Alternating Minimization algorithm for unlabeled sensing and linked linear regression

Ahmed Ali Abbasi\textsuperscript{1}, Shuchin Aeron\textsuperscript{2}, and Abiy Tasissa \textsuperscript{*3}

\textsuperscript{1}Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011, USA.
\textsuperscript{2}Department of Electrical and Computer Engineering, Tufts University, Medford, MA 02155, USA.
\textsuperscript{3}Department of Mathematics, Tufts University, Medford, MA 02155, USA.

December 13, 2024

Abstract

Unlabeled sensing is a linear inverse problem with permuted measurements. We propose an alternating minimization (AltMin) algorithm with a suitable initialization for two widely considered permutation models: partially shuffled/\(k\)-sparse permutations and \(r\)-local/block diagonal permutations. Key to the performance of the AltMin algorithm is the initialization. For the exact unlabeled sensing problem, assuming either a Gaussian measurement matrix or a sub-Gaussian signal, we bound the initialization error in terms of the number of blocks \(s\) and the number of shuffles \(k\). Experimental results show that our algorithm is fast, applicable to both permutation models, and robust to choice of measurement matrix. We also test our algorithm on several real datasets for the ‘linked linear regression’ problem and show superior performance compared to baseline methods.

1 Introduction

The linear inverse problem is given by \(Y = BX^\ast + W\), where \(B \in \mathbb{R}^{n \times d}\) is the measurement matrix, \(Y \in \mathbb{R}^{n \times m}\) represents the linear measurements of the unknown \(X^\ast \in \mathbb{R}^{d \times m}\), and \(W_{ij} \sim \mathcal{N}(0, \sigma^2)\) denotes i.i.d. Gaussian noise. In unlabeled sensing, the measurements \(Y\) are scrambled. Specifically,

\[
Y = P^*BX^\ast + W, \tag{1}
\]

where \(P^* \in \mathbb{R}^{n \times n}\) is the unknown permutation matrix. Given \(Y\) and \(B\), the objective is to estimate \(X^\ast\). In this manuscript, \(X^\ast\) and \(P^*\) denote the underlying unknown parameters of the unlabeled sensing problem. If there is no noise, i.e., \(W = 0\), we refer to (1) as the exact unlabeled sensing problem.

Since estimating \(X^\ast\) and \(P^*\) for a generic permutation is challenging, several works [1, 2, 3, 4, 5, 6] assume a \(k\)-sparse or partially shuffled permutation model, Figure 1. A permutation matrix \(P^*_k\) is \(k\)-sparse if it has \(k\) off-diagonal elements, i.e., \(\langle I, P^*_k \rangle = n - k\), where \(\langle \cdot, \cdot \rangle\) denotes the trace inner product. The \(r\)-local, or block diagonal, permutation model was first proposed in [7] and later considered in [8, 9]. A permutation matrix \(P^*_r\) is \(r\)-local with \(s\) blocks if \(P^*_r = \text{blkdiag}(P_1, \ldots, P_s)\), where \(\text{blkdiag}(\cdot)\) denotes a block diagonal matrix. Fig. 1 illustrates the \(k\)-sparse and \(r\)-local permutation models. These two models are compared by information-theoretic inachievability in [9].

In [9], an alternating minimization algorithm was proposed for the unlabeled sensing problem. The algorithm estimates \(P^*\) and \(X^\ast\) by minimizing the following optimization program:

\[
(\hat{P}, \hat{X}) = \arg\min_{P \in \Pi_n, X} \quad F(P, X) = \|Y - PBX\|_F^2, \tag{2}
\]

where \(\Pi_n\) denotes the set of \(n \times n\) permutations and \(\|\cdot\|_F^2\) denotes the squared Frobenius norm of a matrix, which is the sum of the squares of its entries. Given that (2) is a non-convex optimization problem, a crucial part of the alternating minimization algorithm in [9] is the initialization. When the unknown permutation \(P^*_r\) is \(r\)-local, [9] proposed the collapsed initialization (3). Let \(P_i\) denote each \(r_i \times r_i\) block of \(P^*_r = \text{blkdiag}(P_1, \ldots, P_s)\). Let \(B_i \in \mathbb{R}^{r_i \times d}\) denote blocks of the measurement matrix

\*Corresponding Author: abiy.tasissa@Tufts.edu
Figure 1: Left. Sparse (or partially shuffled) permutation considered in [1, 2, 3, 4], with number of shuffles $k = 10$. Right. The $r$-local permutation structure considered in [7, 9, 10], with block size $r = 10$. In this paper, we propose a general algorithm for both permutation models.

$B = [B_1; \cdots; B_s]$, where ; denotes vertical concatenation. Then, each block of measurements $Y_i$ is expressed as $Y_i = PB_iX^*, i \in [s]$. The shuffled measurements in each block $[P_iB_i : Y_i]$ are summed to extract $s$ unshuffled measurements $(b^T_i : y^T_i)$, where $1_r \in \mathbb{R}^r$ denotes the vector whose entries are all 1. These $s$ measurements are then represented compactly in the collapsed linear system of equations $BX^* = \hat{Y}$, where $B \in \mathbb{R}^{s \times d}$ and $Y \in \mathbb{R}^{s \times m}$. The initialization for $r$-local $P^*_r$ is the minimum-norm solution to the collapsed system and is given by:

$$\hat{X}^{(0)}_r = B^\dagger \hat{Y},$$

where $\hat{X}^{(0)}_r = [\hat{x}^{(0)}_1 | \cdots | \hat{x}^{(0)}_m]$. For the $k$-sparse problem, in this manuscript, we propose the following initialization:

$$\tilde{X}^{(0)}_k = B^\dagger Y.$$  

Note that the above initialization corresponds to the least square solution of (1) when $P^*$ is the identity matrix. The initialization is motivated by the observation that, since $P^*_k$ has $k$ off-diagonal elements, the identity matrix serves as a simple approximation of $P^*_k$. The main goal of this manuscript is to characterize the effectiveness of the initializations $X^{(0)}_r$ and $X^{(0)}_k$ by upper bounding the initialization error for the exact unlabeled sensing problem.

### 1.1 Contributions and outline

This paper presents a theoretical analysis of the initialization for the exact unlabeled sensing problem under two permutation models. The key contributions of this manuscript are summarized as follows:

1. Assuming $B$ is Gaussian, we show that the relative error $\frac{||X^* - \hat{X}^{(0)}_r||_F}{||X^*||_F}$ critically depends on the parameter $d - s$. In particular, Theorem 4.1 establishes that the probability that the relative error exceeds $\sqrt{1 - \frac{s}{d}}$ decays at a sub-Gaussian rate.

2. Assuming $x^*$ is sub-Gaussian, we focus on the error $\left[ \sum_{j=1}^{m} ||x^*_j - \tilde{x}^{(0)}_j||_2 \right]$, which also depends critically on the parameter $d - s$. In particular, Theorem 4.2 in this work shows that the probability that the error exceeds $O(\sqrt{d - s})$ also decays at a sub-Gaussian rate.

3. For the $k$-sparse problem, we analyze the relative error $\frac{||X^* - \tilde{X}^{(0)}_k||_F}{||X^*||_F}$. Assuming $B$ is Gaussian and sufficiently tall (see Definition 4.4), we show that the relative error depends on the parameter $k/n$. Specifically, depending on the problem parameters, Theorem 4.5 shows that the probability of the error exceeding $O(k/n)$ decays at a rate ranging from inverse linear to exponential.

4. We apply the alternating minimization algorithm in [9] with the initializations (3) and (4). Experiments results on real datasets show superior performance compared to baseline methods.
Table 1: Table summarizing the frequently used notation used in this paper. See Figure 1 for definition and examples of r-local \( P^r \) and k-sparse \( P^k \) permutations. The superscript * in \( P^*, P^r, P^k, X^* \) denotes that these are unknown problem parameters which are estimated by the proposed algorithm.

Outline: The outline of the paper is as follows. Section 2 discusses related work. Section 3 presents the alternating minimization algorithms for the r-local and k-sparse permutation models. Section 4 covers the initialization analysis. Experimental results on real datasets showing superior performance compared to baseline methods are given in Section 5. We conclude in Section 6.

1.2 Notation

Boldface lowercase letters denote column vectors and boldface uppercase letters denote matrices. The transpose of a vector \( x \) is noted by \( x^\top \). For a matrix \( X \), its transpose is denoted by \( X^\top \). \( A^\dagger \) denotes the Moore-penrose pseudoinverse of \( A \). \( \text{Tr}(\cdot) \) denotes the trace of a matrix. Given two matrices \( A \) and \( B \), \( (A:B) \) denotes the trace inner product, defined as \( \text{Tr}(A^\top B) \). Given a vector \( x \), the \( i \)-th entry of \( x \) is denoted by \( x_i \). \( [A:B] \) denotes the vertical concatenation of matrices \( A, B \). \( \| \cdot \|_p \) denotes the vector p-norm. \( \| \cdot \|_F \) and \( \| \cdot \| \) denote the Frobenius norm and the operator norm of a matrix, respectively. \( 1_n \in \mathbb{R}^n \) denotes the vector of all ones. \( \mathds{I} \) denotes the identity matrix. \( \mathbb{R} \) denotes the set of real numbers. \( \Pi_n = \{Z : Z \in \{0,1\}^{n \times n}, Z1_n = 1_n, Z^\top 1_n = 1_n \} \) denotes the set of \( n \times n \) permutation matrices. \( \mathbb{E}(\cdot) \) denotes the expectation of a random variable in consideration. \( C, C' \) denote absolute constants > 1, and \( c, c' \) denote absolute constants \( \leq 1 \). \( \text{argmin} \) denotes the set of minima of an objective function in consideration. \( \exp(\cdot) \) denotes the exponential function. We also summarize the frequently used notation in Table 1.

2 Related Work

This section provides a concise overview of related work in unlabeled sensing theory and algorithms, inference problems involving unlabeled data, and applications of unlabeled sensing.

2.1 Theory and algorithms

We note that the problem in (1) is referred to as the single-view (multi-view) unlabeled sensing problem for \( m = 1 \) (\( m > 1 \)). For the single-view problem, we represent the problem as \( y = P^* B x^* + w \) where \( y, x^* \), and \( w \) denote the variables analogous to \( Y, X^* \) and \( W \) respectively. The work in [11] formulated the single-view unlabeled sensing problem and established that \( n = 2d \) measurements are both necessary and sufficient for the recovery of \( x^* \). Subsequent works [12, 13] generalized this result and developed an information-theoretic inachievability analysis.

For single-view unlabeled sensing, algorithms based on branch and bound and expectation maximization are proposed in [14, 15, 16], which are suitable for small problem sizes. A stochastic alternating minimization approach is introduced in [17]. For multi-view unlabeled sensing, the Levsort subspace matching algorithm was proposed in [18], and a subspace clustering approach was presented in [4].
The works [1, 2, 19] propose methods based on bi-convex optimization, robust regression, and spectral initialization, respectively.

2.2 Inference on unlabeled data

The problem of unlabeled sensing has been studied in the context of estimation and detection from unlabeled samples. In [20], the authors consider the problem of signal amplitude estimation and detection from unlabeled quantized binary samples. In the setup they consider, the ordering of the time indexes is unknown. The work in [20] proposes a maximum likelihood estimator to estimate the underlying signal amplitude and permutation, under certain structural assumptions on the quantization and signal profile. Furthermore, an alternating maximization algorithm is studied for the general estimation and detection problem.

In [21], the authors study signal recovery from unlabeled samples by considering a special case of unlabeled sensing, referred to as the unlabeled ordered sampling problem. In this problem, instead of estimating an unknown permutation matrix, the goal is to estimate an unknown selection matrix that preserves the order of the measurements. [21] links this problem to compressive sensing and also proposes an alternating minimization algorithm for solving it. [22] also addresses the unlabeled ordered sampling problem, but with a focus on signal detection, rather than signal recovery as in [21]. A dynamic programming algorithm is proposed therein to estimate the selection matrix.

[23] considers the problem of signal detection, where the observations are independently drawn from a finite alphabet. Focusing on the inference problem, the work in [23] characterizes the information available in the unordered samples by studying a binary hypothesis test. Additionally, it provides a computationally efficient detector to address the detection problem. [24] further studies the signal detection problem in the case of binary unlabeled observations. In the low-detectability regime, [24] gives an analytical characterization of various statistical inference quantities, such as Chernoff information. In [25], the signal detection framework of unlabeled sensing is applied to decision-making in sensor networks, where sensors report measurements to a central node, but the measurements are imprecise or lack labels. [26] considers a related problem, where sensors send their measurements to a central fusion center. Due to an anomaly in one of the sensors, the measurements at the fusion center are assumed to be unordered, and [26] explores the problem of detecting the anomaly with minimal detection delay.

2.3 Applications of unlabeled sensing

The linked linear regression problem [27, 28, 29], also called regression analysis with linked data, is to fit a linear model by minimizing over \( x \) and \( P \) the objective \( \| y - PBx \|_2 \), where \( P \) is an \( r \)-local permutation and \( x \) is the regression vector. For example, consider the statistical model where the weight depends linearly on age and height. Let \( y \in \mathbb{R}^n \) contain the weights of \( n = 10 \) individuals, 5 males, 5 females. Let \( B \in \mathbb{R}^{n \times d} \), with \( d = 2 \) contain the age and height values. 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 \( y, B \) contain the records of male (female) individuals, the unknown \( r \)-local permutation, \( r = 5 \), assigns each weight value (row of \( y \)) to its corresponding age, height (row of \( B \)). Detailed references describing record linkage with blocking are [30], Section 1.1 and [31], Section 2; also, see Fig. 2. Experimental results on real datasets are given in Section 5. Other applications of unlabeled sensing include the pose and correspondence problem [13] and 1-d unassigned distance geometry [9]. In [32], the authors apply the unlabeled sensing framework to the sensor network localization problem. We also note a recent work that extends the unlabeled sensing theory to multi-channel signals, leading to highly structured unlabeled sensing problems [33]. In [33], beyond theoretical analysis of the structured problem, the model is applied to a real-world application in calcium imaging.

2.4 Technical background

This section provides definition of sub-Gaussian and sub-exponential random vectors, and also summarizes the key concentration inequalities used throughout this paper. Let \( Z^* \in \mathbb{R}^{d \times m} \) be a matrix such that the columns \( z^*_j \) are independent, zero-mean sub-Gaussian random vectors, with sub-Gaussian norm \( K \). It follows then that

\[
E[\exp(\alpha^T x^*_j)] \leq \exp(\|\alpha\|^2 K^2/2) \forall \alpha \in \mathbb{R}^d \text{ and } \forall j \in \{1, ..., m\}. \tag{5}
\]
A random variable $X$ is sub-exponential with sub-exponential norm $K_{d-s}$ if
\[
\Pr[|X| \geq t] \leq C \exp(-t/K_{d-s}),
\]
where $C \geq 1, t \geq 0$ and $K_{d-s} \geq 0$.

**Theorem 2.1** (Hansson-Wright Inequality, Theorem 2.1 in [35]). Let $\Sigma = A^T A$ be a positive semi-definite matrix. Let $x = (x_1, \ldots, x_d)$ be a zero-mean sub-Gaussian random vector, i.e., for $\alpha \in \mathbb{R}^d$, $K \geq 0$
\[
\mathbb{E}[\exp(\alpha^T x^\ast)] \leq \exp(||\alpha||_2^2 K^2/2).
\]
For $t \geq 0$,
\[
\Pr[||AX||_2^2 \geq K^2 (\text{Tr}(\Sigma) + 2\sqrt{\text{Tr}(\Sigma^2)}t + 2t||\Sigma||)] \leq e^{-t}.
\]

**Lemma 2.2** (Johnson-Lindenstrauss Lemma, Lemma 5.3.2 in [36]). 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 $1 - 2 \exp(-c t^2 q)$,
\[
(1 - t) \sqrt{\frac{q}{p}} ||z||_2 \leq ||Pz||_2 \leq (1 + t) \sqrt{\frac{q}{p}} ||z||_2.
\]

**Theorem 2.3** (Hoeffding’s inequality, Theorem 2.6.2 in [36]). Let $X_1, \ldots, X_m$ be independent, sub-Gaussian random variables. Then, for every $t \geq 0$,
\[
\Pr[\sum_{i=m}^{i=m} X_i - E[X_i] \geq t] \leq 2 \exp \left( \frac{-c t^2}{\sum_{i=1}^{i=m} ||X_i||_p^2} \right),
\]
where $||X_i||_p$ denotes the sub-Gaussian norm of $X_i$ and $c'$ is an absolute constant.

**Lemma 2.4** (Tail inequality for $\chi_2^2$ distributed random variables, Lemma 1 in [37]). Let $Z_D$ be a $\chi^2$ statistic with $D$ degrees of freedom. For any positive $t$,
\[
\Pr[Z_D \geq D + 2\sqrt{Dt} + 2t] \leq e^{-t},
\]
\[
\Pr[Z_D \leq D - 2\sqrt{Dt}] \leq e^{-t}.
\]

### 3 Algorithm

In this section, we briefly summarize the alternating minimization algorithm first proposed in [9]. To estimate the $P^\ast$ and $X^\ast$ in the optimization objective (1), the alternating minimization (AltMin) updates for (2) are
\[
P^{(t)} = \arg\min_{P \in \Pi_n} \langle -YX^{(t)} \top B \top, P \rangle,
\]
\[
X^{(t+1)} = (P^{(t)} B) \top Y,
\]
where $P^{(t)}$ and $X^{(t)}$ denote the estimate of $P^\ast$ and $X^\ast$ at the $t$-th iteration respectively. The initialization to (12), (13), a linear assignment problem and a least squares problem, has to be chosen carefully because (2) is a non-convex optimization problem. After initialization, we alternate (12), (13) and use the relative change in the objective value $F(X, P)$ as the stopping criterion. For $r$-local $P^\ast_r$, the $P$-update (12) decouples along the blocks of $P^\ast_r$, see Algorithm 1. For $k$-sparse $P^\ast_k$, see Algorithm 2.
Algorithm 1 AltMin for r-local $P_1^*$

**Input:** Convergence threshold $\epsilon$, $Y$, $B$ and block sizes $r_1, r_2, \ldots, r_s$ of $P_1^*, P_2^*, \ldots, P_s^*$ respectively.

1. $\hat{X} \leftarrow$ collapsed initialization in (3)
2. $\hat{Y} \leftarrow BX$
3. while relative change $> \epsilon$ do
   4. for $i \in 1 \cdots s$ do // is the number of blocks
      5. $\hat{P}_i \leftarrow \text{argmin}_{P \in \Pi_s} - \langle Y\hat{Y}_i^T, P \rangle$
   6. $\hat{P} \leftarrow \text{blkdiag}(\hat{P}_1, \ldots, \hat{P}_s)$
   7. $\hat{X} \leftarrow B\hat{P}^TY$
   8. $\hat{Y} \leftarrow BX$
   9. Return $\hat{P}, \hat{X}$

Algorithm 2 AltMin for k-sparse $P_k^*$

**Input:** convergence threshold $\epsilon$, $Y$, $B$
1. $\hat{Y} \leftarrow Y$
2. while relative change $> \epsilon$ do
3. $\hat{P} \leftarrow \text{argmin}_{P \in \Pi_s} - \langle Y\hat{Y}^T, P \rangle$
4. $\hat{X} \leftarrow B\hat{P}^TY$
5. $\hat{Y} \leftarrow BX$
6. Return $\hat{P}, \hat{X}$

4 Initialization analysis

In this section, we derive upper bounds for the initialization error of the alternating minimization algorithm applied to solve the exact unlabeled sensing problem ($(1)$ with $W = 0$), using the initializations in (3) and (4).

4.1 Analysis for r-local permutation

We consider the initialization in (3). Let $\tilde{B} = \tilde{US}\tilde{V}^T$ denote the compact singular value decomposition of $\tilde{B}$. Using this in (3), we obtain $\tilde{B}^T = \tilde{V}\tilde{S}^{-1}\tilde{U}^T$, where $\tilde{B} = US\tilde{V}^T$. Using this substitution, it can be shown that the initialization error $\|X^* - \hat{X}^{(0)}\|_F$ is the projection of $X^*$ onto the orthogonal complement of the row space of $B$, as follows:

$$
\|X^* - \hat{X}^{(0)}\|_F = \|X^* - \tilde{B}^T\tilde{B}X^*\|_F = \|(I - \tilde{V}\tilde{V}^T)X^*\|_F. \quad (14)
$$

Recall that the size of $\tilde{B}$ is $s \times d$. If $s \geq d$, and assuming that $\tilde{B}$ has rank $d$, it can be verified that $\hat{X}^{(0)} = X^*$. For instance, for $s \geq d$, if the entries of $B$ are drawn independently from a continuous distribution, the rank of $B$ is $d$ with high probability. Given that, in this section, we upper bound the error in initialization for the under-determined $s < d$ case. Similar analysis for under-determined systems have been studied in the sketching literature [38], but these results are not applicable here as our sub-sampling strategy via the collapsing initialization is deterministic.

The first key theoretical result is Theorem 4.1. This theorem considers a Gaussian matrix $B \in \mathbb{R}^{n \times d}$, with $X^* \in \mathbb{R}^{d \times m}$ assumed to be a fixed but unknown matrix. To upper bound the relative error $\frac{\|X^* - \hat{X}^{(0)}\|_F}{\|X^*\|_F}$, we apply the Johnson–Lindenstrauss lemma. The key insight is that the relative error depends on the parameter $d - s$. Specifically, the probability that the relative error exceeds $\sqrt{1 - \frac{s}{d}}$ decays at a sub-Gaussian rate. This implies that the initialization in (3) improves as the number of blocks $s$ in the r-local model increases.

**Theorem 4.1.** Let $X^* \in \mathbb{R}^{d \times m}$ be the fixed unknown matrix and let $\hat{X}^{(0)}_r$ be as defined in (3). Consider the exact unlabeled sensing problem. Assuming Gaussian $B \in \mathbb{R}^{n \times d}$ and block-diagonal $P^*_r$ with $s$ blocks, for $\log m \leq ct^2(d - s)$, $s < d$, and $t \geq 0$,

$$
\Pr \left[ \frac{\|X^* - \hat{X}^{(0)}_r\|_F}{\|X^*\|_F} \geq (1 + t)\sqrt{\frac{d - s}{d}} \right] \leq 2\exp(-c(d - s)t^2). \quad (15)
$$
Proof. The error (14) is the column-wise projection of \( \mathbf{X}^* \in \mathbb{R}^{d \times m} \) onto a \((d - s)\)-dimensional uniformly random subspace (14), which can be bounded by the JL Lemma (8). \qed

In Theorem 4.1, we considered a Gaussian matrix \( \mathbf{B} \) and a fixed, unknown matrix \( \mathbf{X}^* \). In the following, we fix the measurement matrix \( \mathbf{B} \) and introduce randomness in \( \mathbf{X} \). Before discussing our main result for this setting, we provide motivation for considering this scenario. One motivating application is the unassigned distance geometry problem (uDGP) [39, 40, 41], which involves recovering the configuration of points from a list of distances. We focus on the 1-dimensional uDGP, which can be framed as a structured unlabeled sensing problem, as described in [9], where \( \mathbf{x}^* \) represents the sought-after positions of the points on the 1D line. In this case, the matrix \( \mathbf{B} \) is deterministic (see equation (3) of [9]). For this setting, without any assumptions on the underlying points \( \mathbf{x}^* \), and assuming that the underlying permutation is \( r \)-local, the collapsed initialization may be suboptimal. To illustrate this, recall the error in the collapsed estimate: \( \| (\mathbf{I} - \mathbf{VV}^T) \mathbf{x}^* \|_F \). For any \( \mathbf{x}^* \) orthogonal to the subspace spanned by \( \mathbf{V} \), the collapsed estimate will be the zero vector. This shows that the following upper bound on the initialization error \( \| \mathbf{x}^*_0 - \mathbf{x}^* \| \leq \| (\mathbf{I} - \mathbf{VV}^T) \mathbf{x}^* \| = \| \mathbf{x}^* \| \), holds with equality for \( \mathbf{x}^* \) orthogonal to the subspace spanned by the columns of \( \mathbf{V} \). The implication of the above discussion is that some structural assumption on the underlying \( \mathbf{x}^* \) is necessary to obtain suitable bounds on the initialization error.

We next consider the squared error in the initialization \( \| \mathbf{A} \mathbf{x}^* \|^2 \) where \( \mathbf{A} = \mathbf{I} - \mathbf{VV}^T \). Note that \( \mathbf{A} \) is an orthogonal projection matrix onto the orthogonal complement of the subspace spanned by \( \mathbf{V} \). Given that \( \mathbf{V} \in \mathbb{R}^{d \times s} \) and we are considering the underdetermined regime \( s < d \), the rank of \( \mathbf{A} \) is \( d - s \). Therefore, the analysis of the error is equivalent to analyzing the quadratic form \( \| \mathbf{A} \mathbf{x}^* \|^2 \). One common approach to control the norm of a random vector is to assume that \( \mathbf{x}^* \) is a random vector with independent, sub-Gaussian coordinates (see, for example, Theorem 3.1.1 in [36]). However, we note that this assumption does not hold after applying the projection operator \( \mathbf{A} \). A more suitable tool in this case is the Hanson-Wright inequality. In fact, the quadratic form in our case is special due to the fact that \( \mathbf{A} \) is a projection operator, allowing us to leverage its linear algebraic properties. To apply the Hanson-Wright inequality, it suffices to assume that \( \mathbf{x}^* \) is sub-Gaussian with mean zero, without requiring the independence of its coordinates. With this assumption, we study the error

\[
\sum_{j=1}^{m} \| \mathbf{x}^*_j - \mathbf{x}^*_0 \|_2^2 \leq 2 \exp \left( \frac{-ct^2}{mK_{d-s}} \right),
\]

Theorem 4.1 shows that this error depends critically on the parameter \( d - s \). In particular, Theorem 4.2 demonstrates that the probability of the error exceeding \( O(\sqrt{d - s}) \) decays at a sub-Gaussian rate.

**Theorem 4.2.** Let \( \mathbf{X}^* \in \mathbb{R}^{d \times m} \) be the unknown matrix such that the columns \( \mathbf{x}^*_j \) are independent, zero-mean sub-Gaussian random vectors, with sub-Gaussian norm \( K \). Let \( \mathbf{X}^{(0)} = [\mathbf{x}^{(0)}_1 \cdots \mathbf{x}^{(0)}_m] \) be as defined in (3). Consider the exact unlabeled sensing problem. For any fixed measurement matrix \( \mathbf{B} \) and block-diagonal \( \mathbf{P}_r \) with \( s \) blocks, \( s < d \) and \( t \geq 0 \),

\[
\Pr \left[ \sum_{j=1}^{m} \| \mathbf{x}^*_j - \mathbf{x}^*_0 \|_2^2 - mC_{d-s} \geq t \right] \leq 2 \exp \left( \frac{-ct^2}{mK_{d-s}} \right),
\]

where \( C_{d-s} = K(d - s + \frac{3}{2} \sqrt{d - s} + 2\frac{1}{2} K_{d-s}^2) \), \( K_{d-s} = K^2 \sqrt{d - s} \).

The strategy for the proof of Theorem 4.2 is to first use the Hanson-Wright inequality to control the quadratic form \( \| \mathbf{A} \mathbf{x}^* \|^2 \). Once that is established, we use Lemma 4.3 to define a modified random variable that is sub-exponential. The rest of the technical proof uses the Hoeffding bound on the modified random variable and standard techniques to relate the concentration inequality of \( \| \mathbf{z} \|_2^2 \) to \( \| \mathbf{z} \|_2 \) [36]. For the complete proof, see the Appendix 7.

**Lemma 4.3.** Let \( \mathbf{x}^{(0)}_r \) be as in (3), \( \mathbf{x}^* \in \mathbb{R}^{d} \) be a zero-mean sub-Gaussian random vector, \( \mathbf{B} \) be a fixed measurement matrix, and \( \mathbf{P}_r \) be a fixed block-diagonal permutation with \( s \) blocks. Let \( \mathbf{x}_{err} \) denote the random variable \( \mathbf{x}_{err} \equiv \| \mathbf{x}^* - \mathbf{x}^{(0)}_r \|_2^2 \).

1. For \( s < d \) and \( t \geq 0 \),

\[
\Pr[|\mathbf{x}_{err} - C_{d-s} | \geq t] \leq \exp(-ct/K_{d-s}),
\]

where \( K_{d-s} = K^2 \sqrt{d - s} \), \( C_{d-s} = K^2(d - s + \frac{1}{2} \sqrt{d - s}) \).
2. Define the random variable $\tilde{x}_{\text{err}}$ as follows:

$$\tilde{x}_{\text{err}} = \begin{cases} 
C_{d-s} & x_{\text{err}} \leq C_{d-s}, \\
 x_{\text{err}} & x_{\text{err}} > C_{d-s}.
\end{cases}$$

(18)

The random variable $\tilde{x}_{\text{err}}$ is a sub-exponential random variable with sub-exponential norm $\|\tilde{x}_{\text{err}} - C_{d-s}\|_{\text{tv}} = C \tilde{K}_{d-s}$. In addition, $\tilde{x}_{\text{err}} \geq x_{\text{err}}$, and $\forall t > 0, \Pr[\tilde{x}_{\text{err}} - C_{d-s} \geq t] = \Pr[x_{\text{err}} - C_{d-s} \geq t]$.  

(19) 

(v) 

Proof. We prove each part separately.

1. To derive (17), set $A = (I - \hat{V} \hat{V}^T)$ in (7), where $\hat{V} \in \mathbb{R}^{s \times d}$ denotes the basis for the row space of the collapsed matrix $\tilde{B} \in \mathbb{R}^{s \times d}$ in (3). It follows that $x_{\text{err}} \equiv \|x^* - \tilde{x}(0)\|_2^2 = \|A x^*\|_2^2$. Here, $A \in \mathbb{R}^{d \times d}$ is a $(d - s)$-dimensional projection matrix, and $\Sigma = A^T A = A A = A$. Moreover, $\operatorname{Tr}(\Sigma) = \operatorname{Tr}(\Sigma^2) = \operatorname{Tr}(A) = d - s$ and the operator norm satisfies $\|\Sigma\| = 1$. We now apply Hanson-Wright inequality in (7), we obtain:

$$\Pr[\|A x\|_2^2 

≥ K^2(d - s + 2\sqrt{(d - s)}t + 2t) \leq e^{-t}.$$ 

(21)

The next step will be to upper bound the term $2\sqrt{(d - s)}t$. To do that, we rely on following inequality: $2\sqrt{b}t \leq 2t + \frac{1}{2}b$, for $b \geq 1$ and $t \geq 0$. To verify this, note that this inequality is equivalent to checking whether $4t \leq 4t^2 + 2ta + a/4$ holds.

- For $a = 1$, this reduces to $0 \leq (2t - 1/2)^2$, which holds for all $t \geq 0$.
- For $a \geq 2$, we check $4t \leq 4t^2 a + 2ta + 1/4$ which is true for $t \geq 0$.

Using this inequality with $a = d - s$, we obtain $2\sqrt{(d - s)}t \leq 2t(d - s) + \frac{1}{2}(d - s)$. Substituting this bound in (21), we simplify the inequality as follows:

$$\Pr[\|A x\|_2^2 \geq K^2(d - s + 2t + (d - s) + 2t) \leq e^{-t}.$$ 

(22)

The above inequality is equivalent to:

$$\Pr[\|A x\|_2^2 - K^2(d - s + \frac{1}{2}(d - s) \geq 2K^2 t((d - s) + 1)] \leq e^{-t}.$$ 

(23)

A change of variable, $u = 2K^2 t((d - s) + 1)$, and minor algebraic manipulation yields the final desired result.

Remark: To derive Theorem 4.2, we first show that $x_{\text{err}}$ is a sub-exponential random variable. It cannot be concluded from (17) that the shifted random variable $x_{\text{err}} - C_{d-s}$ is sub-exponential because the lower tail probability $\Pr[x_{\text{err}} - C_{d-s} \leq 0]$ is not bounded. Since our goal is to upper bound the probability that the error exceeds a certain value, we define the non-negative sub-exponential random variable $\tilde{x}_{\text{err}} - C_{d-s}$, which upper bounds the lower tail as $\{x_{\text{err}} \leq C_{d-s}\} = C_{d-s}$, see (18). We will use the definition of $\tilde{x}_{\text{err}}$ given in (18), (19).

2. Conditioning on the two complementary events in (18), (19), by the law of total probability, $\forall t > 0$, we have

$$\Pr[\tilde{x}_{\text{err}} - C_{d-s} \geq t] = \Pr[\tilde{x}_{\text{err}} - C_{d-s} \geq t | x_{\text{err}} \leq C_{d-s}] \Pr[x_{\text{err}} \leq C_{d-s}] + \Pr[\tilde{x}_{\text{err}} - C_{d-s} \geq t | x_{\text{err}} > C_{d-s}] \Pr[x_{\text{err}} > C_{d-s}]$$

(24)

(24) follows from (18); specifically, $\Pr[\tilde{x}_{\text{err}} - C_{d-s} \geq t | x_{\text{err}} - C_{d-s} = 0] = 0 \forall t > 0$. (25) follows from (19). (26) follows from noting that the event $\{x^* | x_{\text{err}} \geq C_{d-s} + t\}$ is a subset of $\{x^* | x_{\text{err}} > C_{d-s}\}$. From (26), for $t > 0$

$$\Pr[x_{\text{err}} - C_{d-s} \geq t] = \Pr[\tilde{x}_{\text{err}} - C_{d-s} \geq t].$$

(27)
(27) follows from noting that $\hat{x}_{\text{err}} - C_{d-s} = |\hat{x}_{\text{err}} - C_{d-s}|$. Substituting (17) in (27), for $t > 0$,

$$
\Pr[|\hat{x}_{\text{err}} - C_{d-s}| \geq t] \leq \exp \left( \frac{-ct}{D_{d-s}} \right). \tag{28}
$$

Since $\exp(0) = 1$, (28) holds for $t \geq 0$. In (28), we have verified the definition in (6) for $\hat{x}_{\text{err}} - C_{d-s}$ with $K_{d-s} = 2K^2(\sqrt{d} - s + 1) = CK^2\sqrt{d} - s$. By proposition 2.7.1 (a), (d), definition 2.7.5 in [36] and (28), the sub-exponential norm is $\|\hat{x}_{\text{err}} - C_{d-s}\|_{\gamma_1} = CK_{d-s}$.

\[ \square \]

4.2 Analysis for $k$-sparse permutation

In this section, we analyze the initialization step when the underlying permutation is $k$-sparse. Specifically, we study the relative error $\|\hat{X} - \hat{X}^{(0)}\|_F^2 / \|\hat{X}\|_F^2$. Assuming that $B$ is a “tall” Gaussian matrix, Theorem 4.5 establishes that this error depends on the parameter $k/n$, which represents the proportion of shuffled entries relative to the total. More precisely, Theorem 4.5 shows that the probability of the error exceeding $O(k/n)$ decays at a rate ranging from inverse linear to exponential. Central to Theorem 4.5 is the assumption that the measurement matrix $B$ is a “tall” Gaussian matrix, which we define precisely below.

Definition 4.4 ([42]). A Gaussian matrix $B \in \mathbb{R}^{m \times d}$ is considered “tall” if the aspect ratio $\lambda = d/n$ satisfies $\lambda < \lambda_0$ for some sufficiently small constant $\lambda_0 > 0$. In that case, we have $\Pr[\sigma_{\text{min}}(B) \geq c\sqrt{n}] \leq e^{-cn}$, where $c$ is an absolute constant that depends on $\lambda_0$.

Theorem 4.5. Let $P_k^*$ be the fixed unknown $k$-sparse permutation matrix with $\langle I, P_k^* \rangle = n - k$. Let $X^* \in \mathbb{R}^{d \times m}$ be the fixed unknown matrix. Let $B$ be a “tall” Gaussian measurement matrix $B$ as defined in Definition 4.4. For $\frac{2C\log(m)}{\sqrt{k}} \leq \epsilon \leq \frac{8}{\sqrt{k}}C$, where $C = 4(1 + \sqrt{3})$, and $1 \leq k \leq n$, we have

$$
\Pr \left[ \frac{\|X^* - \hat{X}^{(1)}\|_F^2}{\|X^*\|_F^2} \geq \left( \frac{2}{C^2 + \epsilon} \right) \frac{k}{n} \right] \leq 8 \exp \left( -\frac{\epsilon \sqrt{k}}{2C} \right). \tag{29}
$$

For the proof of (29), see Section 7 in the Appendix.

Remark: In the above theorem, we consider the error bound at the lower bounds and upper bounds of $\epsilon$. When $\epsilon = \frac{2C\log(m)}{\sqrt{k}}$, it can be shown that the failure probability is $\frac{8}{m}$. When $\epsilon = \frac{8}{\sqrt{k}}C$, the failure probability is $\exp(-cn)$. (29) gives a $(< 1)$ relative error bound when $k$ grows slowly with $n$, for example $k = n^\beta$, $\beta < 1$.

5 Results

We compare the proposed AltMin algorithm to several benchmark methods on both synthetic (Figure 3) and real (Table 3) datasets. We implemented all algorithms in MATLAB. The linear assignment problem to recover the permutation estimate $P$ is solved by using the MATLAB matchpairs function. The least squares estimate (line 7 of Algorithm 1 and line 4 of Algorithm 2) is solved by computing the Moore-Penrose pseudoinverse using the MATLAB function pinv. The convex optimization problem proposed in [2] is solving using CVX [43, 44]. The specific solver we used is ScDuMi. We also use CVX to project onto the set of doubly stochastic matrices for the benchmark method ‘ds+’. GitHub repository: [github.com/aubbab02/ksparse-and-rlocal].

Baselines. We compare against six baseline methods, ‘ℓ2-regularized’ [2], ‘ds+’ [2], ‘Spectral’ [19], ‘Biconvex’ [1], ‘RLUS’ [7], and ‘Stochastic AltMin’ [17]. The ‘ℓ2-regularized’ method considers the $k$-sparse permutation model and imposes a row-wise group sparse penalty (ℓ2-norm regularization) on $M$, where $\hat{Y} = B\hat{X} + M$. Other methods are discussed in the following paragraphs. For the proposed algorithms, (12), (13) are alternated until the change in the objective value is less than 1 percent.

Synthetic data generation. To simulate $Y = P^*B_n \times X^*_{d \times m} + W$, we generate data by drawing the entries of matrices $B, X^*$ and $W$ from the normal distribution. Subsequently, $W$ is scaled by $\sigma$ to set SNR $\triangleq \|X^*\|_F^2 / (m\sigma^2) = 100$. The permutation matrices $P^*_r$ and $P^*_k$ are sampled uniformly from the set of $r$-local and $k$-sparse permutations, respectively. The results are averaged over 15 Monte-Carlo runs.
Figure 3: \( Y = P^*B_{n \times d}X^*_{k \times m} + W \). In figures (a,b,c,d), the normalized Hamming distortion \( d_H/n \) is plotted on the y-axis against block size \( r \) (a) and the number of shuffles (b,c,d). Hamming distortion \( d_H \) is the number of mismatches in estimate \( \hat{P} \) of \( P^* \) and is defined as \( d_H = \sum_i \mathbb{1}(\hat{P}(i) \neq P^*(i)) \), where \( P(i) \) denotes the column index of the 1 entry in the \( i^{th} \) row of the permutation matrix \( P \). A lower value of Hamming distortion is better.

Table 2: Description of the datasets used to compare the proposed algorithm with baseline methods. The results are given in Table 3. Here, \( n \) is the number of data points, \( d \) is the number of features, and \( m \) is the number of response variables.

| Dataset | \( n \) | \( d \) | \( m \) |
|---------|---------|---------|---------|
| ltp     | 335     | 30      | 6       |
| ames    | 2747    | 6       | 1       |
| scs     | 8966    | 35      | 16      |
| air-qty | 27605   | 27      | 16      |

Table 3: \( r_{\text{max}} \) denotes the largest block size of \( P^*_b \) and \( s \) denotes the number of blocks. For a description of the datasets, see Table 2. The methods are compared by the coefficient of determination \( R^2 \) and relative error (\( R_2 \), relative error). The relative error is \( ||X^* - \tilde{X}||_F/||X^*||_F \), where \( X^* = B^T Y^* \) is the ‘oracle’ regression matrix given unpermuted data \( Y^* \). The coefficient \( R_2(X) = 1 - (||Y^* - BX||_F/||Y^*||_F) \) measures the goodness of fit for the unpermuted data. The ‘naive’ estimate from permuted data \( Y \) is \( \tilde{X} = B^T Y \). The coefficient \( R_2 \) is bounded \( 0 \leq R_2 \leq 1 \), and a higher value of \( R_2 \) indicates a better fit.
Enforcing block-diagonal constraints. 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 block-diagonal:

- The Spectral method in [3] studies the unlabeled sensing model \( \mathbf{Y} = \mathbf{P}_r \mathbf{B} \mathbf{X} + \mathbf{W} \), where \( \mathbf{W} \) is an additive Gaussian noise, with no assumption made on the structure of the permutation, i.e., the underlying permutation is generic. The proposed algorithm in [3] estimates \( \mathbf{P} \) using

\[
\min_{\mathbf{P} \in \Pi_n} \langle \mathbf{P}, \mathbf{Y} \mathbf{Y}^\top \mathbf{B} \mathbf{B}^\top \rangle.
\]

In the case that the underlying permutation is r-local, first recall that \( s \) denotes the number of blocks in the r-local permutation and \( r_i \) is the size of the \( i \)-th block. The unlabeled sensing model reduces to

\[
\mathbf{Y}_i = \mathbf{P}_r \mathbf{B}_i \mathbf{X}_i + \mathbf{W}_i,
\]

where \( \mathbf{Y}_i \in \mathbb{R}^{r_i \times m}, \mathbf{B}_i \in \mathbb{R}^{r_i \times d} \) and \( \mathbf{W}_i \in \mathbb{R}^{r_i \times m} \) respectively denote blocks of the matrices \( \mathbf{Y} \in \mathbb{R}^{n \times m}, \mathbf{B} \in \mathbb{R}^{n \times d} \) and \( \mathbf{W} \in \mathbb{R}^{n \times m} \). Therefore, accounting for the structure of the r-local permutation, we modify (30) as follows:

\[
\min_{\mathbf{P} \in \Pi_{r_i}} \langle \mathbf{P}_i, \mathbf{Y}_i \mathbf{Y}_i^\top \mathbf{B}_i \mathbf{B}_i^\top \rangle \quad \forall i \in [s],
\]

which are \( s \) linear assignment problems over the sets of permutation matrices of size \( r_i \).

- For the ‘Biconvex’ algorithm, we modify the \( \mathbf{P}_1, \mathbf{P}_2 \) updates in Algorithm 1 from [1]. Let \( \mathbf{C}_1 = \mathbf{P}_r \mathbf{D} \mathbf{P}_r^\top + \mu - \rho \mathbf{P}_r^\top \mathbf{P}_r \), where \( \mu \) and \( \rho \) are ADMM parameters, and \( \mathbf{P}_r \) is the projection matrix onto the column-span of \( \mathbf{B} \). Instead of updating \( \mathbf{P}_1^{(t+1)} = \arg\min_{\mathbf{P} \in \Pi_n} \langle \mathbf{P}, \mathbf{C}_1 \rangle \), we enforce the block diagonal constraint by updating \( \mathbf{P}_1^{(t+1)} = \arg\min_{\mathbf{P}_i \in \Pi_{r_i}} \langle \mathbf{P}_i, \mathbf{C}_1, i \rangle \), where \( \mathbf{C}_1, i \in \mathbb{R}^{r_i \times r_i} \) is the matrix comprised of entries from the \( r_i \) rows and \( r_i \) columns in the \( i \)-the block of \( \mathbf{C} \). We update \( \mathbf{P}_2^{(t+1)} \) in a similar manner.

- The ‘ℓ_2-regularized’ method considers the \( k \)-sparse permutation model and imposes a row-wise group sparse penalty on \( \mathbf{M} \), where \( \hat{\mathbf{Y}} = \mathbf{B} \mathbf{X} + \mathbf{M} \). Specifically, the proposed estimate is obtained by solving the minimization problem:

\[
\min_{\mathbf{X}, \mathbf{M}} \| \mathbf{Y} - \mathbf{B} \mathbf{X} \|^2_F + \lambda \sum_{i=1}^n \| \mathbf{M}_i \|_2,
\]

where \( \mathbf{M}_i \) denotes the \( i \)-th row of \( \mathbf{M} \). We do not modify this minimization over \( \mathbf{X} \) and \( \mathbf{M} \). Instead, We substitute this estimate, denoted by \( \mathbf{X} \), and let \( \hat{\mathbf{Y}} = \mathbf{B} \mathbf{X} \). To estimate the underlying permutation, we consider the following minimization problem \( \min_{\mathbf{P}} \| \mathbf{Y} - \mathbf{P} \hat{\mathbf{Y}} \|_F^2 \). We note the following equalities:

\[
\begin{align*}
\arg\min_{\mathbf{P}} \| \mathbf{Y} - \mathbf{P} \hat{\mathbf{Y}} \|_F^2 &= \arg\min_{\mathbf{P}} \| \mathbf{Y} \|_F^2 + \| \mathbf{P} \hat{\mathbf{Y}} \|_F^2 - 2 \langle \mathbf{Y}, \mathbf{P} \hat{\mathbf{Y}} \rangle \\
&= \arg\max_{\mathbf{P}} \langle \mathbf{Y}, \mathbf{P} \hat{\mathbf{Y}} \rangle = \arg\max_{\mathbf{P}} \mathbf{Tr}(\mathbf{Y}^\top \mathbf{P} \hat{\mathbf{Y}}) = \arg\max_{\mathbf{P}} \langle \mathbf{Y} \hat{\mathbf{Y}}^\top, \mathbf{P} \rangle.
\end{align*}
\]

In the case that the underlying permutation is r-local, the above minimization problem decouples, and we obtain a block-wise linear assignment problem:

\[
\hat{\mathbf{P}}_i \leftarrow \arg\min_{\mathbf{P}_i \in \Pi_{r_i}} -\langle \mathbf{Y}_i, \hat{\mathbf{Y}}_i^\top, \mathbf{P}_i \rangle,
\]

where \( \hat{\mathbf{Y}}_i \in \mathbb{R}^{r_i \times m} \) and \( \mathbf{Y}_i \in \mathbb{R}^{r_i \times m} \) denote blocks of the matrices \( \hat{\mathbf{Y}} \) and \( \mathbf{Y} \), respectively.

Figures 3a, 3b. The results show that our algorithm recovers \( \mathbf{P}^* \) with decreasing block size \( r \) and number of shuffles \( k \). This confirms the conclusions of Theorems 4.1, 4.2, and 4.5 as the initialization improves with lower values of \( r \) and \( k \). The proposed AltMin algorithm is also applicable to both models and computationally scalable. For \( r = 125 \) (Fig. 3a) and \( k = 850 \) (Fig. 3b), MATLAB runtime with 16 Gb RAM and 9-th Gen. 4-core processor is less than a second.
The entries of the measurement matrix $B$ are sampled i.i.d. from the uniform $[0, 1]$ distribution. Compared to the case for Gaussian $B$ (Fig. 3b), the performance of the ‘Spectral’ and ‘RLUS’ methods deteriorates significantly. This is because both algorithms consider quadratic measurements $YY^\top$. Specifically, the ‘Spectral’ method is based on the spectral initialization technique [45], which assumes a Gaussian measurement matrix. In contrast, the performance of the proposed and ‘$\ell_2$-regularized’ methods does not deteriorate.

The ‘ds+’ algorithm [2] considers the convex relaxation of (2) by minimizing the objective over the set of doubly stochastic matrices. Assuming a known upper bound on the number of shuffles $k$, ‘ds+’ constrains $(I, P) \geq n - k$. To project onto the set of doubly stochastic matrices, each iteration of ‘ds+’ minimizes a linear program which greatly increases its run-time. The proposed AltMin algorithm optimizes the same objective, but over the set of permutation matrices. This results in a simpler linear assignment problem, which is why AltMin is faster and outperforms ‘ds+’.

The ‘scs’ dataset, one of the 7 response variables is excluded to improve the model fit. For ‘ftp’, ‘ames’, ‘sba’, ‘ames’ and ‘scs’, the values of a feature (column of the design matrix) are rounded and data points with the same feature value are assigned to the same block and permuted. The ‘air-qty’ dataset contains time-stamped readings with year, month, day, and hour information and we follow the experimental setup designed in the ‘Case Study’ in Section 4 of [2]: in row 4 (5) of Table 3, readings with the same month and day (day and hour) are assigned to the same block and permuted. The regression model for ‘air-qty’ is also as defined in [2], and we also use a moving-median filter with window size 32 to remove outliers. The proposed AltMin algorithm outperforms the competing baselines. ‘AltMin-$k$’, i.e., AltMin initialized to $P^{(0)} = I$, is also competitive, possibly because permuting correlated rows does not greatly corrupt the design matrix $B$. Results for ‘Spectral’ and ‘Biconvex’ are omitted because the methods were not competitive.

### Table 3

For the linked linear regression problem (Section 2.3), we report results on the three datasets from [2], available at [46, 47, 48], and the Ames [49] housing datasets. For all datasets, the columns of the response variables $Y$ and the design matrix $B$ are centered (zero-mean). $B$ is also replaced by its top $d$ principal components. For the ‘sba’ dataset, one of the 7 response variables is excluded to improve the model fit. For ‘ftp’, ‘ames’, ‘sba’, ‘ames’ and ‘scs’, the values of a feature (column of the design matrix) are rounded and data points with the same feature value are assigned to the same block and permuted.

The ‘air-qty’ dataset contains time-stamped readings with year, month, day, and hour information and we follow the experimental setup designed in the ‘Case Study’ in Section 4 of [2]: in row 4 (5) of Table 3, readings with the same month and day (day and hour) are assigned to the same block and permuted. The regression model for ‘air-qty’ is also as defined in [2], and we also use a moving-median filter with window size 32 to remove outliers. The proposed AltMin algorithm outperforms the competing baselines. ‘AltMin-$k$’, i.e., AltMin initialized to $P^{(0)} = I$, is also competitive, possibly because permuting correlated rows does not greatly corrupt the design matrix $B$. Results for ‘Spectral’ and ‘Biconvex’ are omitted because the methods were not competitive.

### Conclusion

In this paper, we studied a fast alternating minimization algorithm for the unlabeled sensing problem under two structured permutation models: $k$-sparse and $r$-local. The initialization of this non-convex algorithm plays a crucial role in its performance. To address this, we proposed two initialization strategies tailored to the respective permutation models. Our primary contribution lies in the characterization of the initialization error, providing theoretical insights into its impact on algorithm performance. Additionally, we show the competitive performance of the algorithm on both synthetic and real datasets. While the current work focuses on analyzing the initialization, one direction for future research is to study the rate of convergence and establish conditions under which the algorithm converges provably to the unknown parameters of the unlabeled sensing problem.

### Acknowledgements

Abiy Tasissa thanks Professor R. Adamczak for correspondence on [50]. Abiy Tasissa’s research is partially supported by NSF DMS 2208392.
References

[1] H. Zhang, M. Slawski, and P. Li, “Permutation recovery from multiple measurement vectors in unlabeled sensing,” in 2019 IEEE International Symposium on Information Theory (ISIT), 2019, pp. 1857–1861.

[2] M. Slawski, E. Ben-David, and P. Li, “Two-stage approach to multivariate linear regression with sparsely mismatched data.” J. Mach. Learn. Res., vol. 21, no. 204, pp. 1–42, 2020.

[3] H. Zhang and P. Li, “Optimal estimator for unlabeled linear regression,” in Proceedings of the 37th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 11153–11162. [Online]. Available: https://proceedings.mlr.press/v119/zhang20n.html

[4] M. Slawski, M. Rahmani, and P. Li, “A sparse representation-based approach to linear regression with partially shuffled labels,” in Uncertainty in Artificial Intelligence. PMLR, 2020, pp. 38–48.

[5] X. Shi, X. Li, and T. Cai, “Spherical regression under mismatch corruption with application to automated knowledge translation,” Journal of the American Statistical Association, vol. 116, no. 536, pp. 1953–1964, 2021.

[6] M. Slawski and E. Ben-David, “Linear regression with sparsely permuted data,” Electronic Journal of Statistics, vol. 13, no. 1, pp. 1–36, 2019.

[7] A. A. Abbasi, A. Tasissa, and S. Aeron, “R-local unlabeled sensing: A novel graph matching approach for multiview unlabeled sensing under local permutations,” IEEE Open Journal of Signal Process., vol. 2, pp. 309–317, 2021.

[8] D. Li, A. Aubry, A. De Maio, Y. Fu, and S. Marano, “Detection by block- and band-permuted data,” IEEE Transactions on Signal Processing, vol. 70, pp. 5778–5790, 2022.

[9] A. A. Abbasi, A. Tasissa, and S. Aeron, “r-local unlabeled sensing: Improved algorithm and applications,” in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 5593–5597.

[10] Z. Wang, E. Ben-David, and M. Slawski, “Regularization for shuffled data problems via exponential family priors on the permutation group.” in International Conference on Artificial Intelligence and Statistics. PMLR, 2023, pp. 2939–2959.

[11] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, “Unlabeled sensing with random linear measurements,” IEEE Trans. Inf. Theory, vol. 64, no. 5, pp. 3237–3253, 2018.

[12] I. Dokmanić, “Permutations unlabeled beyond sampling unknown,” IEEE Signal Processing Letters, vol. 26, no. 6, pp. 823–827, 2019.

[13] 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, 2018.

[14] 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), 2014, pp. 1040–1044.

[15] L. Peng and M. C. Tsakiris, “Linear regression without correspondences via concave minimization,” IEEE Signal Process. Letters, vol. 27, pp. 1580–1584, 2020.

[16] L. Peng, X. Song, M. C. Tsakiris, H. Choi, L. Kneip, and Y. Shi, “Algebraically-initialized expectation maximization for header-free communication,” in IEEE Int. Conf. on Acous., Speech and Signal Process. (ICASSP). IEEE, 2019, pp. 5182–5186.

[17] A. Abid and J. Zou, “A stochastic expectation-maximization approach to shuffled linear regression,” in 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2018, pp. 470–477.
[18] A. Pananjady, M. J. Wainwright, and T. A. Courtade, “Denoising linear models with permuted data,” in 2017 IEEE Int. Symposium on Inf. Theory (ISIT), 2017, pp. 446–450.

[19] H. Zhang and P. Li, “Optimal estimator for unlabeled linear regression,” in Int. Conference on Machine Learning. PMLR, 2020, pp. 11 153–11 162.

[20] G. Wang, J. Zhu, R. S. Blum, P. Willett, S. Marano, V. Matta, and P. Braca, “Signal amplitude estimation and detection from unlabeled binary quantized samples,” IEEE Transactions on Signal Processing, vol. 66, no. 16, pp. 4291–4303, 2018.

[21] S. Haghighatshoar and G. Caire, “Signal recovery from unlabeled samples,” IEEE Transactions on Signal Processing, vol. 66, no. 5, pp. 1242–1257, 2017.

[22] Z. Liu and J. Zhu, “Signal detection from unlabeled ordered samples,” IEEE Communications Letters, vol. 22, no. 12, pp. 2431–2434, 2018.

[23] S. Marano and P. K. Willett, “Algorithms and fundamental limits for unlabeled detection using types,” IEEE Transactions on Signal Processing, vol. 67, no. 8, pp. 2022–2035, 2019.

[24] S. Marano and P. Willett, “Making decisions by unlabeled bits,” IEEE Transactions on Signal Processing, vol. 68, pp. 2935–2947, 2020.

[25] Z. Sutton, P. Willett, S. Marano, and Y. Bar-Shalom, “Identity-aware decision network communication budgeting: Is who as important as what?” IEEE Transactions on Aerospace and Electronic Systems, vol. 59, no. 5, pp. 5203–5217, 2023.

[26] Z. Sun and S. Zou, “Quickest anomaly detection in sensor networks with unlabeled samples,” IEEE Transactions on Signal Processing, vol. 71, pp. 873–887, 2023.

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

[28] P. Lahiri and M. D. Larsen, “Regression analysis with linked data,” Journal of the American statistical association, vol. 100, no. 469, pp. 222–230, 2005.

[29] Y. Han and P. Lahiri, “Statistical analysis with linked data,” International Statistical Review, vol. 87, pp. S139–S157, 2019.

[30] G. Kim and R. Chambers, “Regression analysis under incomplete linkage,” Computational Statistics & Data Analysis, vol. 56, no. 9, pp. 2756–2770, 2012.

[31] Z. Wang, E. Ben-David, G. Diao, and M. Slawski, “Regression with linked datasets subject to linkage error,” Wiley Interdisciplinary Reviews: Computational Statistics, vol. 14, no. 4, p. e1570, 2022.

[32] G. Wang, S. Marano, J. Zhu, and Z. Xu, “Target localization by unlabeled range measurements,” IEEE Transactions on Signal Processing, vol. 68, pp. 6607–6620, 2020.

[33] T. Koka, M. C. Tsakiris, M. Muma, and B. B. Haro, “Shuffled multi-channel sparse signal recovery,” Signal Processing, p. 109579, 2024.

[34] P. Ravikumar and W. Cohen, “A hierarchical graphical model for record linkage,” arXiv preprint arXiv:1207.4180, 2012.

[35] D. Hsu, S. Kakade, and T. Zhang, “A tail inequality for quadratic forms of subgaussian random vectors,” Electronic Communications in Probability, vol. 17, pp. 1–6, 2012.

[36] R. Vershynin, High-dimensional probability: An introduction with applications in data science. Cambridge university press, 2018, vol. 47.

[37] B. Laurent and P. Massart, “Adaptive estimation of a quadratic functional by model selection,” The Annals of Statistics, vol. 28, no. 5, pp. 1302 – 1338, 2000. [Online]. Available: https://doi.org/10.1214/aos/1015957395

[38] D. P. Woodruff et al., “Sketching as a tool for numerical linear algebra,” Foundations and Trends® in Theoretical Computer Science, vol. 10, no. 1–2, pp. 1–157, 2014.
A Proof Of Theorem 4.2

Proof. Recall that $\hat{x}_{err}$ defined in (18), (19) is sub-exponential (see Lemma 4.3). For $t \geq 0,$

$$\Pr[|\hat{x}_{err} - E[\hat{x}_{err}]| \geq t] = \Pr[|\hat{x}_{err} - C_{d-s} - E[\hat{x}_{err} - C_{d-s}]| \geq t]$$

$$\leq 2 \exp(-c't/\|\hat{x}_{err} - C_{d-s} - E[\hat{x}_{err} - C_{d-s}]\|_{\varphi_1})$$

$$\leq 2 \exp(-c't/K_{d-s}),$$

where (31) follows from noting that $\hat{x}_{err} - C_{d-s}$ is sub-exponential, shown in Lemma 4.3. (32) is by the centering property, (see proposition 2.7.1 (a), (d) in [36]), which bounds

$$\|\hat{x}_{err} - C_{d-s} - E[\hat{x}_{err} - C_{d-s}]\|_{\varphi_1} \leq C\|\hat{x}_{err} - C_{d-s}\|_{\varphi_1},$$

and that $\|\hat{x}_{err} - C_{d-s}\|_{\varphi_1} = K_{d-s}$, by Lemma 4.3. For $z \geq 0, |z - 1| \geq \delta \implies |z^2 - 1| \geq \max(\delta, \delta^2).$ Therefore,

$$\Pr \left[ \frac{\sqrt{\hat{x}_{err}} - \sqrt{E[\hat{x}_{err}]}}{\sqrt{E[\hat{x}_{err}]}} \geq \delta \right] \leq \Pr \left[ \frac{\hat{x}_{err} - E[\hat{x}_{err}]}{E[\hat{x}_{err}]} \geq \delta^2 \right]$$

$$\leq 2 \exp(-c'E[\hat{x}_{err}]\delta^2/K_{d-s}),$$

(33)
where (33) follows by substituting $t = \delta^2$ in (32). We make a change of variable $t = \delta \sqrt{E[\hat{x}_{err}]}$ in (33) and obtain

$$\Pr \left[ \| \hat{x}_{err} - \sqrt{E[\hat{x}_{err}]} \| \geq t \right] \leq 2 \exp(-c't^2/K_{d-s}).$$

From (34) and proposition 2.5.2 (i), (iv) and definition 2.5.6 in [36], the sub-Gaussian norm squared is bounded as

$$\| \sqrt{x}_{err} - \sqrt{E[x_{err}]} \|^2 = c' K_{d-s}. \quad (35)$$

Applying Hoeffding’s inequality (9) to $\sum_{j=1}^{m} x_{err,j} - \sqrt{E[x_{err,j}]}$ and substituting $E[\sqrt{x_{err,j}}] - \sqrt{E[x_{err,j}]} \leq 0$ (Jensen’s),

$$\Pr \left[ \sum_{j=1}^{m} \sqrt{x_{err,j}} - \sqrt{E[x_{err,j}]} \geq t \right] \leq 2 \exp \left( -\frac{c't^2}{mK_{d-s}} \right). \quad (36)$$

For non-negative random variable $X$, $E[X] = \int_{0}^{\infty} \Pr[X > t]dt$. Applied to $(x_{err} - C_{d-s})/2K^2(\sqrt{d-s} + 1)$,

$$\Pr \left[ \frac{x_{err} - C_{d-s}}{2K^2(\sqrt{d-s} + 1)} \leq \int_{0}^{\infty} \exp(-t)dt = 1. \quad (37)$$

In (37), we have substituted the tail probability upper bounds from Lemma 4.3. Substituting $C_{d-s}$ from (17) in (37),

$$E[x_{err}] \leq K^2(d-s + 5/2\sqrt{d-s} + 2). \quad (38)$$

By definition in (18), (19), $\| x_j - \hat{x}_j \|_2 \leq \sqrt{x_{err,j}} \forall j \in [m]$. Substituting (38) and using $\| x_j - \hat{x}_j \|_2 \leq \sqrt{x_{err}}$ in (36) completes the proof. \hfill \Box

## B Proof of Theorem 4.5

We first provide a high-level road map for the proof. For simplicity of explanation, we focus on the single-view problem. In that case, let $y^* = Bx^*$ and $\hat{y}^{(0)} = P_k^t B y^*$. Here, $y^*$ represents the original measurement, and $\hat{y}^{(0)}$ corresponds to the measurement after being shuffled by a $k$-sparse permutation. The starting point is to consider the quantity $\| y^* - \hat{y}^{(0)} \|^2$. Since this quantity is invariant under permutation, we assume, without loss of generality, that the first $k$ rows of $B$ are shuffled. Lemma B.1 shows that $\| y^* - \hat{y}^{(0)} \|^2$ can be expressed as the difference of two Chi-square random variables. Using a tail inequality for Chi-square distributed random variables (see Lemma 4.3), Lemma B.1 provides a bound for this quantity. With this result established, the main crux of the proof of Theorem 4.5 is to relate $\sum_{j=1}^{m} \| y_j^* - \hat{y}_j^{(0)} \|^2$ to $\sigma_{min}(B)\| x^* - \hat{X}^{(1)} \|_F^2$.

**Proof.** Let $E$ denote the error term such that

$$\hat{Y}^{(0)} = B\hat{X}^{(1)} + E, \quad (39)$$

where $\hat{X}^{(1)} = B^t \hat{Y}^{(0)}$, $\hat{Y}^{(0)} = Y = P_k^t B x^*$ and $E \perp R(B)$ is orthogonal to the range space of $B$ since $\hat{X}^{(1)} = \arg \min_x \| Y^{(0)} - B x \|_F$. Combining $Y^* = Bx^*$ and (39), we obtain:

$$\| Y^* - \hat{Y}^{(0)} \|_F^2 = \| B(x^* - \hat{X}^{(1)}) \|_F^2 + \| E \|_F^2 \geq \sigma_{min}(B)\| x^* - \hat{X}^{(1)} \|_F^2. \quad (40)$$

The inequality in (40) can equivalently be rewritten as:

$$\sigma_{min}(B)\| x^* - \hat{X}^{(1)} \|_F^2 \leq \sum_{j=1}^{m} \| y_j^* - \hat{y}_j^{(0)} \|^2. \quad (41)$$

Let $E_j$ denote the event $E_j = \{ \| x_j^* \| \geq \| y_j^* - \hat{y}_j^{(0)} \|_2, \forall j \in [m] \}$. From (40) and applying the union bound to $\cup_{j=1}^{m} E_j$ using (46),

$$\Pr \left[ \sigma_{min}(B)\| x^* - \hat{X}^{(1)} \|_F^2 \geq 2k + C\sqrt{k}t + 10l \right] \leq 7m e^{-t}. \quad (41)$$

16
For \( t \geq 2\log m \),
\[
Pr \left[ \frac{\|X^* - \hat{X}^{(1)}_F\|^2_F}{\|X^*\|^2_F} - \frac{2k}{\sigma_{\min}^2} \geq \frac{C\sqrt{k}}{\sigma_{\min}} t \right] \leq 7e^{-t/2}. \tag{42}
\]

By Theorem 1.1 in [42], for a “tall” Gaussian matrix \( B \in \mathbb{R}^{n \times d} \), where \( B \) is considered tall if the aspect ratio \( \lambda = d/n \) satisfies \( \lambda < \lambda_0 \) for some sufficiently small constant \( \lambda_0 > 0 \), we have \( \Pr[\sigma_{\min}(B) \leq \epsilon\sqrt{n}] \leq e^{-cn} \) (see Subsection 1.2 of [42]). We consider the union bound of the failure probabilities \( e^{-cn} \) and \( 7e^{-t/2} \). A minor calculation yields that \( e^{-cn} + 7e^{-t/2} \leq 8e^{-t/2} \) for \( t \leq cn \). Consequently, for \( 2\log(m) < t < cn \), we obtain
\[
Pr \left[ \frac{\|X^* - \hat{X}^{(1)}_F\|^2_F}{\|X^*\|^2_F} - \frac{2k}{\epsilon^2 n} \geq \frac{C\sqrt{k}}{n} t \right] \leq 8e^{-t/2}. \tag{43}
\]

We now make a change of variables \( t' = C\sqrt{k}t \). For \( \frac{2C(\log(m))}{\epsilon n} \leq t' \leq C\sqrt{k} \), we have
\[
Pr \left[ \frac{\|X^* - \hat{X}^{(1)}_F\|^2_F}{\|X^*\|^2_F} - \frac{2k}{\epsilon^2 n} \geq t' \right] \leq 8\exp \left( -\frac{t'n}{2Ck} \right). \tag{44}
\]

We make a further change of variable with \( t' = \epsilon k \left( \frac{t}{n} \right) \). For \( \frac{2C(\log(m))}{\epsilon \sqrt{k}} \leq \epsilon \leq \frac{Cn}{\sqrt{k}} \), we obtain
\[
Pr \left[ \frac{\|X^* - \hat{X}^{(1)}_F\|^2_F}{\|X^*\|^2_F} \geq \frac{2}{c^2 + \epsilon} \frac{k}{n} \right] \leq 8\exp \left( -\frac{\epsilon \sqrt{k}}{2C} \right). \tag{45}
\]

(45) provides a useful \((< 1)\) relative error bound when \( k \) grows slowly with \( n \), for example \( k = n^\beta \), \( \beta < 1 \).

\[\square\]

**Lemma B.1.** Let \( P^*_k \) be the fixed unknown \( k \)-sparse permutation matrix with \( \langle I, P^*_k \rangle = n - k \). Consider the exact unlabeled sensing problem. Let \( x^* \) be the fixed unknown vector, \( y^* = Bx^* \), \( \hat{y}^{(0)} = P^*_k Bx^* \). Assuming Gaussian \( B \), for \( k \in [n] \) and \( t \geq 0 \), we have
\[
Pr \left[ \|y^* - \hat{y}^{(0)}\|^2_2 \geq \left( 2k + 4(1 + \sqrt{3})\sqrt{kt} + 10t \right) \|x^*\|^2_2 \right] \leq 7e^{-t}. \tag{46}
\]

**Proof.** Under the \( k \)-sparse assumption on \( P^* \), the known vector \( y \) has \( k \) shuffled entries. Assuming, without loss of generality, that the first \( k \) rows of \( B^* \) are shuffled,
\[
\frac{\|y^* - \hat{y}^{(0)}\|^2_2}{\|x^*\|^2_2} = \frac{1}{\|x^*\|^2_2} \sum_{i=1}^{i=k} (b^\top_{P(i)} x^* - b^\top_{P(i)} x^*)^2 = 2 \sum_{i=1}^{i=k} \frac{(b_{P(i)}^\top x^*)^2}{\|x^*\|^2_2} - 2 \sum_{i=1}^{i=k} \frac{x^* b_{P(i)}^\top b^\top_{P(i)} x^*}{\|x^*\|^2_2} \tag{47}
\]

\( T_1 \), defined in (47), is the sum of \( k \) independent Chi-square random variables and is bounded, using (11), as follows:
\[
Pr \left[ T_1 \geq k + 2\sqrt{kt} + 2t \right] \leq e^{-t}. \tag{48}
\]

The product random variables in \( T_2 \) are distributed as the difference of two independent \( \chi^2 \) random variables.
\[
\frac{b_{P(i)}^\top x^* b_{P(i)}^\top x^*}{\|x^*\|^2_2} \sim \frac{1}{2} Z^2_i - \frac{1}{2} Z^2_i \quad \forall i \in n - k + 1, \ldots, n. \tag{49}
\]

The random variables (rv) in (49) are not mutually independent, but each rv depends on, at most, two other rvs. To see this, let permutation \( P \) such that \( P(i) \mapsto j \), then \( b_{P(i)}^\top x^* b_{P(i)}^\top x^* \) is not independent of
\[
b_{P(i)}^\top x^* b_{P(j)}^\top x^*, \ b_{P(i)}^\top x^* b_{P(j)}^\top x^*. \tag{50}
\]

The \( k \) rvs in (49) 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 \( T_P \), where
\[
T_P \triangleq \frac{1}{\|x^*\|^2_2} \sum_{i \in P} b_{P(i)}^\top x^* b_{P(i)}^\top x^* = \frac{1}{2} \sum_{i \in P} Z^1_i - \frac{1}{2} \sum_{i \in P} Z^2_i, \tag{51}
\]
is upper bounded in probability as

$$\Pr[T_P \leq -2\sqrt{k_1}t - t] \leq 2\exp(-t).$$  \hfill (52)

(52) follows from applying the union bound to probabilities $p_1, p_2$ which are given by

$$p_1 = \Pr \left[ \sum_{i \in P} Z_i^1 \leq k_1 - 2\sqrt{k_1}t \right] \leq e^{-t},$$  \hfill (53)

$$p_2 = \Pr \left[ \sum_{i \in P} Z_i^2 \geq k_1 + 2\sqrt{k_1}t + 2t \right] \leq e^{-t},$$  \hfill (54)

and bounding $p_1, p_2$ using the tail inequalities (11) and (10), respectively. Defining $T_Q, T_R$ similarly to (51), with cardinalities $k_2, k_3$ and applying the union bound as in (53), (54) gives

$$\Pr \left[ T_2 \leq -2(\sqrt{k_1}t + \sqrt{k_2}t + \sqrt{k_3}t) - 3t \right] \leq 6e^{-t},$$  \hfill (55)

where $T_2 = T_P + T_Q + T_R$. Since $\| [\sqrt{k_1} \sqrt{k_2} \sqrt{k_3}]^T \|_2^2 \leq 3\|\cdot\|_2^2 = 3k$, $\sqrt{k_1} + \sqrt{k_2} + \sqrt{k_3} \leq \sqrt{3k}$. Substituting in (55),

$$\Pr \left[ T_2 \leq -2\sqrt{3kt} - 3t \right] \leq 6e^{-t}. \quad (56)$$

Applying the union bound to (48), (56) gives the result in (46) with $C = 4(1 + \sqrt{3})$. \hfill \qed