Sharp information-theoretic thresholds for shuffled linear regression

Leon Lufkin
Department of Statistics and Data Science
Yale University
New Haven, CT, USA
Email: leon.lufkin@yale.edu

Yihong Wu
Department of Statistics and Data Science
Yale University
New Haven, CT, USA
Email: yihong.wu@yale.edu

Jiaming Xu
The Fuqua School of Business
Duke University
Durham, NC, USA
Email: jiaming.xu868@duke.edu

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619515

***Abstract*—This paper studies the problem of shuffled linear regression, where the correspondence between predictors and responses in a linear model is obfuscated by a latent permutation. Specifically, we consider the model $y = \Pi_* X \beta_* + w$, where X is an $n \times d$ standard Gaussian design matrix, w is Gaussian noise with entrywise variance σ^2, Π_* is an unknown $n \times n$ permutation matrix, and β_* is the regression coefficient, also unknown. Previous work has shown that, in the large n-limit, the minimal signal-to-noise ratio (SNR), $\|\beta_*\|_2^2/\sigma^2$, for recovering the unknown permutation exactly with high probability is between n^2 and n^C for some absolute constant C and the sharp threshold is unknown even for $d = 1$. We show that this threshold is precisely $\mathsf{SNR} = n^4$ for exact recovery throughout the sublinear regime $d = o(n)$. As a by-product of our analysis, we also determine the sharp threshold of almost exact recovery to be $\mathsf{SNR} = n^2$, where all but a vanishing fraction of the permutation is reconstructed.**

I. INTRODUCTION

Consider the following linear model, where we observe

$$y = \Pi_* X \beta_* + w, \tag{1}$$

Here $X \in \mathbb{R}^{n\times d}$ is the design matrix, $\beta_* \in \mathbb{R}^d$ is the unknown regression coefficient, Π_* is an unknown $n \times n$ permutation matrix that shuffles the rows of X, and $w \in \mathbb{R}^n$ is observation noise. The goal is to recover Π_* and β_* on the basis of observing X and y.

If Π_* is known, (1) is the familiar linear regression. Otherwise, this problem is known as shuffled regression [1], [2], unlabeled sensing [3]–[5], or linear regression with permuted/mismatched data [6]–[8], as the correspondence between the predictors (the rows x_i's of X) and the responses (y_i's) is lost. As such, it is a much more difficult problem as one needs to jointly estimate the permutation Π_* and the regression coefficients β_*. This is a problem of considerable theoretical and practical interest. For applications in areas such as robotics, data integration, and de-anonymization, we refer the readers to [3, Sec. 1] and [5, Sec. 1.1].

A line of work has studied the minimal signal-to-noise ratio (SNR) that is required to reconstruct Π_*. Following [1], [9], in this paper we consider a random design X with $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and Gaussian noise $w_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$, which are independent from each other. Define

$$\mathsf{SNR} \triangleq \frac{\|\beta_*\|_2^2}{\sigma^2}. \tag{2}$$

It is shown in [1, Theorems 1 and 2] that for exact recovery (namely, $\hat{\Pi} = \Pi_*$ with probability tending to one), the required SNR is between n^2 and n^C for some absolute constant C. Intriguingly, numerical simulation carried out for $d = 1$ (see [1, Fig. 2]) suggests that there is a sharp threshold $\mathsf{SNR} = n^{C_0}$ for some constant C_0 between 3 and 5.

The major contribution of this work is to resolve this question by showing that the sharp threshold for exact recovery is $\mathsf{SNR} = n^4$ for all dimensions satisfying $d = o(n)$. Along the way, we also resolve the optimal threshold for achieving almost exact reconstruction, namely, $\mathsf{overlap}(\hat{\Pi}, \Pi_*) = 1 - o(1)$, where

$$\mathsf{overlap}(\hat{\Pi}, \Pi_*) \triangleq \frac{1}{n}\mathsf{Tr}(\hat{\Pi}^\top \Pi_*)$$

is the fraction of covariants that are correctly unshuffled. In other words, if $\hat{\pi}$ and π_* are permutations corresponding to $\hat{\Pi}$ and Π_*, then $\mathsf{overlap}(\hat{\Pi}, \Pi_*) = \frac{1}{n}|\{i \in [n] : \hat{\pi}(i) = \pi(i)\}|$.

II. MAXIMUM LIKELIHOOD AND QUADRATIC ASSIGNMENT

A natural idea for the joint estimation of (Π_*, β_*) is the maximum likelihood estimator (MLE) [1]:

$$(\hat{\Pi}, \hat{\beta}) = \underset{\Pi \in S_n, \beta \in \mathbb{R}^d}{\arg\min} \|y - \Pi X \beta\|_2^2, \tag{3}$$

where S_n denotes the set of all $n \times n$ permutation matrices. Since β_* has no structural assumptions such as sparsity, $n \geq d$ is necessary even when there is no noise and Π_* is known. By classical theory on linear regression, for a fixed Π the optimal β for (3) is given by

$$\hat{\beta}_\Pi \triangleq (X^\top X)^{-1} X^\top \Pi^\top y \tag{4}$$

and $\|y - \Pi X \hat{\beta}_\Pi\|_2^2 = \|\mathcal{P}_{(\Pi X)^\perp} y\|_2^2$, where

$$\mathcal{P}_{\Pi X} \triangleq \Pi \underbrace{X(X^\top X)^{-1} X^\top}_{\triangleq \mathcal{P}_X} \Pi^\top \tag{5}$$

$$\mathcal{P}_{(\Pi X)^\perp} \triangleq I_n - \mathcal{P}_{\Pi X} = \Pi \underbrace{(I_n - X(X^\top X)^{-1} X^\top)}_{\triangleq \mathcal{P}_{X^\perp}} \Pi^\top \tag{6}$$

A full version of this paper with proofs of all lemmas can be found at arxiv.org/abs/2402.09693.

are the projection matrices onto the column span of ΠX and its orthogonal complement respectively. Thus the ML estimator of Π_* can be written as[1]

$$\hat{\Pi} = \arg\max_{\Pi \in S_n} \|\mathcal{P}_{\Pi X} y\|_2^2. \tag{7}$$

This optimization problem is in fact a special instance of the *quadratic assignment problem* (QAP) [10]:

$$\max_{\Pi \in S_n} \langle A, \Pi^\top B \Pi \rangle, \tag{8}$$

where $A = yy^\top$ is rank-one and $B = \mathcal{P}_X$ is a rank-d projection matrix. For worst-case instances of (A, B), the QAP (8) is known to be NP-hard [11]. Furthermore, even solving the special case (7) is NP-hard provided that $d = \Omega(n)$ [1, Theorem 4]. On the positive side, for constant d it is not hard to see that this can be solved in polynomial time. Indeed, as the proof in Section V shows (see [9, Sec. 2] for a similar result), instead of (8), one can approximate the original (3) by discretizing and restricting β to an appropriate δ-net for $\delta = 1/\text{poly}(n)$. Since for fixed β, (3) becomes a very special case of the *linear assignment problem* (LAP) $\max_\Pi \langle y, \Pi X \beta \rangle$ which can be solved by sorting the vectors y and $X\beta$, the discretized version of (3) can be computed in $n^{O(d)}$-time. In fact, for the special case of $d = 1$, this can be made exact [1, Theorem 4].

III. Main Results

The following theorem on the statistical performance of the estimator (7) is the main result of this paper.

Theorem 1: Fix an arbitrary $\epsilon > 0$. Assume that $d = o(n)$.

(a) Exact recovery: If $\mathsf{SNR} \geq n^{4+\epsilon}$, then $\mathbb{P}[\hat{\Pi} = \Pi_*] = 1 - o(1)$ as $n \to \infty$, where $o(1)$ is uniform in Π_* and β_*.

(b) Almost exact recovery: If $\mathsf{SNR} \geq n^{2+\epsilon}$, then $\mathbb{P}[\mathsf{overlap}(\hat{\Pi}, \Pi_*) = 1 - o(1)] = 1 - o(1)$ as $n \to \infty$, where $o(1)$ is uniform in Π_* and β_*.

The positive results in Theorem 1 are in fact information-theoretically optimal. To see this, for the purpose of the lower bound, consider the case where Π_* is drawn uniformly at random and β_* is a known unit vector. Defining $x \triangleq X\beta_* \sim \mathcal{N}(0, I_n)$, we have $y = \Pi_* x + w$. Then the problem reduces to a special case of the linear assignment model studied in [12]–[14] where the goal is to reconstruct Π_* by observing x and y.[2] Specifically, applying [14, Theorem 3] for one dimension shows that exact (resp. almost exact) reconstruction is impossible unless $\sigma = o(n^{-2})$ (resp. $\sigma = o(n^{-1})$).

Next, let us comment on the role of the dimension d. As lower-dimensional problem instances can be embedded into higher dimensions by padding zeros to β_*, the minimum required SNR for recovery is non-decreasing in d. Theorem 1 shows the optimal thresholds for exact and approximate exact recovery are $\mathsf{SNR} = n^4$ and n^2 in the *sublinear regime* of $d = o(n)$. When the dimension is proportional to the sample size, say $d = \rho n$ for some constant $\rho \in (0, 1)$, we conjecture that the conclusion in Theorem 1 no longer holds and the sharp threshold depends on ρ. In fact, [1, Theorem 1] shows that the estimator (7) achieves exact recovery provided that $\mathsf{SNR} \geq n^{C/(1-\rho)}$ for some unspecified constant C. On the other hand, the simple lower bound argument above does not yield any dependency on ρ, since it assumes β_* is known and reduces the problem to $d = 1$. Determining the optimal threshold in the linear regime remains a challenging question.

IV. Further Related Work

The model (1) has been considered in the compressed sensing literature for zero observation noise ($\sigma = 0$), known as the unlabeled sensing problem, with the goal of recovering $\beta_* \in \mathbb{R}^d$ exactly. The work [3] showed that when the entries of X are sampled iid from some continuous probability distribution, *any* β_*, including adversarial instances (the so-called "for all" guarantee), can be recovered exactly with probability one if and only if one has $n \geq 2d$ observations. The paper shows this using a constructive proof, but it requires a combinatorial algorithm involving exhaustive search.

Moving to the weaker "for any" guarantee, the works [9], [15] also consider the noiseless setting and propose an efficient algorithm based on lattice reduction that recovers an arbitrary fixed β_* with probability one with respect to the random design, provided that $n > d$. Another approach based on method of moments is proposed in [2], where the empirical moments of $X\hat{\beta}$ are matched to those of y.

There is also a line of work on shuffled regression when the latent permutation is partially (or even mostly) known [5]–[8] that has found applications in analyzing census and climate data. This approach permits a robust regression formulation for estimating β_*, wherein the unknown permuted data points are treated as outliers, from which Π_* can then be estimated.

The problem of learning from shuffled data has also been considered in nonparametric settings, e.g., isotonic regression, where $y_i = f(x_i) + w_i$, for some $f : [0,1]^d \to \mathbb{R}$ that is coordinate-wise monotonically increasing, and the goal is to estimate f. When the x_i are permuted, this problem is known as *uncoupled* isotonic regression, which has been studied in [16] for $d = 1$ and in [17] for $d > 1$.

V. Proof of Theorem 1

Throughout the proof, we assume $\Pi_* = I_n$ without loss of generality. The proof of Theorem 1 follows a union bound over $\Pi \neq I_n$ and is divided into two parts: Section V-A deals with those permutations Π whose number of non-fixed points is at least ηn (for some $\eta = o(1)$ depending on d and ϵ). Section V-B deals with those permutations Π whose number of non-fixed points is at most ηn.

Although both [1] and the present paper analyze the estimator (3), the program of our analysis deviates from that in [1]

[1]We note that although $(\hat{\Pi}, \hat{\beta})$ defined in (3) is the MLE for (Π_*, β_*), it is unclear that $\hat{\Pi}$ itself (i.e., (7)) is optimal (that is, minimizing the probability of error $\mathbb{P}[\hat{\Pi} \neq \Pi_*]$ when Π_* is drawn uniformly at random), due to the presence of the nuisance parameter β_*.

[2]These works considered the more general setting where x, y are $n \times m$ Gaussian matrices and the respective threshold for exact and almost exact reconstruction has determined to be $n^{-2/m}$ and $n^{-1/m}$ for $m = o(\log n)$.

in the following two aspects, both of which are crucial for determining the sharp thresholds.

First, a key quantity appearing in the proof is the following moment-generating function (MGF):

$$\mathbb{E}\exp\left(-t\left\|X\beta_*-\Pi X\beta\right\|_2^2\right), \tag{9}$$

for a given Π and β, where $t \propto \frac{1}{\sigma^2}$. While similar quantities have been analyzed in [1], only a crude bound is obtained in terms of the number of fixed points of Π (see Lemma 4 and eq. (25-26) therein). Instead, inspired by techniques in [14] for random graph matching, we precisely characterize (9) in terms of the cycle decomposition of Π and β. In particular, to determine the sharp thresholds, it is crucial to consider *all* cycle types instead of just fixed points.

Second, recall that the MLE (3) involves a double minimization over Π and β. While it is straightforward to solve the inner minimization over β and obtain a closed-form expression for the optimal $\hat{\beta}_\Pi$ (4), directly analyzing the MLE with this optimal $\hat{\beta}_\Pi$ plugged in, namely, the QAP (8), turns out to be challenging. In particular, this requires a tight control of the MGF (9) with β replaced by $\hat{\beta}_\Pi$. While this is doable when Π is close to I_n, the analysis becomes loose when Π moves further away from I_n and requires suboptimally large SNR. Alternatively, we do not work with this optimal $\hat{\beta}_\Pi$ and instead take a union bound over a proper discretization (δ-net) of β. Importantly, the resolution δ needs to be carefully chosen so that the cardinality of the δ-net is not overwhelmingly large compared to (9). This part crucially relies on the sublinearity assumption $d = o(n)$ and the fact that Π has at least ηn non-fixed points.

A. *Proof for permutations with many errors*

In this part, we focus on the permutations that are far away from the ground truth and prove that

$$\mathbb{P}\left\{\mathsf{overlap}(\hat{\Pi}, I_n) \leq (1-\eta)\right\} = o(1), \tag{10}$$

for any fixed ϵ, provided that $\mathsf{SNR} \geq n^{2+\epsilon}$, $d = o(n)$, $\epsilon\eta n \geq 100d$, and $\eta \geq n^{-\epsilon/10}$. Note that here we only require $\mathsf{SNR} \geq n^{2+\epsilon}$ instead of $\mathsf{SNR} \geq n^{4+\epsilon}$. This directly implies the desired sufficient condition for the almost exact recovery and proves Part (b) of Theorem 1, with an appropriate choice of $\eta = o(1)$.

Let $\mathcal{S}(m)$ denote the set of permutation matrices with m fixed points. For a given r, let $B_r(\beta_*) \triangleq \{\beta : \|\beta - \beta_*\|_2 \leq r\}$. The following lemma shows that we can discretize β appropriately without inflating the objective too much.

Lemma 1: There exists a δ-net $N_\delta(r)$ for $B_r(\beta_*)$ such that $|N_\delta(r)| \leq (1+2r/\delta)^d$. Moreover, for any Π, if $\hat{\beta}_\Pi \in B_r(\beta_*)$,

$$\min_{\beta\in N_\delta(r)} \|y - \Pi X\beta\|_2^2 \leq \min_{\beta\in B_r(\beta_*)} \|y - \Pi X\beta\|_2^2 + \|X\|_{\mathrm{op}}^2 \delta^2.$$

Next, we introduce a set of high-probability events to facilitate our analysis of the MLE.

Lemma 2: Suppose $\mathsf{SNR} \geq 1$, $r/\delta \leq n^2$, $\eta \geq n^{-\epsilon/10}$, and $\epsilon\eta n \geq 100d$. The following events hold with probability $1 - o(1)$:

$$\begin{aligned}
\mathcal{E}_1 &\triangleq \{\|X\beta_* - \Pi X\beta\|_2^2 \geq n^{-2-\epsilon}\|\beta_*\|_2^2(n-n_1),\\
&\qquad \forall n_1 \leq (1-\eta)n, \forall \Pi \in \mathcal{S}(n_1), \forall\beta \in N_\delta(r)\},\\
\mathcal{E}_2 &\triangleq \{\|X\|_{\mathrm{op}} \leq C'\sqrt{n}\},\\
\mathcal{E}_3 &\triangleq \{\|\hat{\beta}_\Pi - \beta_*\|_2 \leq c\|\beta_*\|_2, \forall \Pi\},
\end{aligned}$$

for some absolute constants C', c, where $\hat{\beta}_\Pi$ is defined in (4).

Finally, we need a key lemma to bound the MGF of $\|X\beta_* - \Pi X\beta\|_2^2$. The proof crucially relies on the cycle decomposition of the permutation matrix Π. (See Appendices A-A and A-D for details.)

Lemma 3: Suppose $\|\beta_*\|_2/\sigma \geq n^{1+\epsilon/2}$, $\eta \geq n^{-\epsilon/10}$, and C is any constant. Then for all sufficiently large n,

$$\begin{aligned}
&\sum_{n_1=0}^{(1-\eta)n} C^{n-n_1} \sum_{\Pi\in\mathcal{S}(n_1)} \mathbb{E}\exp\left(-\frac{1}{32\sigma^2}\|X\beta_* - \Pi X\beta\|_2^2\right)\\
&\leq n^{-\epsilon\eta n/10}.
\end{aligned}$$

Now, we are ready to prove (10). By the definition of MLE given in (3),

$$\begin{aligned}
&\mathsf{overlap}(\hat{\Pi}, I_n) \leq (1-\eta)\\
&\Rightarrow \min_\beta \|y - \Pi X\beta\|_2^2 \leq \min_\beta \|y - X\beta\|_2^2\\
&\qquad \text{for some } \Pi \in \mathcal{S}(n_1) \text{ and } n_1 \leq (1-\eta)n.
\end{aligned}$$

Recall that $\hat{\beta}_\Pi = \arg\min_\beta \|y - \Pi X\beta\|_2^2$ and the definition of $\mathcal{E}_3$. By letting $r = c\|\beta_*\|_2$, we have

$$\begin{aligned}
&\min_\beta \|y - \Pi X\beta\|_2^2 \leq \min_\beta \|y - X\beta\|_2^2, \mathcal{E}_3\\
&\Rightarrow \min_{\beta\in B_r(\beta_*)} \|y - \Pi X\beta\|_2^2 \leq \|y - X\beta_*\|_2^2 = \|w\|_2^2\\
&\Rightarrow \min_{\beta\in N_\delta(r)} \|y - \Pi X\beta\|_2^2 \leq \|w\|_2^2 + \|X\|_{\mathrm{op}}^2\delta^2,
\end{aligned}$$

where the last implication follows from Lemma 1. Note that for any β,

$$\begin{aligned}
&\|y - \Pi X\beta\|_2^2 \leq \|w\|_2^2 + \|X\|_{\mathrm{op}}^2\delta^2\\
&\Rightarrow \|X\beta_* + w - \Pi X\beta\|_2^2 \leq \|w\|_2^2 + \|X\|_{\mathrm{op}}^2\delta^2\\
&\Rightarrow 2\langle X\beta_* - \Pi X\beta, w\rangle \leq -\|X\beta_* - \Pi X\beta\|_2^2 + \|X\|_{\mathrm{op}}^2\delta^2.
\end{aligned}$$

Now, recalling the definitions of $\mathcal{E}_1, \mathcal{E}_2$, we choose $\delta = C'\sqrt{\eta/2}n^{-1-\epsilon/2}\|\beta_*\|_2$, so that on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, for all $\Pi \in \mathcal{S}(n_1)$ and all $n_1 \leq (1-\eta)n$,

$$\|X\beta_* - \Pi X\beta\|_2^2 \geq 2\|X\|_{\mathrm{op}}^2\delta^2, \forall\beta \in N_\delta(r),$$

and hence

$$\begin{aligned}
&\min_{\beta\in N_\delta(r)} \|y - \Pi X\beta\|_2^2 \leq \|w\|_2^2 + \|X\|_{\mathrm{op}}^2\delta^2, \mathcal{E}_1 \cap \mathcal{E}_2\\
&\Rightarrow \exists\beta \in N_\delta(r) : 2\langle X\beta_* - \Pi X\beta, w\rangle \leq -\frac{1}{2}\|X\beta_* - \Pi X\beta\|_2^2.
\end{aligned}$$

In conclusion, we have shown that

$$\begin{aligned}&\mathsf{overlap}(\hat{\Pi}, I_n) \le (1-\eta), \mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\\ &\Rightarrow 2\langle X\beta_* - \Pi X\beta, w\rangle \le -\frac{1}{2}\|X\beta_* - \Pi X\beta\|_2^2\\ &\qquad \text{for some } \Pi\in\mathcal{S}(n_1), n_1 \le (1-\eta)n, \text{ and } \beta\in N_\delta(r).\end{aligned}$$

Now, for each fixed Π and β, we condition on X and then use the Gaussian tail bound, we get that

$$\begin{aligned}&\mathbb{P}\left\{2\langle X\beta_* - \Pi X\beta, w\rangle \le -\frac{1}{2}\|X\beta_* - \Pi X\beta\|_2^2\right\}\\ &\le \mathbb{E}\exp\left(-\frac{1}{32\sigma^2}\|X\beta_* - \Pi X\beta\|_2^2\right).\end{aligned}$$

It follows from the union bound that

$$\begin{aligned}&\mathbb{P}\left\{\mathsf{overlap}(\hat{\Pi}, I_n) \le (1-\eta), \mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\right\}\\ &\le |N_\delta(r)| \sum_{n_1\le(1-\eta)n}\sum_{\Pi\in\mathcal{S}(n_1)} \mathbb{E}\exp\left(-\frac{\|X\beta_* - \Pi X\beta\|_2^2}{32\sigma^2}\right).\end{aligned}$$

Finally, by Lemma 1, $|N_\delta(r)| \le (1+2r/\delta)^d$. Recall that we set $\delta = C'\sqrt{\eta/2}n^{-1-\epsilon/2}\|\beta_*\|_2$ and $r = c\|\beta_*\|_2$ for constants $c, C' > 0$. Therefore,

$$|N_\delta(r)| \le \left(Cn^{1+\epsilon/2}/\sqrt{\eta}\right)^d$$

for some universal constant $C > 0$. Combining the last displayed equation with Lemma 3 yields that

$$\begin{aligned}&|N_\delta(r)| \sum_{n_1\le(1-\eta)n}\sum_{\Pi\in\mathcal{S}(n_1)} \mathbb{E}\exp\left(-\frac{\|X\beta_* - \Pi X\beta\|_2^2}{32\sigma^2}\right)\\ &\le \left(Cn^{1+\epsilon/2}/\sqrt{\eta}\right)^d n^{-\epsilon\eta n/10} \le n^{-\epsilon\eta n/20},\end{aligned}$$

where the last inequality holds for all sufficiently large n due to the facts that $\epsilon\eta n \ge 100d$ and $\eta \ge n^{-\epsilon/10}$.

Finally, applying Lemma 2, we conclude that

$$\begin{aligned}&\mathbb{P}\left\{\mathsf{overlap}(\hat{\Pi}, I_n) \le (1-\eta)\right\}\\ &\le \mathbb{P}\left\{\mathsf{overlap}(\hat{\Pi}, I_n) \le (1-\eta), \mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\right\}\\ &\quad + \mathbb{P}\{\mathcal{E}_1^c\} + \mathbb{P}\{\mathcal{E}_2^c\} + \mathbb{P}\{\mathcal{E}_3^c\} = o(1).\end{aligned}$$

B. Proof for permutations with few errors

In this part, we focus on the permutations that are close to the ground truth and prove that

$$\mathbb{P}\left\{(1-\eta) \le \mathsf{overlap}(\hat{\Pi}, I_n) \le \frac{n-2}{n}\right\} \le n^{-\Omega(1)}, \quad (11)$$

provided that $\sigma/\|\beta_*\|_2 \le n^{-2-\epsilon}$, $\eta \le \epsilon/8$, and $d = o(n)$.

In this case, we can no longer tolerate the n^d factor arising from the discretization of the β parameter. To address the high-dimensional scenario where $d = o(n)$, we instead adopt the proof strategy outlined by [1] to analyze the QAP formulation (8). However, achieving the sharp threshold necessitates a more meticulous analysis than that employed by [1].

We first state several useful auxiliary lemmas. Recall that $\mathcal{S}(m)$ denotes the set of permutation matrices with m fixed points, and recall the projection matrices $\mathcal{P}_{\Pi X}$ and $\mathcal{P}_{(\Pi X)^\perp}$ as defined in (5)–(6).

Lemma 4: Let $n \ge 2$. Define $\mathcal{E}_4$ such that for all $n_1 \le n-2$ and all $\Pi\in\mathcal{S}(n_1)$,

$$\|\mathcal{P}_{\Pi X}(w)\|_2^2 - \|\mathcal{P}_X(w)\|_2^2 \le 10\sigma^2(n-n_1)\log n.$$

Then $\mathbb{P}\{\mathcal{E}_4\} \ge 1 - 4n^{-2}$.

Lemma 5: Suppose $\eta \le \epsilon/8$. Define $\mathcal{E}_5$ such that for all $(1-\eta)n \le n_1 \le n-2$ and all $\Pi\in\mathcal{S}(n_1)$,

$$\left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*)\right\|_2^2 \ge n^{-4-\epsilon}\|\beta_*\|_2^2(n-n_1).$$

Then $\mathbb{P}\{\mathcal{E}_5\} \ge 1 - n^{-\epsilon/8}$.

Lemma 6: Suppose $\sigma/\|\beta_*\|_2 \le n^{-2-\epsilon/2}$, $\eta \le \epsilon/8$, and C is any fixed constant. Then for all sufficiently large n,

$$\begin{aligned}&\sum_{n_1\ge(1-\eta)n}^{n-2} C^{n-n_1}\sum_{\Pi\in\mathcal{S}(n_1)}\mathbb{E}\exp\left(-\frac{1}{32\sigma^2}\left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*)\right\|_2^2\right)\\ &\le n^{-\epsilon/8}.\end{aligned}$$

Now, we are ready to prove (11). By the definition of the MLE given in (3),

$$\begin{aligned}&(1-\eta) \le \mathsf{overlap}(\hat{\Pi}, I_n) \le \frac{n-2}{n}\\ &\Rightarrow \left\|\mathcal{P}_{(\Pi X)^\perp}(y)\right\|_2^2 \le \|\mathcal{P}_{X^\perp}(y)\|_2^2\\ &\qquad \text{for some } \Pi\in\mathcal{S}(n_1) \text{ and } (1-\eta)n \le n_1 \le n-2.\end{aligned}$$

Since $\mathcal{P}_{X^\perp}(X\beta_*) = 0$, it follows that

$$\begin{aligned}&\left\|\mathcal{P}_{(\Pi X)^\perp}(y)\right\|_2^2 \le \|\mathcal{P}_{X^\perp}(y)\|_2^2\\ &\Leftrightarrow \left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*) + \mathcal{P}_{(\Pi X)^\perp}(w)\right\|_2^2 \le \|\mathcal{P}_{X^\perp}(w)\|_2^2\\ &\Leftrightarrow \left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*)\right\|_2^2 + 2\left\langle\mathcal{P}_{(\Pi X)^\perp}(X\beta_*), \mathcal{P}_{(\Pi X)^\perp}(w)\right\rangle\\ &\qquad \le \|\mathcal{P}_{X^\perp}(w)\|_2^2 - \left\|\mathcal{P}_{(\Pi X)^\perp}(w)\right\|_2^2\\ &\Leftrightarrow \left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*)\right\|_2^2 + 2\left\langle\mathcal{P}_{(\Pi X)^\perp}(X\beta_*), w\right\rangle\\ &\qquad \le \|\mathcal{P}_{\Pi X}(w)\|_2^2 - \|\mathcal{P}_X(w)\|_2^2.\end{aligned}$$

By our assumption that $\sigma/\|\beta_*\|_2 \le n^{-2-\epsilon}$, on event $\mathcal{E}_4\cap\mathcal{E}_5$, for all sufficiently large n, all $(1-\eta)n \le n_1 \le n-2$, and all $\Pi\in\mathcal{S}(n_1)$,

$$\|\mathcal{P}_{X^\perp}(w)\|_2^2 - \left\|\mathcal{P}_{(\Pi X)^\perp}(w)\right\|_2^2 \le \frac{1}{2}\left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*)\right\|_2^2.$$

Thus, on event $\mathcal{E}_4\cap\mathcal{E}_5$,

$$\begin{aligned}&(1-\eta) \le \mathsf{overlap}(\hat{\Pi}, I_n) \le \frac{n-2}{n}\\ &\Rightarrow \frac{1}{2}\left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*)\right\|_2^2 + 2\left\langle\mathcal{P}_{(\Pi X)^\perp}(X\beta_*), w\right\rangle \le 0\\ &\qquad \text{for some } \Pi\in\mathcal{S}(n_1) \text{ and } (1-\eta)n \le n_1 \le n-2.\end{aligned}$$

By the Gaussian tail bound,

$$\begin{aligned}&\mathbb{P}\left\{\frac{1}{2}\left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*)\right\|_2^2 + 2\left\langle\mathcal{P}_{(\Pi X)^\perp}(X\beta_*), w\right\rangle \le 0\right\}\\ &\le \mathbb{E}\exp\left(-\frac{1}{32\sigma^2}\left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*)\right\|_2^2\right).\end{aligned}$$

Therefore, applying union-bound yields that

$$\begin{aligned}&\mathbb{P}\left\{(1-\eta)\leq \mathsf{overlap}(\hat{\Pi},I_n)\leq \frac{n-2}{n},\mathcal{E}_4,\mathcal{E}_5\right\}\\&\leq \sum_{n_1\geq(1-\eta)n}^{n-2}\sum_{\Pi\in\mathcal{S}(n_1)}\mathbb{E}\exp\left(-\frac{1}{32\sigma^2}\left\|\mathcal{P}_{(\Pi X)^\perp}(X\beta_*)\right\|_2^2\right)\\&\leq n^{-\epsilon/8},\end{aligned}$$

where the last inequality holds by Lemma 6 and the assumption that $\sigma/\left\|\beta_*\right\|_2\leq n^{-2-\epsilon}$.

Finally, applying Lemma 4 and Lemma 5, we arrive at

$$\begin{aligned}&\mathbb{P}\left\{(1-\eta)\leq \mathsf{overlap}(\hat{\Pi},I_n)\leq \frac{n-2}{n}\right\}\\&\leq \mathbb{P}\left\{(1-\eta)\leq \mathsf{overlap}(\hat{\Pi},I_n)\leq \frac{n-2}{n},\mathcal{E}_4,\mathcal{E}_5\right\}\\&\quad+\mathbb{P}\left\{\mathcal{E}_4^c\right\}+\mathbb{P}\left\{\mathcal{E}_5^c\right\}\\&\leq 6n^{-\epsilon/8}.\end{aligned}$$

VI. Conclusions and Open Problems

In this paper we resolved the information-theoretically optimal thresholds for exactly and almost exactly recovering the unknown permutation in shuffled linear regression with random design in the sublinear regime of $d=o(n)$. In addition to determining the sharp threshold in the linear regime of $d=\Theta(n)$ mentioned in Section III, a few other problems remain outstanding.

First, the estimator (7) attaining the sharp thresholds involves solving the computationally expensive QAP problem. Although for low dimensions this can be approximately computed in $n^{O(d)}$ time, the resulting algorithm is far from practical as it involves searching over an δ-net in d dimensions. For $d\to\infty$, currently there is no polynomial-time algorithms except in the special case of $\sigma=0$ [9], [15].

Second, it is of interest to extend the current results to multivariate responses where each response y_i is m-dimensional for $m>1$. In other words, $y=\Pi_* X\beta_*+w$, where $\beta_*\in\mathbb{R}^{d\times m}$. This has been considered in several existing works such as [1], [4], [5], [8], where it is observed that multiple responses can significantly reduce the required SNR. Drawing from existing results on related models in LAP and QAP [13], [14], we conjecture that the optimal thresholds for exact and almost exact recovery are given by $\mathsf{SNR}=n^{4/m}$ and $n^{2/m}$, respectively, provided that m is not too large. While one can deduce the lower bound from that in [14] by considering the oracle setting of a known β_*, analyzing the counterpart of (7) remains open.

Acknowledgement

This research was supported in part by an NSF Career Award CCF-2144593.

References

[1] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Linear regression with shuffled data: Statistical and computational limits of permutation recovery," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3286–3300, 2017.

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

[3] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, "Unlabeled sensing with random linear measurements," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3237–3253, 2018.

[4] 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)*. IEEE, 2019, pp. 1857–1861.

[5] H. Zhang and P. Li, "Optimal estimator for unlabeled linear regression," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 153–11 162.

[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] R. Mazumder and H. Wang, "Linear regression with partially mismatched data: local search with theoretical guarantees," *Mathematical Programming*, vol. 197, no. 2, pp. 1265–1303, 2023.

[8] M. Slawski, E. Ben-David, and P. Li, "Two-stage approach to multivariate linear regression with sparsely mismatched data," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 8422–8463, 2020.

[9] D. J. Hsu, K. Shi, and X. Sun, "Linear regression without correspondence," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] T. C. Koopmans and M. Beckmann, "Assignment problems and the location of economic activities," *Econometrica: journal of the Econometric Society*, pp. 53–76, 1957.

[11] S. Sahni and T. Gonzalez, "P-complete approximation problems," *Journal of the ACM (JACM)*, vol. 23, no. 3, pp. 555–565, 1976.

[12] O. E. Dai, D. Cullina, and N. Kiyavash, "Database alignment with gaussian features," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 3225–3233.

[13] D. Kunisky and J. Niles-Weed, "Strong recovery of geometric planted matchings," in *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2022, pp. 834–876.

[14] H. Wang, Y. Wu, J. Xu, and I. Yolou, "Random graph matching in geometric models: the case of complete graphs," in *Conference on Learning Theory (COLT)*, 2022.

[15] A. Andoni, D. Hsu, K. Shi, and X. Sun, "Correspondence retrieval," in *Conference on Learning Theory*. PMLR, 2017, pp. 105–126.

[16] P. Rigollet and J. Weed, "Uncoupled isotonic regression via minimum wasserstein deconvolution," *Information and Inference: A Journal of the IMA*, vol. 8, no. 4, pp. 691–717, 2019.

[17] A. Pananjady and R. J. Samworth, "Isotonic regression with unknown permutations: Statistics, computation and adaptation," *The Annals of Statistics*, vol. 50, no. 1, pp. 324–350, 2022.

[18] B. Baldessari, "The distribution of a quadratic form of normal random variables," *The Annals of Mathematical Statistics*, vol. 38, no. 6, pp. 1700–1704, 1967.

[19] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[20] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices and Banach spaces," in *Handbook of the geometry of Banach spaces, Vol. I*. North-Holland, Amsterdam, 2001, pp. 317–366.

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