Linear Regression with Mismatched Data: A Provably Optimal Local Search Algorithm

Rahul Mazumder(✉) and Haoyue Wang

Massachusetts Institute of Technology, Cambridge, MA 02139, USA
{rahulmaz,haoyuew}@mit.edu

Abstract. Linear regression is a fundamental modeling tool in statistics and related fields. In this paper, we study an important variant of linear regression in which the predictor-response pairs are partially mismatched. We use an optimization formulation to simultaneously learn the underlying regression coefficients and the permutation corresponding to the mismatches. The combinatorial structure of the problem leads to computational challenges, and we are unaware of any algorithm for this problem with both theoretical guarantees and appealing computational performance. To this end, in this paper, we propose and study a simple greedy local search algorithm. We prove that under a suitable scaling of the number of mismatched pairs compared to the number of samples and features, and certain assumptions on the covariates; our local search algorithm converges to the global optimal solution with a linear convergence rate under the noiseless setting.

Keywords: Linear regression · Mismatched data · Local search method · Learning permutations

1 Introduction

Linear regression and its extensions are among the most useful models in statistics and related fields. In the classical and most common setting, we are given n samples with features $x_i \in \mathbb{R}^d$ and response $y_i \in \mathbb{R}$, where i denotes the sample indices. We assume that the features and responses are perfectly matched i.e., x_i and y_i correspond to the same record or sample. An interesting twist to this problem—also the focus of this paper—is when the feature-response pairs are partially mismatched due to errors in the data merging process [6,7,10]. Here, we consider a mismatched linear model with responses $y = [y_1, ..., y_n] \in \mathbb{R}^n$ and covariates $X = [x_1, ..., x_n]^\top \in \mathbb{R}^{n \times d}$ satisfying

$$P^* y = X\beta^* + \epsilon \tag{1.1}$$

Supported by grants from the Office of Naval Research: ONR-N000141812298 (YIP) and National Science Foundation: NSF-IIS-1718258.

M. Singh and D. P. Williamson (Eds.): IPCO 2021, LNCS 12707, pp. 443–457, 2021.
https://doi.org/10.1007/978-3-030-73879-2_31

where $\beta^* \in \mathbb{R}^d$ are the regression coefficients, $\epsilon = [\epsilon_1, ..., \epsilon_n]^\top \in \mathbb{R}^n$ is the noise term, and $P^* \in \mathbb{R}^{n\times n}$ is an unknown permutation matrix. We consider the setting where $n > d$ and X has full rank; and seek to estimate both β^* and P^* based on the n observations $\{(y_i, x_i)\}_1^n$. The main computational difficulty arises in learning the unknown permutation. Linear regression with mismatched/permuted data (model (1.1)) has a long history in statistics dating back to 1960s [6].

Recently, this problem has garnered significant attention from the statistics and machine learning communities. A series of recent works [1–5,7–12,14] have studied the statistical and computational aspects of this model. To learn the coefficients β^* and the matrix P^*, one can consider the following natural optimization problem:

$$\min_{\beta, P} \ \|Py - X\beta\|^2 \quad \text{s.t. } P \in \Pi_n \tag{1.2}$$

where Π_n is the set of permutation matrices in $\mathbb{R}^{n\times n}$. Solving problem (1.2) is difficult as there are combinatorially many choices for $P \in \Pi_n$. Given P, it is easy to estimate β via least squares. [12] shows that in the noiseless setting ($\epsilon = 0$), a solution $(\hat{P}, \hat{\beta})$ of problem (1.2) equals (P^*, β^*) with probability one if $n \geq 2d$ and the entries of X are i.i.d. from a distribution that is absolutely continuous with respect to the Lebesgue measure. [5,8] studies the recovery of (P^*, β^*) under the noisy setting.

It is shown in [8] that Problem (1.2) is NP-hard for $d \geq 2$. A polynomial-time approximation algorithm appears in [5] for a fixed d, though this does not appear to result in a practical algorithm. Several heuristics have been proposed for (1.2): Examples include, alternating minimization [4,14], Expectation Maximization [1] but they lack theoretical guarantees. [10] uses robust regression methods to approximate solutions to (1.2) and discuss statistical properties of the corresponding estimator when the number of mismatched pairs is small.

Problem (1.2) can be formulated as a mixed integer program (MIP) with $O(n^2)$ binary variables. Solving this MIP with off-the-shelf MIP solvers (e.g., Gurobi) becomes computationally expensive for even a small value of n (e.g. $n \approx 50$). To our knowledge, there is no computationally practical algorithm that *provably* solves the original problem (1.2) under suitable statistical assumptions. Addressing this gap is the main focus of this paper: We propose and study a novel greedy local search method for Problem (1.2). Loosely speaking, our algorithm at every step swaps a pair of indices in the permutation in an attempt to improve the cost function. This algorithm is typically efficient in practice based on our preliminary numerical experiments. Suppose r denotes the number of mismatched pairs i.e., the Hamming distance between P^* and the identity matrix I_n. We establish theoretical guarantees on the convergence of the proposed method, under the assumption that r is small compared to n (we make this notion precise later), an assumption appearing in [10] (see also references therein). We consider the noiseless setting (i.e., $\epsilon = 0$) and establish that under some assumptions on the problem data, our local search method converges to an optimal solution of Problem (1.2).

Notation and Preliminaries: For a vector a, we let $\|a\|$ denote the Euclidean norm, $\|a\|_\infty$ the ℓ_∞-norm and $\|a\|_0$ the ℓ_0-pseudo-norm (i.e., number of nonzeros) of a. We let $|||\cdot|||_2$ denote the operator norm for matrices. Let $\{e_1, ..., e_n\}$ be the natural orthogonal basis of $\mathbb{R}^n$. For an interval $[a, b] \subseteq \mathbb{R}$, we let $|[a, b]| = b - a$. For a finite set S, we let $\#S$ denote its cardinality. For any permutation matrix P, let π_P be the corresponding permutation of $\{1, 2, ..., n\}$, that is, $\pi_P(i) = j$ if and only if $e_i^\top P = e_j^\top$ if and only if $P_{ij} = 1$. For two different permutation matrices P and Q, we define the distance between them as

$$\mathsf{dist}(P, Q) = \#\{i \in [n] : \pi_P(i) \neq \pi_Q(i)\}. \tag{1.3}$$

For any permutation matrix $P \in \Pi_n$, we define its support as:

$$\mathsf{supp}(P) := \{i \in [n] : \ \pi_P(i) \neq i\}. \tag{1.4}$$

For a real symmetric matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest eigenvalues of A, respectively.

For two positive scalar sequences $\{a_n\}, \{b_n\}$, we write $a_n = \widetilde{O}(b_n)$ or equivalently, $a_n/b_n = \widetilde{O}(1)$, if a_n/b_n is bounded by a polynomial (of finite degree) in $\log(n)$. In particular, we view any value that can be bounded by a polynomial of $\log(n)$ as a constant.

2 A Local Search Method

Here we present our local search method for (1.2). For any fixed $P \in \Pi_n$, by minimizing the objective function in (1.2) with respect to β, we have an equivalent formulation

$$\min_P \ \|Py - HPy\|^2 \ \text{ s.t. } \ P \in \Pi_n \tag{2.1}$$

where $H = X(X^\top X)^{-1}X^\top$. To simplify the notation, denote $\widetilde{H} := I_n - H$, then Problem (2.1) is equivalent to

$$\min_P \ \|\widetilde{H}Py\|^2 \ \text{ s.t. } P \in \Pi_n \ . \tag{2.2}$$

For a given permutation matrix P, define the R-neighbourhood of P as

$$\mathcal{N}_R(P) := \Big\{Q \in \Pi_n : \ \mathsf{dist}(P, Q) \leq R\Big\}. \tag{2.3}$$

It is easy to check that $\mathcal{N}_1(P) = \{P\}$, and for any $R \geq 2$, $\mathcal{N}_R(P)$ has more than one element. Algorithm 1 introduces our proposed local search method with search width R, which is an upper bound on the number of mismatched pairs.

Algorithm 1. Local search method with search width R for Problem (2.2).

Input: Initial permutation $P^{(0)} = I_n$. Search width R.
For $k = 0, 1, 2, ...$

$$P^{(k+1)} \in \operatorname{argmin} \left\{ \|\widetilde{H} P y\|^2 : \ P \in \mathcal{N}_2(P^{(k)}) \cap \mathcal{N}_R(I_n) \right\}. \tag{2.4}$$

If $\|\widetilde{H} P^{(k+1)} y\|^2 = \|\widetilde{H} P^{(k)} y\|^2$, output $P^{(k)}$.

Algorithm 1 uses an explicit constraint on the search width at every step. When $R \geq n$, we perform local search without any constraint on the search width or neighborhood size. In this paper, we focus on the case where the underlying P^* is close to I_n, i.e., $r \ll n$. Under this assumption, it is reasonable to set $R = cr \ll n$ for some constant $c > 1$. See Sect. 3 for more details.

Let us examine the per-iteration cost of (2.4). The cardinality of $\mathcal{N}_2(P^{(k)})$ is upper bounded by $O(n^2)$. Furthermore, we note that

$$\begin{aligned} &\|\widetilde{H} P y\|^2 = \|\widetilde{H}(P - P^{(k)})y + \widetilde{H} P^{(k)} y\|^2 \\ &= \|\widetilde{H}(P - P^{(k)})y\|^2 + 2\langle (P - P^{(k)})y, \widetilde{H} P^{(k)} y \rangle + \|\widetilde{H} P^{(k)} y\|^2 \ . \end{aligned} \tag{2.5}$$

For each $P \in \mathcal{N}_2(P^{(k)})$, the vector $(P - P^{(k)})y$ has at most two nonzero entries. So the computation of the first term in (2.5) costs $O(1)$ operations. As we retain a copy of $\widetilde{H} P^{(k)} y$ in memory, computing the second term in (2.5) also costs $O(1)$ operations. Therefore, computing (2.4) using the procedure outlined above requires $O(n^2)$ operations.

3 Theoretical Guarantees for Local Search

In this section, we present theoretical guarantees for Algorithm 1. Our theory is based on the assumption that the data X is "well-behaved" (See Assumption 1). In particular, we assume that the projection matrix $\widetilde{H}$ satisfies a "restricted eigenvalue condition" (RE). (We caution the reader that despite nomenclature similarities, our notion of RE is different than what appears in the high-dimensional statistics literature [13]). To give an example, our RE condition is satisfied with high probability, when the rows of X are independent draws from a well-behaved multivariate distribution and when the sample size n is sufficiently large—see Sect. 3.1 for details. Under this RE condition, our analysis is completely deterministic in nature. The RE assumption on $\widetilde{H}$ allows us to relate the objective function $\|\widetilde{H} P^{(k)} y\|^2$ to a simple function $\|(P^{(k)} - P^*)y\|^2$. Then our analysis reduces to an analysis of the local structure of Π_n in terms of minimizing $\|(P^{(k)} - P^*)y\|^2$.

3.1 A Restricted Eigenvalue (RE) Condition

A main building block of our analysis is a RE property of $\widetilde{H}$. Define

$$\mathcal{B}_m := \{ w \in \mathbb{R}^n : \ \|w\|_0 \leq m \}. \tag{3.1}$$

We say that $\widetilde{H}$ satisfies a RE condition with parameter (δ, m) (denoted by the shorthand $\mathrm{RE}(\delta, m)$) if the following holds true

$$\mathrm{RE}(\delta, m): \qquad \|\widetilde{H}u\|^2 \geq (1-\delta)\|u\|^2 \ \ \forall\ u \in \mathcal{B}_m. \tag{3.2}$$

To provide some intuition on the RE condition, we show (cf Lemma 1) that this condition is satisfied with high probability when the rows of X are drawn independently from a mean-zero distribution with finite support and a well-conditioned covariance matrix.

Lemma 1. *(Restricted eigenvalue property) Suppose $x_1, \ldots, x_n$ are i.i.d. zero-mean random vectors in $\mathbb{R}^d$ with covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Suppose there exist constants $\gamma, b, V > 0$ such that $\lambda_{\min}(\Sigma) \geq \gamma$, $\|x_i\| \leq b$ and $\|x_i\|_\infty \leq V$ almost surely. Given any $\tau > 0$, define*

$$\delta_n := 16V^2\Big(\frac{d}{n\gamma}\log(2d/\tau) + \frac{dm}{n\gamma}\log(3n^2)\Big)\ .$$

Suppose n is large enough such that $\sqrt{\delta_n} \geq 2/n$ and $\sqrt{3b^2\log(2d/\tau)/(n\,|||\,\Sigma|||_2)} \leq 1/2$. Then with probability at least $1-2\tau$, condition $RE(\delta_n, m)$ holds true.

The proof of this lemma is presented in Appendix 5.1. For simplicity, we state Lemma 1 for bounded x_i's; though this can be generalized to sub-Gaussian x_i's.

Lemma 1 implies that: given a pre-specified probability level (e.g., $1-2\tau = 0.99$), RE parameters δ, m, and other data parameters d, b, γ, Σ, we can choose $n = \widetilde{O}(dm/\delta)$ such that $\mathrm{RE}(\delta, m)$ holds with high probability. In the following, while presenting the scaling of (n, d, r) in the guarantees for Algorithm 1, when we say that data is generated from the setting of Lemma 1, we make the default assumption that there exist universal constants $\bar{c} > 0$ and $\bar{C} > 0$ such that the parameters $(\gamma, V, b, |||\Sigma|||_2, \tau)$ in Lemma 1 satisfy $\bar{c} \leq \gamma, V, b, |||\Sigma|||_2, \tau \leq \bar{C}$.

In Algorithm 1, we use a constraint on the search width, i.e., $P^{(k)} \in \mathcal{N}_R(I_n)$. Suppose $r = \mathsf{dist}(P^*, I_n) \ll n$ and we set $R = cr$ for some constant $c > 1$, then it holds that $\mathsf{dist}(P^{(k)}, P^*) \leq (c+1)r$. This implies $P^{(k)}y - P^*y \in \mathcal{B}_{(c+1)r}$. In the noiseless setting with $\epsilon = 0$, we have $\widetilde{H}P^*y = 0$, and hence $\|\widetilde{H}P^{(k)}y\|^2 = \|\widetilde{H}(P^{(k)} - P^*)y\|^2$. Suppose the $\mathrm{RE}(\delta_n, (c+1)r)$ condition in (3.2) holds, because $P^{(k)}y - P^*y \in \mathcal{B}_{(c+1)r}$, we have

$$(1-\delta_n)\|(P^{(k)} - P^*)y\|^2 \leq \|\widetilde{H}P^{(k)}y\|^2 \leq \|(P^{(k)} - P^*)y\|^2 \tag{3.3}$$

where, the second inequality is because $|||\widetilde{H}|||_2 \leq 1$. In light of (3.3), when δ_n is small, the objective function $\|\widetilde{H}P^{(k)}y\|^2$ can be approximately replaced by a simpler function $\|(P^{(k)} - P^*)y\|^2$. In what follows, we analyze the local search method on this simple approximation.

3.2 One-Step Decrease

We prove elementary lemmas on the one-step decrease property. Recall that for a given permutation matrix P, $\mathsf{supp}(P) = \{i \in [n]:\ \pi_P(i) \neq i\}$.

Lemma 2. *Given $y \in \mathbb{R}^n$ and a permutation matrix $P \in \Pi_n$, there exists a permutation matrix $\widetilde{P} \in \Pi_n$ such that $\mathsf{dist}(P, \widetilde{P}) = 2$, $\mathsf{supp}(\widetilde{P}) \subseteq \mathsf{supp}(P)$ and*

$$\|Py - y\|^2 - \|\widetilde{P}y - y\|^2 \geq (1/2)\|Py - y\|_\infty^2 \ .$$

The proof of Lemma 2 is presented in Sect. 5.2. Applying Lemma 2 with y replaced by P^*y and P replaced by $P(P^*)^{-1}$, we have the following corollary.

Corollary 1. *Given $y \in \mathbb{R}^n$ and $P, P^* \in \Pi_n$, there exists a permutation matrix $\widetilde{P} \in \Pi_n$ such that $\mathsf{dist}(\widetilde{P}, P) = 2$, $\mathsf{supp}(\widetilde{P}(P^*)^{-1}) \subseteq \mathsf{supp}(P(P^*)^{-1})$ and*

$$\|Py - P^*y\|^2 - \|\widetilde{P}y - P^*y\|^2 \geq (1/2)\|Py - P^*y\|_\infty^2 \ .$$

Corollary 1 provides a lower bound on the change of the (approximate) objective value as one moves from permutation P to $\widetilde{P}$, and will be used in the analysis of the local search algorithm. When $Py - P^*y$ is sparse, Corollary 1 translates to a contraction in the ℓ_2-norm of $Py - P^*y$, as shown below.

Corollary 2. *Let $y \in \mathbb{R}^n$ and $P, P^* \in \Pi_n$; and suppose $\|Py - P^*y\|_0 \leq m$. Let $\widetilde{P} \in \Pi_n$ be the permutation matrix appearing in Corollary 1. Then*

$$\|\widetilde{P}y - P^*y\|^2 \leq (1 - 1/(2m))\, \|Py - P^*y\|^2 \ . \tag{3.4}$$

Proof. Since $\|Py - P^*y\|_0 \leq m$, it holds $\|Py - P^*y\|^2 \leq m\|Py - P^*y\|_\infty^2$. Using Corollary 1, we have:

$$\|Py - P^*y\|^2 - \|\widetilde{P}y - P^*y\|^2 \geq (1/2)\|Py - P^*y\|_\infty^2 \geq (1/(2m))\|Py - P^*y\|^2 \ ,$$

which results in the conclusion (3.4). □

3.3 Main Results

Here we state and prove the main theorem on the convergence of Algorithm 1. We first state the assumptions used in our proof. Recall that $r = \mathsf{dist}(P^*, I_n)$.

Assumption 1. *(1.) We consider a linear model (1.1) with noise term $\epsilon = 0$.*
(2). There exist constants $U > L > 0$ such that $U \geq |(P^*y)_i - y_i| \geq L$ for all $i \in \mathsf{supp}(P^*)$.
(3). In Algorithm 1, we set $R = 10C_1 rU^2/L^2 + 4$ for some constant $C_1 > 1$.
(4). For some $\delta_n < 1/(4(r + R))$ the condition $RE(\delta_n, R + r)$ holds.

Note that the lower bound in Assumption 1 (2) ensures that any two mismatched responses are not too close. Assumption 1 (3) requires that R be set to a constant multiple of r. This constant can be large ($\geq 10U^2/L^2$), and is an artifact of our proof techniques. Our numerical experience however, suggests that this constant can be much smaller in practice. Assumption 1 (4) is a restricted eigenvalue condition. This property holds true under the settings stated in Lemma 1 when $n \geq Cdr^2$ for some constant $C > 0$.

We first present a technical result used in the proof of Theorem 1.

Lemma 3. *Suppose Assumption 1 holds. Let* $\{P^{(k)}\}_k$ *be the permutation matrices generated by Algorithm 1. Suppose* $\|P^{(k)}y - P^*y\|_\infty \geq L$ *for some* $k \geq 1$*. If for all* $t \leq k-1$*, at least one of the two conditions holds: (i)* $t \leq R/2 - 1$*; or (ii)* $\mathsf{supp}(P^*) \subseteq \mathsf{supp}(P^{(t)})$*, then for all* $t \leq k-1$*, we have*

$$\|P^{(t+1)}y - P^*y\|^2 - \|P^{(t)}y - P^*y\|^2 \leq -L^2/5 \ . \tag{3.5}$$

We omit the proof of Lemma 3 due to space constraints. Lemma 3 is used for technical reasons. In our analysis, we make heavy use of the one-step decrease condition in Corollary 2. Note that if the permutation matrix at the current iteration, denoted by $P^{(k)}$, is on the boundary i.e. $\mathsf{dist}(P^{(k)}, I_n) = R$, it is not clear whether the permutation found by Corollary 2 is within the search region $\mathcal{N}_R(I_n)$. Lemma 3 helps address this issue (See the proof of Theorem 1 below for details).

We now state and prove the linear convergence of Algorithm 1.

Theorem 1. *Suppose Assumption 1 holds with* R *being an even number. Let* $\{P^{(k)}\}$ *be the permutation matrices generated by Algorithm 1. Then*

1) For all $k \geq R/2$*, we have that* $\mathsf{supp}(P^*) \subseteq \mathsf{supp}(P^{(k)})$*.*
2) For any $k \geq 0$*,*

$$\|\widetilde{H}P^{(k)}y\|^2 \leq \left(1 - \frac{1}{4(R+r)}\right)^k \|\widetilde{H}P^{(0)}y\|^2 \ .$$

Proof. Part 1) We show this result by contradiction. Suppose that there exists a $k \geq R/2$ such that $\mathsf{supp}(P^*) \not\subseteq \mathsf{supp}(P^{(k)})$. Let $T \geq R/2$ be the first iteration such that $\mathsf{supp}(P^*) \not\subseteq \mathsf{supp}(P^{(T)})$, i.e.,

$$\mathsf{supp}(P^*) \not\subseteq \mathsf{supp}(P^{(T)}) \quad \text{and} \quad \mathsf{supp}(P^*) \subseteq \mathsf{supp}(P^{(k)}) \ \ \forall \ R/2 \leq k \leq T-1 \ .$$

Let $i \in \mathsf{supp}(P^*)$ but $i \notin \mathsf{supp}(P^{(T)})$, then by Assumption 1 (2),

$$\|P^{(T)}y - P^*y\|_\infty \geq |e_i^\top (P^{(T)}y - P^*y)| = |e_i^\top (y - P^*y)| \geq L.$$

By Lemma 3, we have $\|P^{(k+1)}y - P^*y\|^2 - \|P^{(k)}y - P^*y\|^2 \leq -L^2/5$ for all $k \leq T-1$. Summing up these inequalities, we have

$$\|P^{(T)}y - P^*y\|^2 - \|P^{(0)}y - P^*y\|^2 \leq -TL^2/5 \leq -RL^2/10 \tag{3.6}$$

where the last inequality follows from our assumption that $T \geq R/2$. From Assumption 1 (2) and noting that $P^{(0)} = I_n$, we have

$$\|P^{(0)}y - P^*y\|^2 = \|y - P^*y\|^2 \leq rU^2, \tag{3.7}$$

where we use that $(y - P^*y)$ is r-sparse. Using (3.6) and (3.7), we have:

$$\|P^{(T)}y - P^*y\|^2 \leq rU^2 - RL^2/10 \overset{(a)}{\leq} rU^2 - \frac{L^2}{10}\frac{10C_1rU^2}{L^2} = (1-C_1)U^2 \overset{(b)}{<} 0, \tag{3.8}$$

where above, the inequality (a) uses $R = 10C_1 rU^2/L^2 + 4$; and (b) uses $C_1 > 1$. Note that (3.8) leads to a contradiction, so such an iteration counter T does not exist; and for all $k \geq R/2$, we have $\mathsf{supp}(P^*) \subseteq \mathsf{supp}(P^{(k)})$.

Part 2) By Corollary 2, there exists a permutation matrix $\widetilde{P}^{(k)} \in \Pi_n$ such that $\mathsf{dist}(\widetilde{P}^{(k)}, P^{(k)}) \leq 2$, $\mathsf{supp}(\widetilde{P}^{(k)}(P^*)^{-1}) \subseteq \mathsf{supp}(P^{(k)}(P^*)^{-1})$ and

$$\|\widetilde{P}^{(k)}y - P^*y\|^2 \leq \left(1 - \frac{1}{2\|P^{(k)}y - P^*y\|_0}\right) \|P^{(k)}y - P^*y\|^2 \ .$$

Since $\|P^{(k)}y - P^*y\|_0 \leq \mathsf{dist}(P^{(k)}, I_n) + \mathsf{dist}(P^*, I_n) \leq r + R$, we have

$$\|\widetilde{P}^{(k)}y - P^*y\|^2 \leq \left(1 - \frac{1}{2(R+r)}\right) \|P^{(k)}y - P^*y\|^2. \tag{3.9}$$

Note that $\widetilde{H}P^*y = \widetilde{H}X\beta^* = 0$ and $|||\widetilde{H}|||_2 \leq 1$, so we have

$$\|\widetilde{H}\widetilde{P}^{(k)}y\|^2 = \|\widetilde{H}(\widetilde{P}^{(k)}y - P^*y)\|^2 \leq \|\widetilde{P}^{(k)}y - P^*y\|^2 \ . \tag{3.10}$$

In the following, we use the shorthand notation $\widetilde{R} = R + r$. Combining (3.9) and (3.10) we have

$$\|\widetilde{H}\widetilde{P}^{(k)}y\|^2 \leq \|\widetilde{P}^{(k)}y - P^*y\|^2 \leq (1 - (2\widetilde{R})^{-1})\|P^{(k)}y - P^*y\|^2. \tag{3.11}$$

By Assumption 1 (4), we have $\|\widetilde{H}(P^{(k)} - P^*)y\|^2 \geq (1 - \delta_n)\|(P^{(k)} - P^*)y\|^2$. Combining this with (3.11) we have

$$\begin{aligned} \|\widetilde{H}\widetilde{P}^{(k)}y\|^2 &\leq (1-\delta_n)^{-1}(1 - (2\widetilde{R})^{-1})\|\widetilde{H}(P^{(k)} - P^*)y\|^2 \\ &= (1-\delta_n)^{-1}(1 - (2\widetilde{R})^{-1})\|\widetilde{H}P^{(k)}y\|^2, \end{aligned} \tag{3.12}$$

where the last line uses $\widetilde{H}P^*y = 0$. Since $\delta_n \leq 1/(4\widetilde{R})$, we have

$$(1-\delta_n)^{-1}(1 - (2\widetilde{R})^{-1}) \leq 1 - (4\widetilde{R})^{-1}$$

which when used in (3.12) leads to:

$$\|\widetilde{H}\widetilde{P}^{(k)}y\|^2 \leq (1 - (4\widetilde{R})^{-1})\|\widetilde{H}P^{(k)}y\|^2. \tag{3.13}$$

To complete the proof, we will make use of the following claim, the proof of this claim is presented in Appendix 5.3.

Claim. For any $k \geq 0$ it holds $\widetilde{P}^{(k)} \in \mathcal{N}_R(I_n) \cap \mathcal{N}_2(P^{(k)})$. (3.14)

Starting with the definition of $P^{(k+1)}$, we have the following inequalities:

$$\begin{aligned} \|\widetilde{H}P^{(k+1)}y\|^2 = \min_{P \in \mathcal{N}_2(P^{(k)}) \cap \mathcal{N}_R(I_n)} \|\widetilde{H}Py\|^2 &\overset{(a)}{\leq} \|\widetilde{H}\widetilde{P}^{(k)}y\|^2 \\ &\overset{(b)}{\leq} (1 - (4\widetilde{R})^{-1})\|\widetilde{H}P^{(k)}y\|^2, \end{aligned}$$

where, (a) makes use of the above claim $\widetilde{P}^{(k)} \in \mathcal{N}_R(I_n) \cap \mathcal{N}_2(P^{(k)})$; and (b) uses inequality (3.13). Therefore, we have:

$$\|\widetilde{H}P^{(k+1)}y\|^2 \leq (1 - (4\widetilde{R})^{-1})\|\widetilde{H}P^{(k)}y\|^2,$$

which leads to the conclusion in part 2. □

Theorem 1 shows that the sequence of objective values generated by Algorithm 1 converges to zero (the optimal objective value of (2.2)) at a linear rate. The parameter for the linear rate of convergence depends upon $r = \mathsf{dist}(P^*, I_n)$ and the search width R. The proof is based on the assumption that the RE condition holds (Assumption 1 (4)) with some $\delta_n \leq 1/(4(R+r))$. This RE condition holds under the setting of Lemma 1 when $n \geq Cdr^2$ for some constant $C > 0$ (See Sect. 3.1). The sample-size requirement is more stringent than that needed in order for the model to be identifiable ($n \geq 2d$) [12]. In particular, when $n/d = \widetilde{O}(1)$, the number of mismatched pairs r needs to be bounded by a constant. While our theory appears to suggest that n needs to be quite large to learn P^*, numerical evidence presented in Sect. 4 suggests that one can recover P^* with a smaller sample size.

4 Experiments

We numerically study the convergence performance of Algorithm 1. We consider the noiseless setup $P^*y = X\beta^*$ where entries of $X \in \mathbb{R}^{n\times d}$ are iid $N(0,1)$; all coordinates of $\beta^* \in \mathbb{R}^d$ are iid $N(0,1)$ (β^* is independent of X). To generate P^*, we fix $r \geq 1$ and select r coordinates uniformly from $\{1,\ldots,n\}$, then generate a uniformly distributed random permutation on these r coordinates[1].

We test the performance of Algorithm 1 with different combinations of (d, r, n). We simply set $R = n$ in Algorithm 1. Even though this setting is not covered by our theory, in practice when r is small, the algorithm converges to optimality with the number of iterations being bounded by a small constant multiple of r (e.g., for $r = 50$, the algorithm converges to optimality within around 60 iterations). We set the maximum number of iterations as 1000. For the results presented below, we consider 50 independent trials and present the averaged results.

Figure 1 presents the results on examples with $n = 500$, $d \in \{20, 50, 100, 200\}$, and 40 roughly equispaced values of $r \in [10, 400]$. In Fig. 1 [left panel], we plot the Hamming distance of the solution $\hat{P}$ computed by Algorithm 1 and the underlying permutation P^* (i.e. $\mathsf{dist}(\hat{P}, P^*)$) versus r. In Fig. 1 [right panel], we present error in estimating β versus r. More precisely, let $\hat{\beta}$ be the solution of computed by Algorithm 1 (i.e. $\hat{\beta} = (X^\top X)^{-1}X^\top \hat{P}y$), then the *"beta error"* is defined as $\|\hat{\beta} - \beta^*\|/\|\beta^*\|$. For each choice of (r, d), the point on the line is the average of 50 independent replications, and the vertical error bar shows the

[1] This permutation P^* may not satisfy $\mathsf{dist}(P^*, I_n) = r$, but $\mathsf{dist}(P^*, I_n)$ will be close to r.

standard deviation of the mean (the error bars are small and hardly visible in the figures). As shown in Fig. 1, when r is small, the underlying permutation P^* can be exactly recovered, and thus the corresponding beta error is also 0. As r becomes larger, Algorithm 1 fails to recover P^* exactly; and $\mathsf{dist}(P^*, \hat{P})$ is close to the maximal possible value 500. In contrast, the estimation error of β^* behaves in a continuous way: As the value of r increases, the value of $\|\hat{\beta} - \beta^*\|/\|\beta^*\|$ increases continuously. We also observe that the recovery of P^* depends upon the number of covariates d. This is consistent with our analysis that the performance of our algorithm depends upon both r and d.

Figure 2 presents similar results where we exchange the roles of r and d. It shows examples with $n = 500$, $r \in \{20, 50, 100, 200\}$, and 40 different values of d ranging from 10 to 400. When d is small, Algorithm 1 is able to recover P^* exactly. But when d exceeds a certain threshold, $\mathsf{dist}(\hat{P}, P^*)$ increases quickly. The threshold for larger r is smaller. From Fig. 2 [left panel], it is interesting to note a non-monotone behavior of the Hamming distance as d increases. In contrast, the beta error increases continuously as d increases (see Fig. 2 [right panel]).

In terms of the speed of Algorithm 1, we note that for an instance with $n = 500$, $d = 100$ and $r = 50$, Algorithm 1 outputs the solution within around 60 iterations and 0.25 s on the Julia 1.2.0 platform. The total computational time scales approximately as $O(n^2 r)$ when exact recovery is achieved.

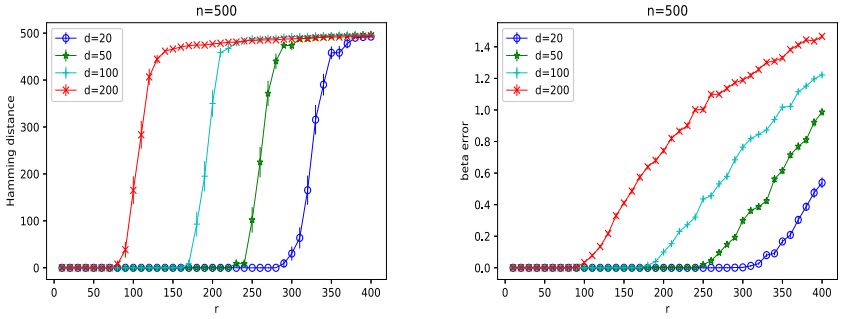

Fig. 1. Left: values hamming distance $\mathsf{dist}(\hat{P}, P^*)$ versus r. Right: values of beta error $\|\hat{\beta} - \beta^*\|/\|\beta^*\|$ versus r.

5 Appendix: Proofs and Technical Results

Lemma 4. *Suppose rows $x_1, ..., x_n$ of the matrix of covariates X are i.i.d. zero-mean random vectors in $\mathbb{R}^d$ with covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Suppose $\|x_i\| \le b$ almost surely. Then for any $t > 0$, it holds*

$$\mathbb{P}\Big(\|\!|\!| \frac{1}{n} X^\top X - \Sigma \|\!|\!|_2 \ge t \, \|\!|\!| \, \Sigma \, |\!|\!|_2 \Big) \le 2d \exp\Big(- \frac{nt^2 \, \|\!|\!| \, \Sigma \, |\!|\!|_2}{2b^2(1+t)} \Big).$$

See e.g. Corollary 6.20 of [13] for a proof.

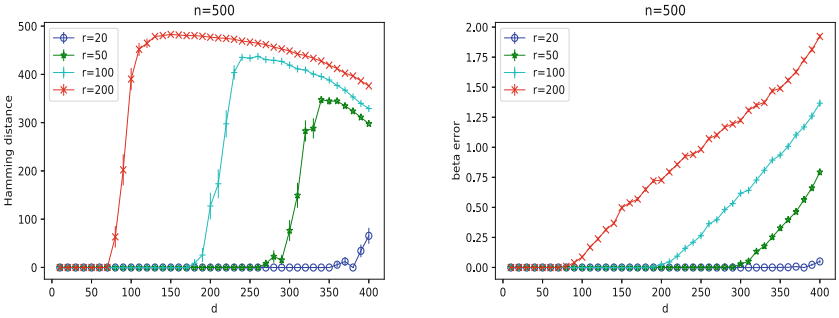

Fig. 2. Left: values of hamming distance $\mathsf{dist}(\hat{P}, P^*)$ vs r. Right: values of beta error $\|\hat{\beta} - \beta^*\|/\|\beta^*\|$ vs r.

5.1 Proof of Lemma 1

Proof. It suffices to prove that for any $u \in \mathcal{B}_m$ (cf definition (3.1)),

$$\|Hu\|^2 = \|X(X^\top X)^{-1}X^\top u\|^2 \le \delta_n \|u\|^2 \ . \tag{5.1}$$

Take $t_n := \sqrt{3b^2 \log(2d/\tau)/(n \,|||\, \Sigma \,|||_2)}$. When n is large enough, we have $t_n \le 1/2$, then from Lemma 4 and some simple algebra we have

$$||| \frac{1}{n} X^\top X - \Sigma |||_2 \le t_n \,|||\, \Sigma |||_2 \tag{5.2}$$

with probability at least $1 - \tau$. When (5.2) holds, we have

$$\lambda_{\min}(X^\top X)/n \ge (1 - t_n)\lambda_{\min}(\Sigma) \ge (1 - t_n)\gamma \ge \gamma/2$$

where, we use $t_n \le 1/2$. Hence we have $\lambda_{\max}((X^\top X)^{-1}) \le 2/(n\gamma)$ and

$$||| X(X^\top X)^{-1} |||_2 = \sqrt{\lambda_{\max}((X^\top X)^{-1})} \le \sqrt{2/(n\gamma)} \ . \tag{5.3}$$

Let $\mathcal{B}_m(1) := \{u \in \mathcal{B}_m : \|u\| \le 1\}$, and let $u^1, ..., u^M$ be an $(\sqrt{\delta_n}/2)$-net of $\mathcal{B}_m(1)$, that is, for any $u \in \mathcal{B}_m(1)$, there exists some u^j such that $\|u^j - u\| \le \sqrt{\delta_n}/2$. Since the $(\sqrt{\delta_n}/2)$-covering number of $\mathcal{B}_m(1)$ is bounded by $(6/\sqrt{\delta_n})^m \binom{n}{m}$, we can take

$$M \le (6/\sqrt{\delta_n})^m \binom{n}{m} \le (3n)^m n^m = (3n^2)^m$$

where the second inequality is from our assumption that $\sqrt{\delta_n} \ge 2/n$. By Hoeffding inequality, for each fixed $j \in [M]$, and for all $k \in [d]$, we have

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \left| e_k^\top X^\top u^j \right| > t \right) \le 2 \exp\left(-\frac{nt^2}{2\|u^j\|^2 U^2} \right) \ .$$

Therefore, for any $\rho > 0$, with probability at least $1 - \rho$, we have

$$\left|e_k^\top X^\top u^j\right| / \sqrt{n} \le \sqrt{2\log(2d/\rho)/n} V \|u^j\| \le V\sqrt{2\log(2d/\rho)/n} \ ,$$

where the second inequality is because each $u^j \in \mathcal{B}_m(1)$. As a result,

$$\frac{1}{\sqrt{n}}\|X^\top u^j\| = \Big(\sum_{k=1}^{d} \left(|e_k^\top X^\top u^j|/\sqrt{n}\right)^2\Big)^{1/2} \le V\sqrt{2d\log(2d/\rho)/n} \ .$$

Take $\rho = \tau/M$, then by the union bound, with probability at least $1-\tau$, it holds

$$\|X^\top u^j\|/\sqrt{n} \le V\sqrt{2d\log(2dM/\tau)/n} \quad \forall\ j \in [M] \ . \tag{5.4}$$

Combining (5.4) with (5.3), we have that for all $j \in [M]$,

$$\begin{aligned} \|X(X^\top X)^{-1}X^\top u^j\| \le& \||\ X(X^\top X)^{-1}\ \||_2 \cdot \|X^\top u^j\| \\ \le& 2V\sqrt{(d/n\gamma)\log(2dM/\tau)} \ . \end{aligned} \tag{5.5}$$

Recall that $M \le (3n^2)^m$, so we have

$$2V\sqrt{(d/n\gamma)\log(2dM/\tau)} \le 2V\Big(\frac{d}{n\gamma}\log(2d/\tau) + \frac{dm}{n\gamma}\log(3n^2)\Big)^{1/2} \le \frac{\sqrt{\delta_n}}{2}.$$

where the last inequality follows the definition of δ_n. Using the above bound in (5.5), we have

$$\|X(X^\top X)^{-1}X^\top u^j\| \le \sqrt{\delta_n}/2.$$

For any $u \in \mathcal{B}_m(1)$, there exists some $j \in [M]$ such that $\|u - u^j\| \le \sqrt{\delta_n}/2$, hence

$$\begin{aligned} \|X(X^\top X)^{-1}X^\top u\| \le& \|X(X^\top X)^{-1}X^\top u^j\| + \|X(X^\top X)^{-1}X^\top (u - u^j)\| \\ \le& \sqrt{\delta_n}/2 + \|u - u^j\|_2 \le \sqrt{\delta_n} \ . \end{aligned} \tag{5.6}$$

Since both (5.2) and (5.4) have failure probability of at most τ, we know that (5.6) holds with probability at least $1 - 2\tau$. This proves the conclusion for all $u \in \mathcal{B}_m(1)$. For a general $u \in \mathcal{B}_m$, $u/\|u\| \in \mathcal{B}_m(1)$, hence we have

$$\|Hu\| = \|X(X^\top X)^{-1}X^\top u\| \le \sqrt{\delta_n}\|u\|$$

which is equivalent to what we had set out to prove (5.1). □

5.2 Proof of Lemma 2

Proof. For any $k \in [n]$, let $k_+ := \pi_P(k)$. Let i be an index such that

$$\left(y_{i_+} - y_i\right)^2 = \|Py - y\|_\infty^2 \ .$$

Without loss of generality, we can assume $y_{i_+} > y_i$. Denote $i_0 = i$ and $i_1 = i_+$. By the structure of a permutation, there exists a cycle that

$$i_0 \xrightarrow{P} i_1 \xrightarrow{P} \cdots \xrightarrow{P} i_t \xrightarrow{P} \cdots \xrightarrow{P} i_S = i_0 \tag{5.7}$$

where $q_1 \xrightarrow{P} q_2$ means $q_2 = \pi_P(q_1)$. By moving from y_i to y_{i_+}, the first step in the cycle (5.7) "upcrosses" the value $(y_i + y_{i_+})/2$. Since the cycle (5.7) returns to i_0 finally, there must exist one step that "downcrosses" the value $(y_i + y_{i_+})/2$. In other words, there exists $j \in [n]$ with $(j, j_+) \neq (i, i_+)$ such that $y_{j_+} < y_j$ and $(y_i + y_{i_+})/2 \in [y_{j_+}, y_j]$. Define $\widetilde{P}$ as follows:

$$\pi_{\widetilde{P}}(i) = j_+, \quad \pi_{\widetilde{P}}(j) = i_+, \quad \pi_{\widetilde{P}}(k) = \pi_P(k) \ \ \forall k \neq i, j \ .$$

We immediately know $\mathsf{dist}(P, \widetilde{P}) = 2$ and $\mathsf{supp}(\widetilde{P}) \subseteq \mathsf{supp}(P)$. Since

$$y_{i_+} - y_i = \|Py - y\|_\infty \geq y_j - y_{j_+},$$

there are 3 cases depending upon the ordering of $y_i, y_{i_+}, y_j, y_{j_+}$. We consider these cases to arrive at the final inequality in Lemma 2.

Case 1: $(y_j \geq y_{i_+} \geq y_{j_+} \geq y_i)$ In this case, let $a = y_j - y_{i_+}$, $b = y_{i_+} - y_{j_+}$ and $c = y_{j_+} - y_i$. Then $a, b, c \geq 0$, and

$$\begin{aligned} \|Py - y\|^2 - \|\widetilde{P}y - y\|^2 &= (y_i - y_{i_+})^2 + (y_j - y_{j_+})^2 - (y_i - y_{j_+})^2 - (y_j - y_{i_+})^2 \\ &= (b + c)^2 + (a + b)^2 - a^2 - c^2 \\ &= 2b^2 + 2ab + 2bc \ . \end{aligned}$$

Since $(y_i + y_{i_+})/2 \in [y_{j_+}, y_j]$, we have

$$b = y_{i_+} - y_{j_+} \geq y_{i_+} - \frac{y_i + y_{i_+}}{2} = \frac{y_{i_+} - y_i}{2} \ ,$$

and hence

$$\|Py - y\|^2 - \|\widetilde{P}y - y\|^2 \geq 2b^2 \geq \frac{(y_{i_+} - y_i)^2}{2} = \frac{1}{2}\|Py - y\|_\infty^2 \ .$$

Case 2: $(y_{i_+} \geq y_j \geq y_i \geq y_{j_+})$. In this case, let $a = y_{i_+} - y_j$, $b = y_j - y_i$ and $c = y_i - y_{j_+}$. Then $a, b, c \geq 0$, and

$$\begin{aligned} \|Py - y\|^2 - \|\widetilde{P}y - y\|^2 &= (y_i - y_{i_+})^2 + (y_j - y_{j_+})^2 - (y_i - y_{j_+})^2 - (y_j - y_{i_+})^2 \\ &= (a + b)^2 + (b + c)^2 - a^2 - c^2 \\ &= 2b^2 + 2ab + 2bc \ . \end{aligned}$$

Since $(y_i + y_{i_+})/2 \in [y_{j_+}, y_j]$, we have

$$b = y_j - y_i \geq \frac{y_i + y_{i_+}}{2} - y_i = \frac{y_{i_+} - y_i}{2} \ ,$$

and hence

$$\|Py - y\|^2 - \|\widetilde{P}y - y\|^2 \geq 2b^2 \geq \frac{(y_{i_+} - y_i)^2}{2} = \frac{1}{2}\|Py - y\|_\infty^2 \ .$$

Case 3: $(y_{i_+} \geq y_j \geq y_{j_+} \geq y_i)$. In this case, let $a = y_{i_+} - y_j$, $b = y_j - y_{j_+}$ and $c = y_{j_+} - y_i$. Then $a, b, c \geq 0$, and

$$\begin{aligned}\|Py - y\|^2 - \|\widetilde{P}y - y\|^2 &= (y_i - y_{i_+})^2 + (y_j - y_{j_+})^2 - (y_i - y_{j_+})^2 - (y_j - y_{i_+})^2 \\ &= (a+b+c)^2 + b^2 - a^2 - c^2 \\ &= 2b^2 + 2ab + 2bc + 2ac \ .\end{aligned}$$

Note that $\|Py-y\|_\infty^2 = (y_i - y_{i_+})^2 = (a+b+c)^2$. Because $(y_i + y_{i_+})/2 \in [y_{j_+}, y_j]$, we know that $a \leq (a+b+c)/2$ and $c \leq (a+b+c)/2$. So we have

$$\|Py - y\|^2 - \|\widetilde{P}y - y\|^2 \geq w\|Py - y\|_\infty^2 \ ,$$

where

$$w := \min\left\{\frac{2b^2 + 2ab + 2bc + 2ac}{(a+b+c)^2} \ : \ a, b, c \geq 0; \ a, c \leq (a+b+c)/2\right\}.$$

This is equivalent to

$$\begin{aligned}w &= \min\left\{2b^2 + 2ab + 2bc + 2ac \ : \ a, b, c \geq 0; \ a, c \leq 1/2; \ a+b+c = 1\right\} \\ &= \min\left\{2b + 2ac \ : \ a, b, c \geq 0; \ a, c \leq 1/2; \ a+b+c = 1\right\} \\ &= \min\left\{2(1-a-c) + 2ac \ : \ a, c \geq 0; \ a, c \leq 1/2\right\} \\ &= \min\left\{2(1-a)(1-c) \ : \ a, c \geq 0; \ a, c \leq 1/2\right\} \\ &= 1/2\end{aligned}$$

□

5.3 Proof of Claim (3.14) in Theorem 1

Proof. To prove this claim, we just need to prove that $\widetilde{P}^{(k)} \in \mathcal{N}_R(I_n)$, i.e. $\mathsf{dist}(\widetilde{P}^{(k)}, I_n) \leq R$. If $k \leq R/2 - 1$, because $\mathsf{dist}(P^{(t+1)}, P^{(t)}) \leq 2$ for all $t \geq 0$ and $P^{(0)} = I_n$, we have $\mathsf{dist}(P^{(k)}, I_n) \leq 2k \leq R - 2$. Hence

$$\mathsf{dist}(\widetilde{P}^{(k)}, I_n) \leq \mathsf{dist}(\widetilde{P}^{(k)}, P^{(k)}) + \mathsf{dist}(P^{(k)}, I_n) \leq R \ .$$

We consider the case when $k \geq R/2$. By Part (1) of Theorem 1, it holds $\mathsf{supp}(P^*) \subseteq \mathsf{supp}(P^{(k)})$. We will show that $\mathsf{supp}(\widetilde{P}^{(k)}) \subseteq \mathsf{supp}(P^{(k)})$. Equivalently, we just need to show that for any $i \notin \mathsf{supp}(P^{(k)})$, we have $i \notin \mathsf{supp}(\widetilde{P}^{(k)})$. Let $i \notin \mathsf{supp}(P^{(k)})$, then $e_i^\top P^{(k)} = e_i^\top$. Since $\mathsf{supp}(P^*) \subseteq \mathsf{supp}(P^{(k)})$, we also have $e_i^\top P^* = e_i^\top$. So it holds $e_i^\top P^{(k)}(P^*)^{-1} = e_i^\top$ or equivalently $i \notin \mathsf{supp}(P^{(k)}(P^*)^{-1})$. Because $\mathsf{supp}(\widetilde{P}^{(k)}(P^*)^{-1}) \subseteq \mathsf{supp}(P^{(k)}(P^*)^{-1})$, we have $i \notin \mathsf{supp}(\widetilde{P}^{(k)}(P^*)^{-1})$, or equivalently $e_i^\top \widetilde{P}^{(k)}(P^*)^{-1} = e_i^\top$. This implies $e_i^\top \widetilde{P}^{(k)} = e_i^\top P^* = e_i^\top$, or equivalently, $i \notin \mathsf{supp}(\widetilde{P}^{(k)})$. □

References

1. Abid, A., Zou, J.: Stochastic EM for shuffled linear regression. arXiv preprint arXiv:1804.00681 (2018)
2. Dokmanić, I.: Permutations unlabeled beyond sampling unknown. IEEE Signal Process. Lett. **26**(6), 823–827 (2019)
3. Emiya, V., Bonnefoy, A., Daudet, L., Gribonval, R.: Compressed sensing with unknown sensor permutation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1040–1044. IEEE (2014)
4. Haghighatshoar, S., Caire, G.: Signal recovery from unlabeled samples. IEEE Trans. Signal Process. **66**(5), 1242–1257 (2017)
5. Hsu, D.J., Shi, K., Sun, X.: Linear regression without correspondence. In: Advances in Neural Information Processing Systems, pp. 1531–1540 (2017)
6. Neter, J., Maynes, E.S., Ramanathan, R.: The effect of mismatching on the measurement of response errors. J. Am. Stat. Assoc. **60**(312), 1005–1027 (1965)
7. Pananjady, A., Wainwright, M.J., Courtade, T.A.: Denoising linear models with permuted data. In: 2017 IEEE International Symposium on Information Theory (ISIT), pp. 446–450. IEEE (2017)
8. Pananjady, A., Wainwright, M.J., Courtade, T.A.: Linear regression with shuffled data: statistical and computational limits of permutation recovery. IEEE Trans. Inf. Theory **64**(5), 3286–3300 (2017)
9. Shi, X., Li, X., Cai, T.: Spherical regression under mismatch corruption with application to automated knowledge translation. J. Am. Stat. Assoc., 1–12 (2020)
10. Slawski, M., Ben-David, E., Li, P.: Two-stage approach to multivariate linear regression with sparsely mismatched data. J. Mach. Learn. Res. **21**(204), 1–42 (2020)
11. Tsakiris, M.C., Peng, L., Conca, A., Kneip, L., Shi, Y., Choi, H., et al.: An algebraic-geometric approach to shuffled linear regression. arXiv preprint arXiv:1810.05440 (2018)
12. Unnikrishnan, J., Haghighatshoar, S., Vetterli, M.: Unlabeled sensing with random linear measurements. IEEE Trans. Inf. Theory **64**(5), 3237–3253 (2018)
13. Wainwright, M.J.: High-Dimensional Statistics: A Non-asymptotic Viewpoint, vol. 48. Cambridge University Press, Cambridge (2019)
14. Wang, G., et al.: Signal amplitude estimation and detection from unlabeled binary quantized samples. IEEE Trans. Signal Process. **66**(16), 4291–4303 (2018)