbf{X}^*\|_F^2] \qquad (13)$$

$$= \Pr[\sum_{i=1}^{i=m}\|\mathbf{x}_i^* - \hat{\mathbf{x}}_i\|_2^2 \leq \frac{(1+t)^2(d-s)}{d}\sum_{i=1}^{i=m}\|\mathbf{x}_i^*\|_2^2] \qquad (14)$$

$$\geq 1 - \sum_{i=1}^{i=m} \Pr[\mathcal{E}_i] \qquad (15)$$

$$\geq 1 - 2\exp(-c't^2(d-s)/2). \qquad (16)$$

(14) substitutes $\|\mathbf{X}^*\|_F^2 = \sum_{i=1}^{i=m}\|\mathbf{x}_i^*\|_2^2$. (15) follows from noting that $\Pr[\mathcal{E}^c] = 1 - \Pr[\mathcal{E}]$, and applying the union bound to $\mathcal{E}$, defined in (12). (16) is by substituting (10) and bounding $m\exp(-c't^2(d-s)/2) \leq 1$ for $\log m \leq c't^2(d-s)/2$.

Lemma 2. *Let $\hat{\mathbf{x}}$ be as defined in (7). Let $\mathbf{B}$ be a fixed sensing matrix, and let $\mathbf{x}^*$ be a zero-mean sub-Gaussian random vector (32). For $t \geq 0$,*

$$\Pr_{\mathbf{x}^*}\left[\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 \geq c + 2K^2(\sqrt{d-s}+1)t\right] \leq \exp(-t), \quad (17)$$

where $c = K^2(d - s + \frac{1}{2}\sqrt{d-s})$, and K is as given in (32).

Proof of Lemma 2. The result in (17) follows from applying the Hanson-Wright inequality (64) to (8).

To derive the result in Theorem 2, it is necessary to show that $\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2$ is a sub-exponential random variable. A random variable X is sub-exponential if for $t \geq 0, K_1 > 0$

$$X \sim \text{subExp} \iff \Pr[|X| \geq t] \leq c'\exp(-t/K_1), \quad (18)$$

where $c' \geq 1$. It cannot be concluded that the shifted random variable $\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c$ is sub-exponential because the lower tail probability $\Pr[\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c \leq -t], t > 0$ is not upper bounded. However, our goal is to upper bound the probability that the error *exceeds* a certain value. We therefore define the non-negative random variable $\tilde{\mathbf{e}} - c$, which upper bounds the lower tail as $\{\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c \leq -t, t > 0\} \leq 0$, (19). The definition of $\tilde{\mathbf{e}} - c$ is given in (19), (20). Fig. 2 is a visual description of the definition. We show that $\tilde{\mathbf{e}} - c$ is a sub-exponential random variable in Lemma 3.

Lemma 3. *The random variable $\tilde{\mathbf{e}} - c$, defined in (19), (20), is a sub-exponential random variable*

$$\tilde{\mathbf{e}} - c \mid \mathcal{E}_1 \triangleq 0, \qquad (19)$$

$$\tilde{\mathbf{e}} - c \mid \mathcal{E}_2 \triangleq \|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c, \qquad (20)$$

where events $\mathcal{E}_1, \mathcal{E}_2$ are defined as

$$\mathcal{E}_1 = \{\mathbf{x}^* \mid \|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c \leq 0\}, \qquad (21)$$

$$\mathcal{E}_2 = \mathcal{E}_1^c = \{\mathbf{x}^* \mid \|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c > 0\}. \qquad (22)$$
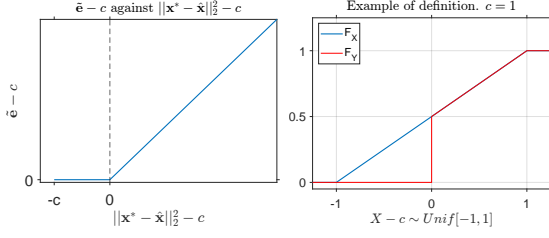
Fig. 2: **Left.** The derived random variable (rv) $\tilde{\mathbf{e}} - c$ is plotted against the error rv $\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c$. Note that i) $\tilde{\mathbf{e}} - c \geq 0$, ii) $\tilde{\mathbf{e}} - c \geq \|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c$, iii) $\tilde{\mathbf{e}} - c \mid \mathcal{E}_1 = 0$, and iv) $\tilde{\mathbf{e}} - c \mid \mathcal{E}_2 = \|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c$. **Right.** Example. The cdfs of $Y(\tilde{\mathbf{e}} - c)$ and $X(\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c) \sim \mathrm{Unif}[-1, 1]$ are plotted.

Moreover, for $t > 0$

$$\Pr\left[\tilde{\mathbf{e}} - c \geq t\right] = \Pr\left[\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c \geq t\right]. \quad (23)$$

Proof of Lemma 3. We first prove the claim in (23). For $t > 0$,

$\Pr[\tilde{\mathbf{e}} - c \geq t]$
$= \Pr[\tilde{\mathbf{e}} - c \geq t \mid \mathcal{E}_1] \cdot \Pr[\mathcal{E}_1] + \Pr[\tilde{\mathbf{e}} - c \geq t \mid \mathcal{E}_2] \cdot \Pr[\mathcal{E}_2]$
$= \Pr[\tilde{\mathbf{e}} - c \geq t \mid \mathcal{E}_2] \cdot \Pr[\mathcal{E}_2] \quad (24)$
$= \Pr[\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c \geq t \mid \mathcal{E}_2] \cdot \Pr[\mathcal{E}_2] \quad \text{(By eq. (20))}$
$= \Pr[\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c \geq t], \quad (25)$

where (24) follows from $\Pr[\tilde{\mathbf{e}} - c \geq t \mid \mathcal{E}_1] = 0 \, \forall t > 0$. (25) follows from

$\Pr[\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c \geq t \mid \mathcal{E}_2] \Pr[\mathcal{E}_2] = \Pr[\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c \geq t \cap \mathcal{E}_2],$

and that the event $\{\mathbf{x}^* \mid \|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c \geq t, t > 0\}$ is a subset of $\mathcal{E}_2$. The claim in (23) is shown in (25). From (17), for $t \geq 0$

$$\Pr\left[\frac{\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c}{2K^2(\sqrt{d-s}+1)} \geq t\right] \leq \exp(-t). \quad (26)$$

From (23), for $t > 0$

$$\Pr\left[\frac{\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c}{2K^2(\sqrt{d-s}+1)} \geq t\right] = \Pr\left[\frac{\tilde{\mathbf{e}} - c}{2K^2(\sqrt{d-s}+1)} \geq t\right], \quad (27)$$

where $\sqrt{d-s}+1 > 0$ is a positive number since $s < d$, see (8). From (19) and (20), $\tilde{\mathbf{e}} - c$ is non-negative so that $\tilde{\mathbf{e}} - c = |\tilde{\mathbf{e}} - c|$,

$$\Pr\left[\frac{\tilde{\mathbf{e}} - c}{2K^2(\sqrt{d-s}+1)} \geq t\right] = \Pr\left[\frac{|\tilde{\mathbf{e}} - c|}{2K^2(\sqrt{d-s}+1)} \geq t\right]. \quad (28)$$

Combining (27) and (28), for $t > 0$

$$\Pr\left[\frac{|\tilde{\mathbf{e}} - c|}{2K^2(\sqrt{d-s}+1)} \geq t\right] = \Pr\left[\frac{\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2 - c}{2K^2(\sqrt{d-s}+1)} \geq t\right]. \quad (29)$$

Substituting (26) in (29), for $t > 0$,

$$\Pr\left[\frac{|\tilde{\mathbf{e}} - c|}{2K^2(\sqrt{d-s}+1)} \geq t\right] \leq \exp(-t) \quad \forall t > 0. \quad (30)$$

Since $\exp(0) = 1$, (30) also holds at $t = 0$

$$\Pr\left[\frac{|\tilde{\mathbf{e}} - c|}{2K^2(\sqrt{d-s}+1)} \geq t\right] \leq \exp(-t) \quad \forall t \geq 0. \quad (31)$$

In (31), we have verified the definition in (18) for $\tilde{\mathbf{e}} - c$ with $K_1 = 2K^2(\sqrt{d-s}+1)$.

Theorem 2. *Let $\mathbf{X}^* \in \mathbb{R}^{d \times m}$ be the unknown matrix such that the columns $\mathbf{x}_i^*$ are identical, independent, zero-mean sub-Gaussian random vectors, i.e., for $\alpha \in \mathbb{R}^d$, $K \geq 0$*

$$\mathbb{E}[\exp(\alpha^\mathsf{T} \mathbf{x}_i^*)] \leq \exp(\|\alpha\|_2^2 K^2 / 2). \quad (32)$$

Let $\widehat{\mathbf{X}}$ be as defined in (7). Let $\mathbf{B}$ be a fixed sensing matrix. For $t \geq 0$,

$$\Pr\left[\sum_{i=1}^{i=m} \|\mathbf{x}_i^* - \hat{\mathbf{x}}_i\|_2 - mc_2 \geq t\right] \leq 2 \exp\left(\frac{-c't^2}{m(d-s)}\right), \quad (33)$$

where $c_2 = K^2(d - s + \frac{5}{2}\sqrt{d-s} + 2)^{\frac{1}{2}}$ and c' is an absolute constant.

Proof of Theorem 2. Since $\tilde{\mathbf{e}} - c$ is a sub-exponential random variable (Lemma 3), by (66), $\sqrt{\tilde{\mathbf{e}} - c}$ is a sub-Gaussian random variable. To apply Hoeffding's inequality to the sum $\sum_i \sqrt{\tilde{\mathbf{e}}_i}$, we first upper bound the expectation $\mathbb{E}[\sqrt{\tilde{\mathbf{e}}}]$ and the sub-Gaussian norm $\|\sqrt{\tilde{\mathbf{e}}}\|_{\varphi_2}$. For non-negative random variable X, $\mathbb{E}[X] = \int_0^\infty \Pr[X > t]dt$. Applied to $\tilde{e} - c/2(\sqrt{d-s}+1)$,

$$\mathbb{E}\left[\frac{\tilde{\mathbf{e}} - c}{2K^2(\sqrt{d-s}+1)}\right] \leq \int_0^\infty \Pr\left[\frac{\tilde{\mathbf{e}} - c}{2K^2(\sqrt{d-s}+1)} \geq t\right]dt \quad (34)$$

$$\leq \int_0^\infty \exp(-t)dt = 1. \quad (35)$$

In (35), we have substituted the tail probability upper bounds from (31). By (35), and substituting c from (17), the expectation is upper bounded as

$$\mathbb{E}[\tilde{\mathbf{e}}] \leq 2K^2(\sqrt{d-s}+1) + c = K^2(d - s + \frac{5}{2}\sqrt{d-s} + 2). \quad (36)$$

By Jensen's inequality $\mathbb{E}[\sqrt{\tilde{\mathbf{e}}}] \leq \sqrt{\mathbb{E}[\tilde{\mathbf{e}}]}$, and therefore

$$\mathbb{E}[\sqrt{\tilde{\mathbf{e}}}] \leq \left(K^2(d - s + \frac{5}{2}\sqrt{d-s} + 2)\right)^{1/2}. \quad (37)$$

For sub-Gaussian X, $\|X\|_{\varphi_2} \leq c'\mathbb{E}[X]$ (Proposition 2.7.1 (ii), (iv) [26]). By (66), the sub-Gaussian norm squared is upper bounded as

$$\|\tilde{\mathbf{e}}\|_{\varphi_2}^2 = \|\tilde{\mathbf{e}}\|_{\varphi_1} \leq c'(K^2(d - s + \frac{5}{2}\sqrt{d-s} + 2))^{1/2}$$
$$\leq c'K\sqrt{d-s}. \quad (38)$$

Given the upper bound on the sub-Gaussian norm in (38), we apply Hoeffding's inequality (67) to the sum $\sum_i \sqrt{\tilde{\mathbf{e}}_i}$

$$\Pr\left[\sum_{i=1}^{i=m}(\sqrt{\tilde{\mathbf{e}}_i} - \mathbb{E}[\sqrt{\tilde{\mathbf{e}}_i}]) \geq t\right] \leq 2\exp\left(\frac{-c't^2}{mK\sqrt{d-s}}\right), \quad (39)$$

where c' is an absolute constant. Let c_2 denote

$$c_2 = c'(K^2(d - s + \frac{5}{2}\sqrt{d-s} + 2))^{1/2}. \quad (40)$$

Substituting (37) into (39)

$$\Pr\left[\sum_{i=1}^{i=m} \sqrt{\tilde{\mathbf{e}}_i} - mc_2 \geq t\right] \leq 2\exp\left(\frac{-c't^2}{mK\sqrt{d-s}}\right). \quad (41)$$

By definition in (19), (20) (also see Fig. 2), $\|\mathbf{x}_i^* - \hat{\mathbf{x}}_i\|_2 \leq \sqrt{\tilde{\mathbf{e}}_i} \ \forall \ i \in [m]$. Therefore,

$$\Pr[\sum_{i=1}^{i=m} \|\mathbf{x}_i^* - \hat{\mathbf{x}}_i\|_2 - mc_2 \geq t] \leq \Pr[\sum_{i=1}^{i=m} \sqrt{\tilde{\mathbf{e}}_i} - mc_2 \geq t]$$
$$\leq 2\exp(\frac{-c't^2}{mK\sqrt{d-s}}). \quad (42)$$

In (42), we have substituted the bound in (41). We conclude with the final result

$$\Pr[\sum_{i=1}^{i=m} \|\mathbf{x}_i^* - \hat{\mathbf{x}}_i\|_2 - mc_2 \geq t] \leq 2\exp(\frac{-c't^2}{mK\sqrt{d-s}}). \quad (43)$$

Remark. Assuming Gaussian $\mathbf{B}$, Theorem 1 shows the probability that the relative error in (7) exceeds $\sqrt{(d-s)/d}$ decays at a sub-Gaussian rate. Assuming sub-Gaussian $\mathbf{x}^*$ instead, Theorem 2 shows the probability that the error exceeds mc_2, where c_2 as in (40), also decays at a sub-Gaussian rate. Both results agree that the estimate (7) improves as the number of measurements s in the under-determined system (6) increases. The quantity K in (33) depends linearly on the sub-Gaussian norm of $\mathbf{x}^*$ and can be estimated in practice. For $\mathbf{x}^* \sim \mathcal{N}(0, \mathbf{I})$, $K = 1$. For $\mathbf{x}^*$ uniformly distributed on the Euclidean ball, $K = c'$, where c' is an absolute constant.

B. k-sparse model

For the standard (un-permuted) linear inverse problem $\mathbf{y} = \mathbf{B}\mathbf{x}^* + \mathbf{w}$, the error in the solution $\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2$ is upper bounded in terms of the noise $\|\mathbf{w}\|_2$ and $\sigma_{\min}(\mathbf{B})$, the smallest singular value of $\mathbf{B}$. In this section, we bound the error in the proposed initialization for the k-sparse permutation model in terms of the value of the objective function F, $\sigma_{\min}(\mathbf{B})$, and the number of shuffles k.

Lemma 4. *Let* $\mathbf{P}_k^* \in \mathbb{R}^{n \times n}$ *be the fixed unknown permutation matrix. Let* $\mathbf{x}^*$ *be the fixed unknown vector. Assuming Gaussian* $\mathbf{B}$, *let* $\mathbf{y}^* = \mathbf{B}\mathbf{x}^*$, $\hat{\mathbf{y}}^{(0)} = \mathbf{P}_k^*\mathbf{B}\mathbf{x}^*$. *For* $k = 1, \cdots, n-1$, *and* $t \geq 0$,

$$\Pr\left[\|\mathbf{y}^* - \hat{\mathbf{y}}^{(0)}\|_2^2 \geq 2\|\mathbf{y}^*\|_2^2 - 2\|\mathbf{x}^*\|_2^2(n - k - c_1\sqrt{t} + 6t)\right]$$
$$\leq 7\exp(-t), \quad (44)$$

where $c_1 = 2\sqrt{n-k} + 4\sqrt{3k}$.

Proof of Lemma 4. Under the k-sparse assumption on $\mathbf{P}^*$, the known vector $\mathbf{y}$ has k shuffled entries. For $\hat{\mathbf{P}}^{(0)} = \mathbf{I}$, *i.e.*, $\hat{\mathbf{y}}^{(0)} = \mathbf{y}$, the forward error is

$$\|\mathbf{y}^* - \hat{\mathbf{y}}^{(0)}\|_2^2 = 2\|\mathbf{y}^*\|_2^2 - 2\langle \mathbf{y}^*, \hat{\mathbf{y}}^{(0)} \rangle, \quad (45)$$

where $\mathbf{y}^* = \mathbf{B}\mathbf{x}^*$ is the unknown vector of unshuffled measurements. Since $\|\mathbf{y}^*\|_2 = \|\hat{\mathbf{y}}^{(0)}\|_2$, and initialization $\hat{\mathbf{y}}^{(0)} = \mathbf{P}_k^*\mathbf{B}\mathbf{x}^*$ is known, the only unknown term in (45)

is the inner product $\langle \mathbf{y}^*, \hat{\mathbf{y}}^{(0)} \rangle$. The scaled inner product $\langle \mathbf{y}^*, \hat{\mathbf{y}}^{(0)} \rangle / \|\mathbf{x}^*\|_2^2$ is expanded as

$$\frac{\langle \mathbf{y}^*, \hat{\mathbf{y}}^{(0)} \rangle}{\|\mathbf{x}^*\|_2^2} = \frac{1}{\|\mathbf{x}^*\|_2^2} \sum_{i=1}^{i=n} \mathbf{b}_i^\mathsf{T}\mathbf{x}^*\mathbf{b}_{\mathbf{P}^*(i)}^\mathsf{T}\mathbf{x}^* \quad (46)$$

$$= \underbrace{\frac{1}{\|\mathbf{x}^*\|_2^2} \sum_{i=1}^{i=n-k} (\mathbf{b}_i^\mathsf{T}\mathbf{x}^*)^2}_{\triangleq T_1} + \underbrace{\frac{1}{\|\mathbf{x}^*\|_2^2} \sum_{i=n-k+1}^{n} \mathbf{b}_i^\mathsf{T}\mathbf{x}^*\mathbf{b}_{\mathbf{P}^*(i)}^\mathsf{T}\mathbf{x}^*}_{\triangleq T_2}.$$

We have assumed in (46), *wlog*, that the last k rows of $\mathbf{P}_k^*$ are shuffled. Assuming Gaussian $\mathbf{B}$, $(\mathbf{b}_i^\mathsf{T}\mathbf{x}^*/\|\mathbf{x}^*\|_2)^2 \sim \chi^2$ is Chi-squared distributed with 1 degree of freedom. T_1, defined in (46), is the sum of $n-k$ independent Chi-square random variables, and is bounded, using (70), as

$$\Pr\left[T_1 \leq n - k - 2\sqrt{(n-k)t}\right] \leq \exp(-t). \quad (47)$$

The product random variables in T_2 are distributed as the difference of two independent χ^2 random variables Z_i^1, Z_i^2

$$\frac{\mathbf{b}_i^\mathsf{T}\mathbf{x}^*\mathbf{b}_{\mathbf{P}(i)}^\mathsf{T}\mathbf{x}^*}{\|\mathbf{x}^*\|_2^2} \sim \frac{1}{2}Z_i^1 - \frac{1}{2}Z_i^2 \ \forall \ i \in n-k+1, \cdots, n. \quad (48)$$

The random variables (rv) in (48) are not mutually independent, but each rv depends on, at most, two rvs. To see this, let permutation $\mathbf{P}$ such that $\mathbf{P}(i) \mapsto j$, then $\mathbf{b}_i^\mathsf{T}\mathbf{x}^*\mathbf{b}_j^\mathsf{T}\mathbf{x}^*$ is not independent of

$$\mathbf{b}_j^\mathsf{T}\mathbf{x}^*\mathbf{b}_{\mathbf{P}(j)}^\mathsf{T}\mathbf{x}^*, \ \mathbf{b}_{\mathbf{P}^\mathsf{T}(i)}^\mathsf{T}\mathbf{x}^*\mathbf{b}_i^\mathsf{T}\mathbf{x}^*. \quad (49)$$

The k rvs in (48) can therefore be partitioned into three sets P, Q, R such that the rvs within each set are independent. Let k_1 be the number of rvs in set P. The sum T_P, defined as

$$T_P \triangleq \sum_{i \in P} \mathbf{b}_i^\mathsf{T}\mathbf{x}^*\mathbf{b}_{\mathbf{P}(i)}^\mathsf{T}\mathbf{x}^* = \sum_{i \in P} Z_i^1 - \sum_{i \in P} Z_i^2, \quad (50)$$

is upper bounded in probability as

$$\Pr[T_P \geq 4\sqrt{k_1 t} + 2t] \leq 2\exp(-t). \quad (51)$$

(51) follows from applying the union bound to probabilities p_1, p_2

$$p_1 = \Pr\left[\sum_{i \in P} Z_i^1 \leq k_1 - 2\sqrt{k_1 t}\right] \leq \exp(-t), \quad (52)$$

$$p_2 = \Pr\left[\sum_{i \in P} Z_i^2 \geq k_1 + 2\sqrt{k_1 t} + 2t\right] \leq \exp(-t), \quad (53)$$

and bounding p_1, p_2 using tail inequalities (70), (69), respectively. Defining T_Q, T_R similarly to (50), with cardinalities k_2, k_3 respectively, and applying the union bound as in (52), (53) gives

$$\Pr\left[T_2 \geq 4(\sqrt{k_1 t} + \sqrt{k_2 t} + \sqrt{k_3 t}) + 6t\right] \leq 6\exp(-t), \quad (54)$$

where $T_2 = T_P + T_Q + T_R$. Since $k_1 + k_2 + k_3 = k$, and $\sqrt{k_1} + \sqrt{k_2} + \sqrt{k_3} \leq \sqrt{3k}$,

$$\Pr\left[T_2 \geq 4\sqrt{3kt} + 6t\right] \leq 6\exp(-t). \quad (55)$$

Applying the union bound to (47), (55) gives the result in (44).

Lemma 5. *For the same assumptions as in Lemma 4, $k = 1, \cdots, n-1$, and $t \geq 0$,*

$$\Pr\left[\sigma_{\min}^2(\mathbf{B})\|\mathbf{x}^* - \hat{\mathbf{x}}^{(1)}\|_2^2 \geq \right. \tag{56}$$
$$\left. 2\|\mathbf{y}^*\|_2^2 - 2\|\mathbf{x}^*\|_2^2 \left(n - k - c_1\sqrt{t} - 6t\right) - F^{(1)}\right] \leq 7e^{-t},$$

where $\hat{\mathbf{x}}^{(1)} = \mathbf{B}^{\dagger}\hat{\mathbf{y}}^{(0)}$, $F^{(1)} = \|\mathbf{y} - \mathbf{B}\hat{\mathbf{x}}^{(1)}\|_2^2$, and $c_1 = 2\sqrt{n-k} + 4\sqrt{3k}$.

Proof of Lemma 5. Let $\mathbf{e}$ denote the error term such that

$$\hat{\mathbf{y}}^{(0)} = \mathbf{B}\hat{\mathbf{x}}^{(1)} + \mathbf{e}, \tag{57}$$

where $\hat{\mathbf{x}}^{(1)} = \mathbf{B}^{\dagger}\hat{\mathbf{y}}^{(0)}$, $\hat{\mathbf{y}}^{(0)} = \mathbf{P}_k^*\mathbf{B}\mathbf{x}^*$, and $\mathbf{e}$ is orthogonal to the range space of $\mathbf{B}$, i.e., $\mathbf{e} \perp \mathcal{R}(\mathbf{B})$. Substituting $\mathbf{y}^* = \mathbf{B}\mathbf{x}^*$ and (57) below.

$$\|\mathbf{y}^* - \hat{\mathbf{y}}^{(0)}\|_2^2 \triangleq \|\mathbf{B}(\mathbf{x}^* - \hat{\mathbf{x}}^{(1)}) - \mathbf{e}\|_2^2. \tag{58}$$

Using Pythogras' theorem,

$$\|\mathbf{B}(\mathbf{x}^* - \hat{\mathbf{x}}^{(1)}) - \mathbf{e}\|_2^2 = \|\mathbf{B}(\mathbf{x}^* - \hat{\mathbf{x}}^{(1)})\|_2^2 + \|\mathbf{e}\|_2^2. \tag{59}$$

Substituting the lower bound $\sigma_{\min}^2(\mathbf{B})\|\mathbf{x}^* - \hat{\mathbf{x}}^{(1)}\|_2^2 \leq \|\mathbf{B}(\mathbf{x}^* - \hat{\mathbf{x}}^{(1)})\|_2^2$ in (59),

$$\sigma_{\min}^2(\mathbf{B})\|\mathbf{x}^* - \hat{\mathbf{x}}^{(1)}\|_2^2 \leq \|\mathbf{y}^* - \hat{\mathbf{y}}^{(0)}\|_2^2 - \|\mathbf{e}\|_2^2. \tag{60}$$

(61) follows by substituting the probability upper bound for $\|\mathbf{y}^* - \hat{\mathbf{y}}^{(0)}\|_2^2$ from (44) into (60), and noting that $\|\mathbf{e}\|_2^2 = F^{(1)} = F(\hat{\mathbf{P}}^{(0)}, \hat{\mathbf{x}}^{(1)})$.

Theorem 3. *Let $\mathbf{P}_k^* \in \mathbb{R}^{n \times n}$ be the fixed unknown permutation matrix. Assuming Gaussian $\mathbf{B}$, let $\mathbf{X}^*$ be the fixed unknown matrix. Let $\mathbf{Y}^* = \mathbf{B}\mathbf{X}^*$, and $\hat{\mathbf{Y}}^{(0)} = \mathbf{P}_k^*\mathbf{B}\mathbf{X}^*$. For $k = 1, \cdots, n-1$, and $t \geq \log m^2$,*

$$\Pr\left[\sigma_{\min}^2(\mathbf{B})\|\mathbf{X}^* - \hat{\mathbf{X}}^{(1)}\|_F^2 \geq \right. \tag{61}$$
$$\left. 2\|\mathbf{Y}^*\|_F^2 - 2\|\mathbf{X}^*\|_F^2\left(n - k - c_1\sqrt{t} - 6t\right) - F^{(1)}\right] \leq 7e^{-t},$$

where $\hat{\mathbf{X}}^{(1)} = \mathbf{B}^{\dagger}\hat{\mathbf{Y}}^{(0)}$, $F^{(1)} = \|\mathbf{Y} - \mathbf{B}\hat{\mathbf{X}}^{(1)}\|_F^2$, and $c_1 = 2\sqrt{n-k} + 4\sqrt{3k}$.

An immediate upper bound that does not depend on the number of diagonal entries of $\mathbf{P}_k^*$ is

$$\sigma_{\min}^2(\mathbf{B})\|\mathbf{X}^* - \hat{\mathbf{X}}^{(1)}\|_F^2 \leq 4\|\mathbf{Y}^*\|_F^2 - F^{(1)}. \tag{62}$$

(61) shows that the probability that the error exceeds $\|\mathbf{Y}^*\|_F^2 - c$, $c = \Omega(n-k)$, and $n-k$ is the number of diagonal entries in $\mathbf{P}_k^*$, decays exponentially. Furthermore, the term $\|\mathbf{Y}^*\|_F^2$ is known, and $\|\mathbf{X}^*\|_F$ can be lower bounded, in terms of known quantities, as $\|\mathbf{X}^*\|_F^2 \leq \|\mathbf{Y}^*\|_F^2 / \sigma_{\max}^2(\mathbf{B})$.

Proof of Theorem 3. The proof follows by applying the union bound to (56).

IV. RESULTS

MATLAB code and datasets are at the author's github [1].
Benchmarks. We compare against six benchmark methods. We refer to these methods as 'ℓ_2-regularized' [5], 'ds+' [5], 'Spectral' [17], 'Biconvex' [4], 'RLUS' [10], and 'Stochastic alt-min' [15]. The 'ℓ_2-regularized' method considers the k-sparse permutation model. The method incorporates the model

---

[1] https://github.com/aabbas02/ksparse-and-rlocal

assumption by imposing a row-wise group sparse penalty on $\|\hat{\mathbf{Y}}_{i,:}\|_2$, where $\hat{\mathbf{Y}} = \mathbf{B}\hat{\mathbf{X}}$. The other benchmark methods are discussed in the following paragraphs. For all experiments, we set $\epsilon = 0.01$ in Algorithm 1.

Figs. 3a, 3b. We compare our algorithm for the r-local and k-sparse permutation models to benchmarks. To adapt the 'Spectral', 'ℓ_2-regularized', and 'Biconvex' methods to the r-local model, we add a constraint enforcing the permutation estimates to be r-local. The results show that our algorithm recovers $\mathbf{P}^*$ with decreasing block size r and number of shuffles k. This observation confirms the conclusions of Theorems 1,2,3 as the initialization to our algorithm improves with lower values of r and k. The results also show that our algorithm is the only algorithm applicable to both models. For $r = 125$ in Fig. 3a and $k = 850$ in Fig. 3b, the MATLAB run-times of our algorithm, with 16 Gb RAM, are less than one second.

Fig. 3c. The entries of the design matrix $\mathbf{B}$ are sampled i.i.d. from the uniform $[0,1]$ distribution. Compared to the case for Gaussian $\mathbf{B}$ (Fig. 3b), the performance of the 'Spectral' and 'RLUS' methods deteriorates significantly. This is because both algorithms consider quadratic measurements $\mathbf{Y}\mathbf{Y}^{\intercal}$. Specifically, the 'Spectral' method is based on the spectral initialization technique [27], which assumes that the design matrix is Gaussian. In contrast, the performance of the proposed algorithm and the 'ℓ_2-regularized' method does not deteriorate.

Fig. 3d. The 'ds+' algorithm [5] considers the convex relaxation of (2) by minimizing the objective over the set of doubly stochastic matrices. Assuming an upper bound on the number of shuffles k is known, 'ds+' also constrains $\langle \mathbf{I}, \mathbf{P} \rangle \geq n - k$. Each iteration of 'ds+' minimizes a linear program, greatly increasing the run-time of the algorithm. The results show that the proposed algorithm outperforms 'ds+', possibly because, for our algorithm, the objective is optimized directly over the set of permutations.

Fig. 3e. We compare to the method in [15] which considers the $m = 1$ single-view setup and proposes stochastic alternating minimization (S.alt-min) to optimize (2). The run-time for S.alt-min is 50 times the run-time of the proposed algorithm because S.alt-min updates $\mathbf{P}$ multiple times in each iteration and retains the best update. [15] also proposes alt-min with multiple initializations for $\hat{\mathbf{P}}^{(0)}$. The results in Fig. 3e show that our algorithm (alt-min with $\hat{\mathbf{P}}^{(0)} = \mathbf{I}$ initialization) outperforms both S.alt-min and alt-min with multiple initializations.

Real data. For the linked linear regression problem introduced in Section I-B, we report results in Table I on the four datasets considered in [5]. The design matrix and the regression model are also defined as in [5]. For 'ftp', 'scm', and 'scs', mismatch error is casued by an erroneous identifier. The values of a feature (or column of the design matrix) are rounded and data points with the same feature value are assigned to the same block and permuted, also as done in [5]. The 'air-qty' dataset comprises time-stamped readings with year, month, day, and hour information. In row 4 (5) of Table I, readings with the same month and day (day and hour) are assigned to the same block and permuted. The results show that the proposed alt-min algorithm outperforms the competing
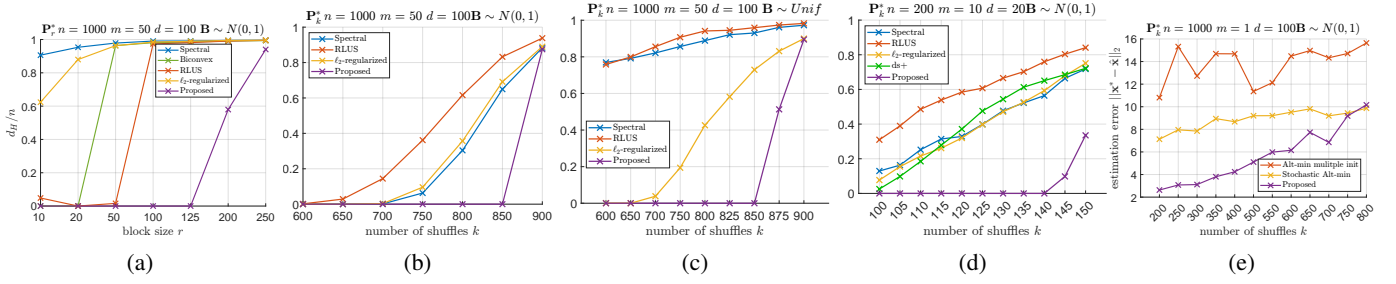
Fig. 3: **Synthetic simulations.** $\mathbf{Y} = \mathbf{P}^*\mathbf{B}_{n \times d}\mathbf{X}^*_{d \times m} + \mathbf{W}$. The entries of $\mathbf{X}^*$ are drawn from the normal distribution. The permutation matrix $\mathbf{P}^*_r(\mathbf{P}^*_k)$ is sampled uniformly from the set of permutations $\Pi_r(\Pi_n)$. The results are averaged over 15 Monte-Carlo runs. (a). Figure plots the fractional Hamming distortion d_H/n against block size r. (b), (c). Figure plots d_H/n against number of shuffles k. (d). Figure plots the estimation error against k.

Dataset	n	d	m	$r_{\max}$	s	Oracle	Naive	RLUS	ℓ_2-reg	Alt-min-k	Alt-min-r
ftp	335	30	6	43	46	(0.88, 0)	(0.44, 0.70)	(0.62, 0.01)	(0.57, 0.59)	(0.73, 0.41)	(**0.85, 0.17**)
scm	8966	35	16	424	1137	(0.58, 0)	(0.21, 0.79)	(0.52, 0.29)	(0.46, 0.49)	(0.54, 0.22)	(**0.55, 0.22**)
scs	44484	10	6	3553	356	(0.81, 0)	(0.01, 0.99)	(0.74, 0.24)	(0.02, 0.98)	(0.73, 0.33)	(**0.77, 0.21**)
air-qty	27605	27	16	95	366	(0.69, 0)	(0.29, 0.78)	(0.65, 0.20)	(0.64, 0.21)	(**0.66, 0.17**)	(0.65, 0.19)
air-qty	27605	27	16	46	744	(0.69, 0)	(0.10, 0.94)	(0.60, 0.34)	(0.26, 0.79)	(0.57, 0.42)	(**0.62, 0.32**)

TABLE I: $r_{\max}$ denotes the largest block size of $\mathbf{P}^*_r$ and s denotes the number of blocks. The methods are compared based on the coefficient of determination R_2 and relative error (R_2, relative error). The relative error is $\|\mathbf{X}^* - \widehat{\mathbf{X}}\|_F/\|\mathbf{X}^*\|_F$, where $\mathbf{X}^* = \mathbf{B}^\dagger\mathbf{Y}^*$ is the 'oracle' regression matrix given unpermuted data $\mathbf{Y}^*$. The 'naive' estimate from permuted data $\mathbf{Y}$ is $\widehat{\mathbf{X}} = \mathbf{B}^\dagger\mathbf{Y}$. The coefficient $R_2(\widehat{\mathbf{X}}) = 1 - (\|\mathbf{Y}^* - \mathbf{B}\widehat{\mathbf{X}}\|_F/\|\mathbf{Y}^*\|_F)$ measures the goodness of fit for the unpermuted data.

baselines. 'alt-min-k', i.e. alt-min initialized to $\widehat{\mathbf{P}}^{(0)} = \mathbf{I}$, is also competitive, possibly because permuting correlated rows does not greatly corrupt the design matrix. Detailed description of the datasets and the setup is provided in the appendix.

V. CONCLUSION

We propose and analyze the alternating minimization algorithm with a model-dependent initialization for the unlabeled sensing problem. Assuming a random measurement matrix or random signal, we upper bound the initialization error for k-sparse and r-local permutation models. Numerical experiments show that, compared to base-line methods, our algorithm achieves superior performance on both real and synthetic datasets. As part of future work, we plan to i) characterize the rate of convergence, and ii) analyze the second-order convergence properties of the proposed algorithm.

VI. APPENDIX

Hanson Wright Inequality. *From Theorem 2.1 in [28], let* $\Sigma = \mathbf{A}^\intercal\mathbf{A}$ *be a positive semi-definite matrix. Let* $\mathbf{x} = (x_1, \cdots, x_d)$ *be a zero-mean sub-Gaussian random vector. i.e., for* $\alpha \in \mathbb{R}^d$, $K \geq 0$

$$\mathbb{E}[\exp(\alpha^\intercal\mathbf{x}^*)] \leq \exp(\|\alpha\|_2^2 K^2/2). \tag{63}$$

For $t \geq 0$,

$$\Pr[\|\mathbf{A}\mathbf{x}\|_2^2 \geq K^2(tr(\Sigma) + 2\sqrt{tr(\Sigma^2)t} + 2t\|\Sigma\|)] \leq e^{-t}. \tag{64}$$

To derive the result in (17), set $\mathbf{A} = (\mathbf{I} - \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\intercal)$, as in (8). Since $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a $(d-s)$-dimensional projection matrix, the matrix $\Sigma = \mathbf{A}^\intercal\mathbf{A} = \mathbf{A}\mathbf{A} = \mathbf{A}$ is also a projection matrix. Therefore, $tr(\Sigma) = tr(\Sigma^2) = tr(\mathbf{A}) = d - s$, and $\|\Sigma\| = 1$. Next, upper bound $2\sqrt{tr(\Sigma^2)t} \leq 2t\sqrt{tr(\Sigma^2)} + \frac{1}{2}\sqrt{tr(\Sigma^2)}$.

Johnson Lindenstrauss Lemma. *Let* P *be a projection in* $\mathbb{R}^p$ *onto a uniformly distributed random q-dimensional subspace. Let* $z \in \mathbb{R}^p$ *be a fixed point and* $t > 0$. *Then, with probability at least* $1 - 2\exp(-ct^2q)$,

$$(1-t)\sqrt{\frac{q}{p}}\|z\|_2 \leq \|Pz\|_2 \leq (1+t)\sqrt{\frac{q}{p}}\|z\|_2. \tag{65}$$

Lemma 2.7.6 in [26]. *A random variable* X *is sub-Gaussian if and only if* X^2 *is sub-exponential. Moreover, the sub-Gaussian norm* $\|X\|_{\varphi_2}$ *and the sub-exponential norm* $\|X\|_{\varphi_1}$ *of* X *are such that*

$$\|X^2\|_{\varphi_1} = \|X\|_{\varphi_2}^2. \tag{66}$$

Hoeffdings' inequality. Theorem 2.6.2 in [26]. *Let* $X_1, \cdots, X_m$ *be independent, sub-Gaussian random variables. Then, for every* $t \geq 0$,

$$\Pr[|\sum_{i=1}^{i=n} X_i - \mathbb{E}[X_i]| \geq t] \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{i=m}\|X_i\|_{\varphi_2}^2}\right), \tag{67}$$

where $\|X_i\|_{\varphi_2}$ *denotes the sub-Gaussian norm of* X_i, *and* c *is an absolute constant.*

Sub-exponential norm definition. *The sub-exponential norm a random variable* X *is defined as*

$$\|X\|_{\varphi_1} \triangleq \inf\{t > 0 \mid \mathbb{E}[\exp(|X|/t)] \leq 2\}. \tag{68}$$

Tail inequality for χ_D^2 **distributed random variables [29].** *Let* Z_D *be a* χ^2 *statistic with* D *degrees of freedom. For any positive* t,

$$\Pr[Z_D \geq D + 2\sqrt{Dt} + 2t] \leq \exp(-t), \tag{69}$$
$$\Pr[Z_D \leq D - 2\sqrt{Dt}] \leq \exp(-t). \tag{70}$$

REFERENCES

[1] Jayakrishnan Unnikrishnan, Saeid Haghighatshoar, and Martin Vetterli. Unlabeled sensing with random linear measurements. IEEE Trans. Inf. Theory, 64(5):3237–3253, 2018.

[2] I. Dokmanić. Permutations unlabeled beyond sampling unknown. IEEE Signal Processing Letters, 26(6):823–827, 2019.

[3] 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, 64(5):3286–3300, 2018.

[4] Hang Zhang, Martin Slawski, and Ping Li. Permutation recovery from multiple measurement vectors in unlabeled sensing. In 2019 IEEE Int. Symposium on Inf. Theory (ISIT), pages 1857–1861, 2019.

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

[6] Hang Zhang and Ping Li. Optimal estimator for unlabeled linear regression. In Proceedings of the 37th Int. Conference on Machine Learning (ICML), 2020.

[7] Martin Slawski, Mostafa Rahmani, and Ping Li. A sparse representation-based approach to linear regression with partially shuffled labels. In Uncertainty in Artificial Intelligence, pages 38–48. PMLR, 2020.

[8] Xu Shi, Xiaoou Li, and Tianxi Cai. Spherical regression under mismatch corruption with application to automated knowledge translation. Journal of the American Statistical Association, (just-accepted):1–28, 2020.

[9] Martin Slawski and Emanuel Ben-David. Linear regression with sparsely permuted data. Electronic Journal of Statistics, 13(1):1–36, 2019.

[10] Ahmed Ali Abbasi, Abiy Tasissa, and Shuchin Aeron. R-local unlabeled sensing: A novel graph matching approach for multiview unlabeled sensing under local permutations. IEEE Open Journal of Signal Process., 2:309–317, 2021.

[11] Ahmed Ali Abbasi, Abiy Tasissa, and Shuchin Aeron. r-local unlabeled sensing: Improved algorithm and applications. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5593–5597, 2022.

[12] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval. Compressed sensing with unknown sensor permutation. In 2014 IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1040–1044, 2014.

[13] Liangzu Peng and Manolis C. Tsakiris. Linear regression without correspondences via concave minimization. IEEE Signal Proces. Letters, 27:1580–1584, 2020.

[14] Liangzu Peng, Xuming Song, Manolis C Tsakiris, Hayoung Choi, Laurent Kneip, and Yuamming Shi. Algebraically-initialized expectation maximization for header-free communication. In IEEE Int. Conf. on Acous., Speech and Signal Process. (ICASSP), pages 5182–5186. IEEE, 2019.

[15] Abubakar Abid and James Zou. A stochastic expectation-maximization approach to shuffled linear regression. In 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 470–477. IEEE, 2018.

[16] A. Pananjady, M. J. Wainwright, and T. A. Courtade. Denoising linear models with permuted data. In 2017 IEEE Int. Symposium on Inf. Theory (ISIT), pages 446–450, 2017.

[17] Hang Zhang and Ping Li. Optimal estimator for unlabeled linear regression. In Int. Conference on Machine Learning, pages 11153–11162. PMLR, 2020.

[18] David P Woodruff et al. Sketching as a tool for numerical linear algebra. Foundations and Trends® in Theoretical Computer Science, 10(1–2):1–157, 2014.

[19] Jared S Murray. Probabilistic record linkage and deduplication after indexing, blocking, and filtering. Journal of Privacy and Confidentiality 7 (1)., 2016.

[20] Partha Lahiri and Michael D Larsen. Regression analysis with linked data. Journal of the American statistical association, 100(469):222–230, 2005.

[21] Ying Han and Partha Lahiri. Statistical analysis with linked data. International Statistical Review, 87:S139–S157, 2019.

[22] Manuel Marques, Marko Stošić, and João Costeira. Subspace matching: Unique solution to point matching with geometric constraints. In 2009 IEEE 12th Int. Conference on Computer Vision, pages 1288–1294, 2009.

[23] Guanyu Wang, Jiang Zhu, Rick S. Blum, Peter Willett, Stefano Marano, Vincenzo Matta, and Paolo Braca. Signal amplitude estimation and detection from unlabeled binary quantized samples. IEEE Transactions on Signal Processing, 66(16):4291–4303, 2018.

[24] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks, 2018.

[25] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. Electronic Communications in Probability, 18(none):1 – 9, 2013.

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

[27] Wangyu Luo, Wael Alghamdi, and Yue M. Lu. Optimal spectral initialization for signal recovery with applications to phase retrieval. IEEE Transactions on Signal Processing, 67(9):2347–2356, 2019.

[28] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. Electronic Communications in Probability, 17:1–6, 2012.

[29] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. The Annals of Statistics, 28(5):1302 – 1338, 2000.