Linear Regression with an Unknown Permutation: Statistical and Computational Limits

Ashwin Pananjady[†], Martin J. Wainwright[†,⋆], Thomas A. Courtade[†]
Department of Electrical Engineering and Computer Sciences[†]
Department of Statistics[⋆]
University of California, Berkeley
Email: {ashwinpm, wainwrig, courtade}@eecs.berkeley.edu

***Abstract*—Consider a noisy linear observation model with an unknown permutation, based on observing $y = \Pi^* A x^* + w$, where $x^* \in \mathbb{R}^d$ is an unknown vector, Π^* is an unknown $n \times n$ permutation matrix, and $w \in \mathbb{R}^n$ is additive Gaussian noise. We analyze the problem of permutation recovery in a random design setting in which the entries of the matrix A are drawn i.i.d. from a standard Gaussian distribution, and establish sharp conditions on the SNR, sample size n, and dimension d under which Π^* is exactly and approximately recoverable. On the computational front, we show that the maximum likelihood estimate of Π^* is NP-hard to compute, while also providing a polynomial time algorithm when $d = 1$.**

I. Introduction

Recovery of a vector based on noisy linear measurements is the classical problem of linear regression, and is arguably the most basic problem in statistical inference. A variant, the "errors-in-variables" model [1], allows for errors in the measurement matrix, but mainly in the form of additive or multiplicative noise [2]. In this paper, we study a form of errors-in-variables in which the measurement matrix is perturbed by an unknown permutation of its rows.

More concretely, we study an observation model of the form

$$y = \Pi^* A x^* + w, \tag{1}$$

where $x^* \in \mathbb{R}^d$ is an unknown vector, $A \in \mathbb{R}^{n \times d}$ is a measurement (or design) matrix, Π^* is an unknown $n \times n$ permutation matrix, and $w \in \mathbb{R}^n$ is observation noise. We refer to the setting where $w = 0$ as the *noiseless case*. As with linear regression, there are two settings of interest, corresponding to whether the design matrix is **(i)** deterministic (the fixed design case), or **(ii)** random (the random design case).

There are also two complementary problems of interest – recovery of the unknown Π^*, and recovery of the unknown x^*. In this paper, we focus on the former problem; the latter problem is also known as unlabelled sensing [3].

The observation model (1) is frequently encountered in scenarios where there is uncertainty in the order in which measurements are taken. An illustrative example is that of sampling in the presence of jitter [4], in which the uncertainty about the instants at which measurements are taken results in an unknown permutation of the measurements. A similar synchronization issue occurs in timing and molecular channels [5]. Here, identical molecular tokens are received at the receptor at different times, and their signatures are indistinguishable. The vectors of transmitted and received times correspond to the signal and the observations, respectively, where the latter is some permuted version of the former with additive noise.

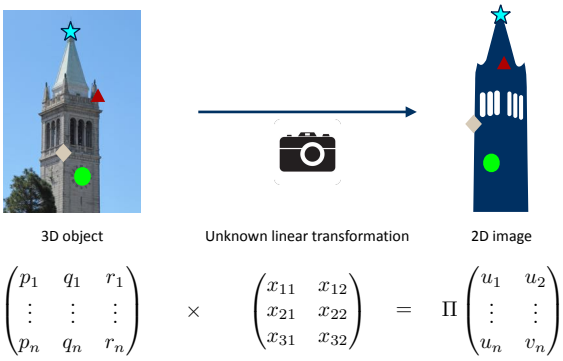
Another such scenario arises in multi-target tracking problems [6]. For example, SLAM tracking [7] is a classical problem in robotics where the environment in which measurements are made is unknown, and part of the problem is to infer relative permutations between measurements. Archaeological measurements [8] also suffer from an inherent lack of ordering, which makes inference of chronology hard. Another compelling example of such an observation model is in data anonymization, in which the order, or "labels", of measurements are intentionally deleted to preserve privacy. The inverse problem of data de-anonymization [9] is to infer these labels from the observations.

Also, in large sensor networks, it is often the case that the number of bits of information that each sensor records and transmits to the server is exceeded by the number of bits it transmits in order to identify itself to the server [10]. In applications where sensor measurements are linear, model (1) corresponds to the case where each sensor only sends its measurement but not its identity. The server is then tasked with recovering sensor identites, or equivalently, with determining the unknown permutation.

The pose and correspondence estimation problem in image processing [11], [12] is also related to the observation model (1). The capture of a 3D object by a 2D image can be modelled by an unknown linear transformation called the "pose", and an unknown permutation representing the "correspondence" between points in the two spaces. One of the central goals in image processing is to identify this correspondence information, which in this case is equivalent to permutation estimation in the linear model. An illustration of the problem is provided in Figure 1. Image stitching from multiple camera angles [13] also involves the resolution of unknown correspondence information between point clouds.

The discrete analog of the model (1) in which the vectors x^* and y, and the matrix A are all constrained to belong to some finite field corresponds to the permutation channel studied by Schulman and Zuckerman [14], with A representing the (linear) encoding matrix. However, techniques for the discrete problem do not carry over to the continuous problem (1).

Another line of work that is related in spirit to the observa-



$$\begin{pmatrix} p_1 & q_1 & r_1 \\ \vdots & \vdots & \vdots \\ p_n & q_n & r_n \end{pmatrix} \times \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} = \Pi \begin{pmatrix} u_1 & u_2 \\ \vdots & \vdots \\ u_n & v_n \end{pmatrix}$$

Fig. 1. Example of pose and correspondence estimation. The camera introduces an unknown linear transformation corresponding to the pose. The unknown permututation represents the correspondence between points, which is shown in the picture via coloured shapes, and needs to be estimated.

tion model (1) is the genome assembly problem from shotgun reads [15], in which an underlying vector $x^* \in \{A, T, G, C\}^d$ must be assembled from an unknown permutation of its continuous sub-vector measurements, called "reads". Two aspects, however, render it a particularization of our observation model, besides the fact that x^* in the genome assembly problem is constrained to a finite alphabet: (i) in genome assembly, the matrix A is fixed and consists of shifted identity matrices that select sub-vectors of x^*, and (ii) the permutation matrix of genome assembly is in fact a block permutation matrix that permutes sub-vectors instead of coordinates as in equation (1).

A. Related work

Previous work related to the observation model (1) can be broadly classified into two categories – those that focus on x^* recovery, and those focussed on recovering the underlying permutation. We discuss the most relevant results below.

1) Latent vector estimation: The observation model (1) appears in the context of compressed sensing with an unknown sensor permutation [16]. The authors consider the matrix-based observation model $Y = \Pi^* A X^* + W$, where X^* is a matrix whose columns are composed of multiple unknown vectors. Their contributions include a branch and bound algorithm to recover the underlying X^*, which they show to perform well empirically for small instances under the setting in which the entries of the matrix A are drawn i.i.d. from a Gaussian distribution.

For pose and correspondence estimation, the paper [12] considers the noiseless observation model (1), and shows that if the permutation matrix maps a sufficiently large number of positions to themselves, then x^* can be recovered reliably.

In the context of molecular channels, the model (1) has been analyzed for the case when x^* is some random vector, $A = I$, and w represents non-negative noise that models delays introduced between emitter and receptor. Rose et al. [5] provide lower bounds on the capacity of such channels. In particular, their results yield closed-form lower bounds for some noise distributions, e.g., exponentially random noise.

A more recent paper [3] that is most closely related to our model considers the question of when the equation (1) has a unique solution x^*, i.e., the identifiability of the noiseless model. The authors show that if the entries of A are sampled i.i.d. from any continuous distribution with $n \geq 2d$, then equation (1) has a unique solution x^* with probability 1. They also provide a converse showing that if $n < 2d$, any matrix A whose entries are sampled i.i.d. from a continuous distribution does not (with probability 1) have a unique solution x^* to equation (1). While the paper shows uniqueness, the question of designing an efficient algorithm to recover a solution, unique or not, is left open. The paper also analyzes the stability of the noiseless solution, and establishes that x^* can be recovered exactly when the SNR goes to infinity.

We also briefly compare the model (1) with the problem of vector recovery in unions of subspaces, studied widely in the compressive sensing literature [17], [18]. In the compressive sensing setup, the vector x^* lies in the union of finitely many subspaces, and must be recovered from linear measurements with a random matrix, without a permutation. In our model, on the other hand, the vector x^* is unrestricted, and the observation y lies in the union of $n!$ subspaces – one for each permutation. While the two models share a superficial connection, results do not carry over from one to the other in any obvious way. In fact, our model is fundamentally different from traditional compressive sensing, since the unknown permutation acts on the *row space* of the design matrix A. In contrast, restricting x^* to a union of subspaces (or restricting its sparsity) influences the column space of A.

2) Latent permutation estimation: While our paper seems to be the first to consider permutation recovery in the linear regression model (1), there are many related problems for which permutation recovery has been studied. We mention only those that are most closely related to our work.

The feature matching problem in machine learning [19] bears a superficial resemblance to our observation model. There, observations take the form $Y = X^* + W$ and $Y' = \Pi^* X^* + W'$, with (X^*, Y, Y', W, W') representing matrices of appropriate dimensions, and the goal is to recover Π^* from the tuple (Y, Y'). The paper [19] establishes minimax rates on the separation between the rows of X^* (as a function of the parameters n, d, σ) required for exact permutation recovery.

The problem of statistical seriation [20] involves an observation model of the form $Y = \Pi^* X^* + W$, with the matrix X^* obeying some shape constraint. In particular, if the columns of X^* are unimodal (or, as a special case, monotone), then Flammarion et al. [20] establish minimax rates for the problem in the prediction error metric $\|\widehat{\Pi}\widehat{X} - \Pi^* X^*\|_F^2$ by analyzing the least squares estimator. The seriation problem (without noise) was also considered by Fogel et al. [21] in the context of designing convex relaxations to permutation problems.

Permutation estimation has also been considered in other observation models involving matrices with structure, particularly in the context of ranking [22], [23], or even more generally, in the context of *identity management* [24]. While we mention both of these problems because are related in spirit

to permutation recovery, the problem setups do not bear too much resemblance to our linear model (1).

Algorithmic approaches to solving for Π^* in equation (1) are related to the multi-dimensional assignment problem. In particular, while finding the correct permutation mapping between two vectors minimizing some loss function between them corresponds to the 1-dimensional assignment problem, here we are faced with an assignment problem between subspaces. While we do not elaborate on the vast literature that exists on solving variants on assignment problems, we note that broadly speaking, assignment problems in higher dimensions are much harder than the 1-D assignment problem. A survey on the quadratic assignment problem [25] and references therein provide examples and methods that are currently used to solve these problems.

B. Contributions

Our primary contribution addresses permutation recovery in the noisy version of observation model (1), with a random design matrix A. In particular, when the entries of A are drawn i.i.d. from a standard Gaussian matrix, we show sharp conditions on the SNR under which exact permutation recovery is possible. We also derive necessary conditions for approximate permutation recovery to within a prescribed Hamming distortion. We also briefly address the computational aspect of the permutation recovery problem. We show that the information theoretically optimal estimator we propose for exact permutation recovery is NP-hard to compute in the worst case. For the special case of $d = 1$, however, we show that it can be computed in polynomial time. Our results are corroborated by numerical simulations.

C. Organization

The paper is organized as follows. In the next section, we set up notation and formally state the problem. In Section III, we state our main results and discuss some of their implications. We provide proofs of the main results in Section IV, deferring the more technical lemmas to the appendices.

II. Background and problem setting

In this section, we set up notation and formally state the problem we wish to solve.

A. Notation

Since most of our analysis involves metrics involving permutations, we introduce all the relevant notation in this section. Permutations are denoted by π and permutation matrices by Π. We use $\pi(i)$ to denote the image of an element i under the permutation π. With a minor abuse of notation, we let $\mathcal{P}_n$ denote both the set of permutations on n objects as well as the corresponding set of permutation matrices. We sometimes use the compact notation y_π (or y_Π) to denote the vector y with entries permuted according to the permutation π (or Π).

We let $\mathsf{d_H}(\pi, \pi')$ denote the Hamming distance between two permutations. More formally, we have $\mathsf{d_H}(\pi, \pi') := \#\{i \mid \pi(i) \neq \pi'(i)\}$. For convenience, we let $\mathsf{d_H}(\Pi, \Pi')$ denote the Hamming distance between two permutation matrices, which is to be interpreted as the Hamming distance between the corresponding permutations.

The notation v_i denotes the ith entry of a vector v. We denote the ith standard basis vector in $\mathbb{R}^d$ by e_i. We use the notation $a_i^\top$ to refer to the ith row of A. We also use the standard shorthand notation $[n] := \{1, 2, \ldots, n\}$.

We also make use of standard asymptotic $\mathcal{O}$ notation. Specifically, for two real sequences f_n and g_n, $f_n = \mathcal{O}(g_n)$ means that $f_n \leq C g_n$ for a universal constant $C > 0$. Lastly, all logarithms denoted by $\log$ are to the base e, and we use c_1, c_2, etc. to denote absolute constants that are independent of other problem parameters.

B. Formal problem setting and permutation recovery

As mentioned in the introduction, we focus exclusively on the noisy observation model in the random design setting. In other words, we obtain an n-vector of observations y from the model (1) with $n \geq d$ to ensure identifiability, and with the following assumptions:

Signal model: The vector $x^* \in \mathbb{R}^d$ is fixed, but unknown. We note that this is different from the *adversarial* signal model of Unnikrishnan et al. [3].

Measurement matrix: The measurement matrix $A \in \mathbb{R}^{n \times d}$ is a random matrix of i.i.d. standard Gaussian variables chosen without knowledge of x^*. Our assumption on i.i.d. standard Gaussian designs easily extends to accommodate the more general case when rows of A are drawn i.i.d. from the distribution $\mathcal{N}(0, \Sigma)$. In particular, writing $A = W\sqrt{\Sigma}$, where W in an $n \times d$ standard Gaussian matrix and $\sqrt{\Sigma}$ denotes the symmetric square root of the (non-singular) covariance matrix Σ, our observation model takes the form

$$y = \Pi^* W \sqrt{\Sigma} x^* + w,$$

and the unknown vector is now $\sqrt{\Sigma}x^*$ in the model (1).

Noise variables: The vector $w \sim \mathcal{N}(0, \sigma^2 I_n)$ represents uncorrelated noise variables, each of (possibly unknown) variance σ^2. As will be made clear in the analysis, our assumption that the noise is Gaussian also readily extends to accommodate i.i.d. σ-sub-Gaussian noise. Additionally, the permutation noise represented by the unknown permutation matrix Π^* is arbitrary.

The main recovery criterion we address is that of exact permutation recovery, which we describe below. Following that, we also discuss two other relevant recovery criteria.

Exact permutation recovery: The problem of exact permutation recovery is to recover Π^*, and the risk of an estimator is evaluated on the 0-1. More formally, given an estimator of Π^* denoted by $\widehat{\Pi} : (y, A) \to \mathcal{P}_n$, we evaluate its risk by

$$\Pr\{\widehat{\Pi} \neq \Pi^*\} = \mathbb{E}\left[\mathbf{1}\{\widehat{\Pi} \neq \Pi^*\}\right], \tag{2}$$

where the probability in the LHS is taken over the randomness in y induced by both A and w.

Approximate permutation recovery: It is reasonable to think that recovering Π^* up to some distortion is sufficient for many applications. Such a relaxation of exact permutation recovery allows the estimator to output a $\widehat{\Pi}$ such that $\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \leq D$, for some distortion D to be specified. The risk of such an estimator is again evaluated on the 0-1 loss of this error metric, given by $\Pr\{\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \geq D\}$, with the probability again taken over both A and w. While our results are derived mainly in the context of exact permutation recovery, they can be suitably modified to also yields results for approximate permutation recovery.

Recovery with side information: In this variation, the unknown permutation matrix is not arbitrary, but known to be in some Hamming ball around the identity matrix. In other words, the estimator is provided with side information that $\mathsf{d_H}(\Pi^*, I) \leq \bar{h}$, for some $\bar{h} < n$. In many applications, this may constitute a prior that leads us to believe that the permutation matrix is not arbitrary. In multi-target tracking, for example, we may be sure that at any given time, a certain number of measurements correspond to the true sensors that made them (that are close to the target, perhaps). Our results also address exact permutation recovery with side information.

We are now in a position to state our main results.

III. Main results

In this section, we state our main theorems and discuss their consequences. Proofs can be found in Section IV.

A. Statistical limits of exact permutation recovery

Our main theorems in this section provide necessary and sufficient conditions under which the probability of error in exactly recovering the true permutation goes to zero.

In brief, provided that d is sufficiently small, we establish a threshold phenomenon that characterizes how the signal-to-noise ratio $\mathsf{snr} := \frac{\|x^*\|_2^2}{\sigma^2}$ must scale relative to n in order to ensure identifiability. More specifically, defining the ratio

$$\Gamma(n, \mathsf{snr}) := \frac{\log(1 + \mathsf{snr})}{\log n},$$

we show that the maximum likelihood estimator recovers the true permutation with high probability provided $\Gamma(n, \mathsf{snr}) \gg c$, where c denotes an absolute constant. Conversely, if $\Gamma(n, \mathsf{snr}) \ll c$, then exact permutation recovery is impossible. For illustration, we have plotted the behaviour of the maximum likelihood estimator for the case when $d = 1$ in Figure 2. Evidently, there is a sharp phase transition between error and exact recovery as the ratio $\Gamma(n, \mathsf{snr})$ varies from 3 to 5.

Let us now turn to more precise statements of our results. We first define the maximum likelihood estimator (MLE) as

$$(\widehat{\Pi}_{\mathsf{ML}}, \widehat{x}_{\mathsf{ML}}) = \arg \min_{\substack{\Pi \in \mathcal{P}_n \\ x \in \mathbb{R}^d}} \|y - \Pi A x\|_2^2. \tag{3}$$

The following theorem provides an upper bound on $\Pr\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\}$, with (c_1, c_2) denoting absolute constants.

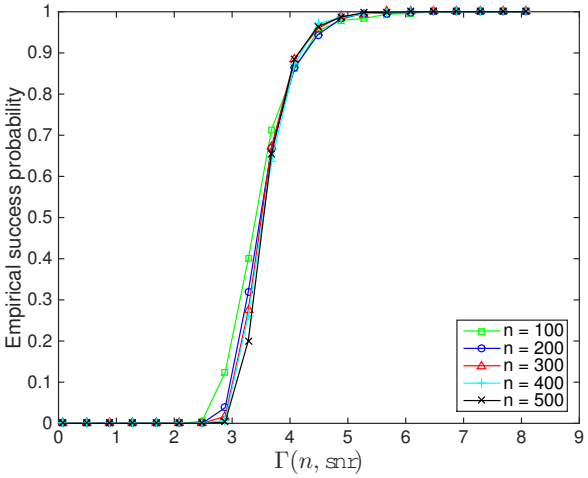


Fig. 2. Empirical frequency of the event $\{\widehat{\Pi}_{\mathsf{ML}} = \Pi^*\}$ over 1000 independent trials with $d = 1$, plotted against $\Gamma(n, \mathsf{snr})$ for different values of n. The probability of successful permutation recovery undergoes a phase transition as $\Gamma(n, \mathsf{snr})$ varies from 3 to 5. This is consistent with the prediction of Theorems 1 and 2.

Theorem 1. *For any $d < n$ and $\epsilon < \sqrt{n}$, if*

$$\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) \geq \left(c_1 \frac{n}{n-d} + \epsilon\right) \log n, \tag{4}$$

then $\Pr\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} \leq c_2 n^{-2\epsilon}$.

Theorem 1 provides conditions on the signal-to-noise ratio $\mathsf{snr} = \frac{\|x^*\|_2^2}{\sigma^2}$ that are sufficient for permutation recovery in the non-asymptotic, noisy regime. In contrast, the results of Unnikrishnan et al. [3] are stated in the limit $\mathsf{snr} \to \infty$, without an explicit characterization of the scaling behavior.

We also note that Theorem 1 holds for all values of $d < n$, whereas the results of Unnikrishnan et al. [3] require $n \geq 2d$ for identifiability of x^* in the noiseless case. Although the recovery of Π^* and x^* are not directly comparable, it is worth pointing out that the discrepancy also arises due to the difference between our fixed and unknown signal model, and the adversarial signal model assumed in the paper [3].

We now turn to the following converse result, which complements Theorem 1.

Theorem 2. *For any $\delta \in (0, 2)$, if*

$$2 + \log\left(1 + \frac{\|x^*\|_2^2}{\sigma^2}\right) \leq (2 - \delta) \log n, \tag{5}$$

then $\Pr\{\widehat{\Pi} \neq \Pi^*\} \geq 1 - c_3 e^{-c_4 n \delta}$ *for any estimator* $\widehat{\Pi}$.

Theorem 2 serves as a "strong converse" for our problem, since it guarantees that if condition (5) is satisfied, then the probability of error of any estimator goes to 1 as n goes to infinity. Indeed, it is proved using the strong converse argument for the Gaussian channel [26], which yields a converse result for any *fixed* design matrix A (see (18)). In fact, we are also able to show the following "weak converse" for Gaussian designs in the presence of side information.

Proposition 1. *If $n \geq 9$ and*

$$\log\left(1+\frac{\|x^*\|_2^2}{\sigma^2}\right) \leq \frac{8}{9}\log\left(\frac{n}{8}\right),$$

then $\Pr\{\widehat{\Pi} \neq \Pi^*\} \geq 1/2$ for any estimator $\widehat{\Pi}$, even if it is known a-priori that $\mathsf{d_H}(\Pi^*, I) \leq 2$.

As mentioned earlier, restriction of Π^* constitutes some application-dependent prior; the strongest such prior restricts it to a Hamming ball of radius 2 around the identity. Proposition 1 asserts that even this side information does not substantially change the statistical limits of permutation recovery.

It is also worth noting that the converse results hold uniformly over d. In particular, if $d \leq pn$ for some fixed $p < 1$, Theorems 1 and 2 together yield the threshold behavior of identifiability in terms of $\Gamma(n, \mathsf{snr})$ that was discussed above. In the next section, we find that a similar phenomenon occurs even with approximate permutation recovery.

B. Limits of approximate permutation recovery

The techniques we used to prove results for exact permutation recovery can be suitably modified to obtain results for approximate permutation recovery to within a Hamming distortion D. In particular, we show the following converse result for approximate recovery.

Theorem 3. *For any $2 < D \leq n-1$, if*

$$\log\left(1+\frac{\|x^*\|_2^2}{\sigma^2}\right) \leq \frac{n-D+1}{n}\log\left(\frac{n-D+1}{2e}\right), \quad (6)$$

then $\Pr\{\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \geq D\} \geq 1/2$ for any estimator $\widehat{\Pi}$.

Note that for any $D \leq pn$ with $p \in (0,1)$, Theorems 1 and 3 provide a set of sufficient and necessary conditions for approximate permutation recovery that match up to constant factors. In particular, the necessary condition resembles that for exact permutation recovery, and the same SNR threshold behaviour is observed. We remark that a corresponding converse with side information can also be proved for approximate permutation recovery using techniques similar to the proof of Proposition 1. It is also worth mentioning the following:

Remark 1. *The converse results given by Theorem 2, Proposition 1, and Theorem 3 hold even when the estimator has exact knowledge of x^*.*

C. Computational aspects

In the previous sections, we considered the MLE given by equation (3) and analyzed its statistical properties. However, since equation (3) involves a combinatorial minimization over $n!$ permutations, it is unclear if $\widehat{\Pi}_{\mathsf{ML}}$ can be computed efficiently. The following theorem addresses this question.

Theorem 4. *For $d = 1$, the MLE $\widehat{\Pi}_{\mathsf{ML}}$ can be computed in time $\mathcal{O}(n \log n)$ for any choice of the measurement matrix A. In contrast, if $d > 1$, then $\widehat{\Pi}_{\mathsf{ML}}$ is NP-hard to compute.*

The algorithm used to prove the first part of the theorem involves a simple sorting operation, which introduces the $\mathcal{O}(n \log n)$ complexity. We emphasize that the algorithm assumes no prior knowledge about the distribution of the data; for every given A and y, it returns the optimal solution to problem (3).

The second part of the theorem asserts that the algorithmic simplicity enjoyed by the $d = 1$ case does not extend to general d. The proof proceeds by a reduction from the NP-complete partition problem. We stress here that the NP-hardness claim holds over worst case input instances. In particular, it does not preclude the possibility that there exists a polynomial time algorithm that solves problem (3) with high probability when A is chosen randomly as in our original setting. However, we conjecture that solving problem (3) over random A is also a computationally hard problem, conditioned on an average-case hardness assumption.

IV. Proof Sketches of Some Results

Due to space constraints, we only sketch proofs of some of our main results. In particular, we prove Theorem 1 for the special case when $d = 1$, and Theorem 2. Complete proofs of all of our results can be found in the full version of the paper [27].

Throughout the proofs, we assume that n is larger than some universal constant. The case where n is smaller can be handled by changing the constants in our proofs appropriately. We also use the notation c, c' to denote absolute constants that can change from line to line.

A. Proof sketch of Theorem 1: $d = 1$ case

We prove Theorem 1 by bounding the probability that a fixed permutation is preferred to Π^* by the estimator. The analysis requires precise control on the lower tails of χ^2-random variables, which are proved in the full version [27].

For a fixed $\Pi \in \mathcal{P}_n$, consider the random variable

$$\Delta(\Pi, \Pi^*) := \|P_{\Pi}^{\perp} y\|_2^2 - \|P_{\Pi^*}^{\perp} y\|_2^2. \quad (7)$$

For any permutation Π, the estimator (3) prefers the permutation Π to Π^* if $\Delta(\Pi, \Pi^*) \leq 0$. The overall error event occurs when $\Delta(\Pi, \Pi^*) \leq 0$ for some Π, meaning that

$$\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} = \bigcup_{\Pi \in \mathcal{P}_n \setminus \Pi^*} \{\Delta(\Pi, \Pi^*) \leq 0\}. \quad (8)$$

In order to prove the theorem, we first bound the probability of each error event in the RHS of equation (8) using the following key lemma. Recall the definition of $\mathsf{d_H}(\Pi, \Pi')$, the Hamming distance between two permutation matrices.

Lemma 1. *For $d = 1$ and any two permutation matrices Π and Π^*, we have*

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \exp\left(-c\, \mathsf{d_H}(\Pi, \Pi^*) \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right).$$

Lemma 1 is proved in Section IV-A1. Taking it as given for the moment, we can then prove Theorem 1 for the $d = 1$

case by a union bound. In particular, begin by fixing $\epsilon > 0$ and assume that

$$c \log \left(\frac{\|x^*\|_2^2}{\sigma^2} \right) \geq (1+\epsilon) \log n, \tag{9}$$

where c is the same as in Lemma 1. Now, observe that

$$\begin{aligned}
\Pr\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} &\leq \sum_{\Pi \in \mathcal{P}_n \setminus \Pi^*} \Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \\
&\overset{\text{(i)}}{\leq} c' \sum_{2 \leq k \leq n} n^k \exp\left(-c\, k \log \left(\frac{\|x^*\|_2^2}{\sigma^2} \right) \right) \\
&\overset{\text{(ii)}}{\leq} c' \sum_{2 \leq k \leq n} n^{-\epsilon k} \\
&\leq c' \frac{1}{n^\epsilon (n^\epsilon - 1)}.
\end{aligned}$$

where step (i) follows from Lemma 1 and since $\#\{\Pi : \mathsf{d_H}(\Pi, \Pi^*) = k\} \leq n^k$, and step (ii) follows from condition (9). Relabelling the constants in condition (9) proves the theorem. □

It remains to prove Lemma 1.

1) Proof sketch of Lemma 1: We first evaluate the probability over the randomness in w holding A fixed, and then consider the randomness in A. We begin by splitting the quantity $\Delta(\Pi, \Pi^*)$ into two and analyzing the terms individually. In particular, for each $\delta > 0$, define the events

$$\mathcal{F}_1(\delta) = \left\{ \left| \|P_{\Pi^*}^\perp y\|_2^2 - \|P_\Pi^\perp w\|_2^2 \right| \geq \delta \right\}, \text{ and} \tag{10a}$$

$$\mathcal{F}_2(\delta) = \left\{ \|P_\Pi^\perp y\|_2^2 - \|P_\Pi^\perp w\|_2^2 \leq 2\delta \right\}. \tag{10b}$$

It is easy to verify that the following inclusion holds for any $\delta > 0$:

$$\{\Delta(\Pi, \Pi^*) \leq 0\} \subseteq \mathcal{F}_1(\delta) \cup \mathcal{F}_2(\delta). \tag{11}$$

Accordingly, we bound the probability of the two events $\mathcal{F}_1(\delta)$ and $\mathcal{F}_2(\delta)$ individually, and then use the union bound to prove the lemma for a fixed $\delta = \delta^*$.

Lemma 2. *For any $\delta > 0$ and with $\delta^* = \frac{1}{3}\|P_\Pi^\perp \Pi^* A x^*\|_2^2$, we have*

$$\Pr_w\{\mathcal{F}_1(\delta)\} \leq c' \exp\left(-c \frac{\delta}{\sigma^2} \right), \text{ and} \tag{12a}$$

$$\Pr_w\{\mathcal{F}_2(\delta^*)\} \leq c' \exp\left(-c \frac{\delta^*}{\sigma^2} \right). \tag{12b}$$

The proofs of both claims use algebraic manipulation and basic sub-Gaussian and sub-exponential tail bounds, and can be found in Appendix B of the full version [27].

Applying Lemma 2 and using the union bound then yields

$$\begin{aligned}
\Pr_w\{\Delta(\Pi, \Pi^*) \leq 0\} &\leq \Pr_w\{\mathcal{F}_1(\delta^*)\} + \Pr_w\{\mathcal{F}_2(\delta^*)\} \\
&\leq c' \exp\left(-c \frac{T_\Pi}{\sigma^2} \right),
\end{aligned} \tag{13}$$

where we have introduced the shorthand $T_\Pi := \|P_\Pi^\perp \Pi^* A x^*\|_2^2$.

It remains to evaluate the probability over the randomness in A. As a first step, we provide a tail bound on the random variable T_Π. We let $h := \mathsf{d_H}(\Pi, \Pi^*)$ denote the Hamming distance between Π and Π^*.

Lemma 3. *For $d = 1$ and any two permutation matrices Π and Π^* at Hamming distance h, we have*

$$\Pr_A\{T_\Pi \leq t \|x^*\|_2^2\} \leq 6 \exp\left(-\frac{h}{10} \left[\log \frac{h}{t} + \frac{t}{h} - 1 \right] \right), \tag{14}$$

for all $t \in [0, h]$.

Lemma 3 is proved in Section IV-A2. Assuming it to be true for the moment, we can then combine it with the inequality (13) to write

$$\begin{aligned}
\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} &\leq c' \exp\left(-c \frac{t\|x^*\|_2^2}{\sigma^2} \right) \Pr_A\{T_\Pi \geq t\|x^*\|_2^2\} \\
&\quad + \Pr_A\{T_\Pi \leq t\|x^*\|_2^2\} \\
&\leq c' \exp\left(-c \frac{t\|x^*\|_2^2}{\sigma^2} \right) \\
&\quad + 6 \exp\left(-\frac{h}{10} \left[\log \frac{h}{t} + \frac{t}{h} - 1 \right] \right),
\end{aligned} \tag{15}$$

where the last inequality holds provided that $tin[0, h]$, and the probability in the LHS is now taken over randomness in both w *and* A.

Minimizing the RHS of inequality (15) over $t \in [0, h]$ (details can be found in the full version [27]) yields that for all $\frac{\|x^*\|_2^2}{\sigma^2} \geq 1$, we have

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \exp\left(-ch \log \left(\frac{\|x^*\|_2^2}{\sigma^2} \right) \right). \tag{16}$$

This completes the proof of Lemma 1. □

The only remaining piece is the proof of Lemma 3.

2) Proof sketch of Lemma 3: In the case $d = 1$, the matrix A is composed of one column, which we denote by a. Recall the random variable $T_\Pi = \|P_\Pi^\perp \Pi^* A x^*\|_2^2$, which can be written as

$$\begin{aligned}
T_\Pi &= (x^*)^2 \left(\|a\|_2^2 - \frac{1}{\|a\|_2^2} \langle a_\Pi, a \rangle^2 \right) \\
&\overset{\text{(i)}}{\geq} (x^*)^2 \left(\|a\|_2^2 - |\langle a, a_\Pi \rangle| \right) \\
&= \frac{(x^*)^2}{2} \min\left(\|a - a_\Pi\|_2^2, \|a + a_\Pi\|_2^2 \right),
\end{aligned}$$

where step (i) follows from the Cauchy Schwarz inequality. Applying the union bound then yields

$$\begin{aligned}
\Pr\{T_\Pi \leq t(x^*)^2\} &\leq \Pr\{\|a - a_\Pi\|_2^2 \leq 2t\} \\
&\quad + \Pr\{\|a + a_\Pi\|_2^2 \leq 2t\}.
\end{aligned}$$

Let Z_ℓ and $\tilde{Z}_\ell$ denote (not necessarily independent) χ^2 random variables with ℓ degrees of freedom. For $h \geq 3$,

applying Lemma 4 from the Appendix guarantees that

$$\frac{\|a-a_\Pi\|_2^2}{2} \stackrel{d}{=} Z_{h_1} + Z_{h_2} + Z_{h_3}, \text{ and} \tag{17a}$$

$$\frac{\|a+a_\Pi\|_2^2}{2} \stackrel{d}{=} \widetilde{Z}_{h_1} + \widetilde{Z}_{h_2} + \widetilde{Z}_{h_3} + \widetilde{Z}_{n-h}, \tag{17b}$$

where $\stackrel{d}{=}$ denotes equality in distribution and $h_1, h_2, h_3 \geq \frac{h}{5}$ with $h_1 + h_2 + h_3 = h$. An application of the union bound then yields

$$\Pr\{\|a-a_\Pi\|_2^2 \leq 2t\} \leq \sum_{i=1}^{3} \Pr\left\{Z_{h_i} \leq t\frac{h_i}{h}\right\}.$$

Similarly, provided $h \geq 3$, we have

$$\begin{aligned}\Pr\{\|a+a_\Pi\|_2^2 \leq 2t\} &\leq \Pr\{\widetilde{Z}_{h_1} + \widetilde{Z}_{h_2} + \widetilde{Z}_{h_3} + \widetilde{Z}_{n-h} \leq t\} \\ &\stackrel{\text{(ii)}}{\leq} \Pr\{\widetilde{Z}_{h_1} + \widetilde{Z}_{h_2} + \widetilde{Z}_{h_3} \leq t\} \\ &\stackrel{\text{(iii)}}{\leq} \sum_{i=1}^{3} \Pr\left\{\widetilde{Z}_{h_i} \leq t\frac{h_i}{h}\right\},\end{aligned}$$

where inequality (ii) follows by monotonicity of the CDF since $\widetilde{Z}_{n-h} \geq 0$, and inequality (iii) by the union bound. Finally, bounds on the lower tails of χ^2 random variables (see Lemma 5 in the Appendix) yield

$$\begin{aligned}\Pr\left\{Z_{h_i} \leq t\frac{h_i}{h}\right\} &= \Pr\left\{\widetilde{Z}_{h_i} \leq t\frac{h_i}{h}\right\} \\ &\stackrel{\text{(iv)}}{\leq} \left(\frac{t}{h}\exp\left(1-\frac{t}{h}\right)\right)^{h_i/2} \\ &\stackrel{\text{(v)}}{\leq} \left(\frac{t}{h}\exp\left(1-\frac{t}{h}\right)\right)^{h/10}.\end{aligned}$$

Here, inequality (iv) is valid provided $\frac{th_i}{h} \leq h_i$, or equivalently, if $t \leq h$; inequality (v) follows since $h_i \geq h/5$ and the function $f(x) = xe^{1-x} \in [0,1]$ for all $x \in [0,1]$. Combining the pieces proves Lemma 3 for $h \geq 3$.

If $h = 2$, we have

$$\frac{\|a-a_\Pi\|_2^2}{2} \stackrel{d}{=} 2Z_1, \quad \text{and} \quad \frac{\|a+a_\Pi\|_2^2}{2} \stackrel{d}{=} 2\widetilde{Z}_1 + \widetilde{Z}_{n-2}.$$

Proceeding as before by applying the union bound and Lemma 5, we have that for $h = 2$ and $t \leq 2$, the random variable T_Π obeys the tail bound

$$\begin{aligned}\Pr\{T_\Pi \leq t(x^*)^2\} &\leq 2\left(\frac{t}{2}\exp\left(1-\frac{t}{2}\right)\right)^{1/2} \\ &\leq 6\left(\frac{t}{h}\exp\left(1-\frac{t}{h}\right)\right)^{h/10}.\end{aligned}$$

The proof of Lemma 3 is thus complete. □

B. Proof sketch of Theorem 2

We begin by assuming that the design matrix A is fixed, and that the estimator has knowledge of x^* a-priori, since the latter cannot make the estimation task any easier. We can also assume that the entries of Ax^* are distinct, since otherwise, perfect permutation recovery is impossible.

Given this setup, we now cast the problem as one of coding over a Gaussian channel. Toward this end, consider the codebook

$$\mathcal{C} = \{\Pi Ax^* \mid \Pi \in \mathcal{P}_n\}.$$

We may view ΠAx^* as the codeword corresponding to the permutation Π, where each permutation is associated to one of $n!$ equally likely messages. Note that each codeword has power $\|Ax^*\|_2^2$.

The codeword is then sent over a Gaussian channel with noise power equal to $\sum_{i=1}^n \sigma^2 = n\sigma^2$. The decoding problem is to ascertain from the noisy observations which message was sent, or in other words, to identify the correct permutation.

We now use the non-asymptotic strong converse for the Gaussian channel [29]. In particular, using the result [27, Lemma 12] with code rate $R = \frac{\log n!}{n}$ then yields that for any $\delta' > 0$, if

$$\frac{\log n!}{n} > \frac{1+\delta'}{2}\log\left(1+\frac{\|Ax^*\|_2^2}{n\sigma^2}\right),$$

then for any estimator $\widehat{\Pi}$, we have $\Pr\{\widehat{\Pi} \neq \Pi\} \geq 1 - 2\cdot 2^{-n\delta'}$. For the choice $\delta' = \delta/(2-\delta)$, we have that if

$$(2-\delta)\log\left(\frac{n}{e}\right) > \log\left(1+\frac{\|Ax^*\|_2^2}{n\sigma^2}\right), \tag{18}$$

then $\Pr\{\widehat{\Pi} \neq \Pi\} \geq 1 - 2\cdot 2^{-n\delta/2}$. Note that the only randomness assumed so far was in the noise w and the random choice of Π.

We now specialize the result for the case when A is Gaussian. Toward that end, define the event

$$\mathcal{E}(\delta) = \left\{1+\delta \geq \frac{\|Ax^*\|_2^2}{n\|x^*\|_2^2}\right\}.$$

Conditioned on the event $\mathcal{E}(\delta)$, it can be verified that condition (5) implies condition (18). We also have

$$\begin{aligned}\Pr\{\mathcal{E}(\delta)\} &= 1 - \Pr\left\{\frac{\|Ax^*\|_2^2}{n\|x^*\|_2^2} > 1+\delta\right\} \\ &\stackrel{\text{(i)}}{\geq} 1 - c'e^{-cn\delta},\end{aligned}$$

where step (i) follows by using the sub-exponential tail bound (see, e.g., [27, Lemma 10]), since $\frac{\|Ax^*\|_2^2}{\|x^*\|_2^2} \sim \chi_n^2$.

Putting together the pieces, we have that provided condition (5) holds,

$$\begin{aligned}\Pr\{\widehat{\Pi} \neq \Pi^*\} &\geq \Pr\{\widehat{\Pi} \neq \Pi^* | \mathcal{E}(\delta)\}\Pr\{\mathcal{E}(\delta)\} \\ &= (1 - 2\cdot 2^{-n\delta/2})(1 - c'e^{-cn\delta}) \\ &\geq 1 - c'e^{-cn\delta}.\end{aligned}$$

□

V. Discussion

We analyzed the problem of exact permutation recovery in the linear regression model, and provided necessary and sufficient conditions that are tight in most regimes of n and d. We also provided a converse for the problem of approximate permutation recovery to within some Hamming distortion. It

is still an open problem to characterize the fundamental limits of exact and approximate permutation recovery for all regimes of n, d and the allowable distortion D. In the context of exact permutation recovery, we believe that the limit suggested by Theorem 1 is tight for all regimes of n and d, but showing this will likely require a different technique. In particular, as pointed out in Remark 1, all of our lower bounds assume that the estimator is provided with x^* as side information; it is an interesting question as to whether stronger lower bounds can be obtained without this side information.

On the computational front, many open questions remain. The primary question concerns the design of computationally efficient estimators that succeed in similar SNR regimes. We have already shown that the maximum likelihood estimator, while being statistically optimal for moderate d, is computationally hard to compute in the worst case. Showing a corresponding hardness result for random A is also an open problem. Finally, while this paper mainly addresses the problem of permutation recovery, the complementary problem of recovering x^* is also interesting, and we plan to investigate its fundamental limits in future work.

Acknowledgements

This work was partially supported by NSF Grants CCF-1528132 and CCF-0939370 (Center for Science of Information), Office of Naval Research MURI grant DOD-002888, Air Force Office of Scientific Research Grant AFOSR-FA9550-14-1-001, Office of Naval Research grant ONR-N00014, as well as National Science Foundation Grant CIF-31712-23800.

REFERENCES

[1] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[2] P. Loh and M. J. Wainwright, "Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 2601–2605.

[3] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli, "Unlabeled sensing with random linear measurements," *preprint arXiv:1512.00115*, 2015.

[4] A. V. Balakrishnan, "On the problem of time jitter in sampling," *IRE Transactions on Information Theory*, vol. 8, no. 3, pp. 226–236, 1962.

[5] C. Rose, I. S. Mian, and R. Song, "Timing channels with multiple identical quanta," *arXiv preprint arXiv:1208.1070*, 2012.

[6] A. B. Poore and S. Gadaleta, "Some assignment problems arising from multiple target tracking," *Mathematical and Computer Modelling*, vol. 43, no. 9, pp. 1074–1091, 2006.

[7] S. Thrun and J. J. Leonard, "Simultaneous localization and mapping," in *Springer Handbook of Robotics*. Springer, 2008, pp. 871–889.

[8] W. S. Robinson, "A method for chronologically ordering archaeological deposits," *American Antiquity*, pp. 293–301, 1951.

[9] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008, pp. 111–125.

[10] L. Keller, M. J. Siavoshani, C. Fragouli, K. Argyraki, and S. Diggavi, "Identity aware sensor networks," in *INFOCOM 2009, IEEE*. IEEE, 2009, pp. 2177–2185.

[11] P. David, D. Dementhon, R. Duraiswami, and H. Samet, "Softposit: Simultaneous pose and correspondence determination," *International Journal of Computer Vision*, vol. 59, no. 3, pp. 259–284, 2004.

[12] M. Marques, M. Stošić, and J. Costeira, "Subspace matching: Unique solution to point matching with geometric constraints," in *Computer Vision, IEEE 12th International Conference on*. IEEE, 2009, pp. 1288–1294.

[13] S. Mann, "Compositing multiple pictures of the same scene," in *Proceedings of the 46th Annual IS&T Conference*, vol. 2, 1993, pp. 319–25.

[14] L. J. Schulman and D. Zuckerman, "Asymptotically good codes correcting insertions, deletions, and transpositions," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2552–2557, 1999.

[15] X. Huang and A. Madan, "CAP3: A DNA sequence assembly program," *Genome Research*, vol. 9, no. 9, pp. 868–877, 1999.

[16] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval, "Compressed sensing with unknown sensor permutation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1040–1044.

[17] Y. M. Lu and M. N. Do, "A theory for sampling signals from a union of subspaces," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2334–2345, 2008.

[18] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4660–4671, 2011.

[19] O. Collier and A. S. Dalalyan, "Minimax rates in permutation estimation for feature matching," *Journal of Machine Learning Research*, vol. 17, no. 6, pp. 1–31, 2016.

[20] N. Flammarion, C. Mao, and P. Rigollet, "Optimal rates of statistical seriation," *arXiv preprint arXiv:1607.02435*, 2016.

[21] F. Fogel, R. Jenatton, F. Bach, and A. d'Aspremont, "Convex relaxations for permutation problems," in *Advances in Neural Information Processing Systems*, 2013, pp. 1016–1024.

[22] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainright, "Stochastically transitive models for pairwise comparisons: Statistical and computational issues," *arXiv preprint arXiv:1510.05610*, 2015.

[23] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *The Annals of Statistics*, vol. 43, no. 1, pp. 177–214, 2015.

[24] J. Huang, C. Guestrin, and L. Guibas, "Fourier theoretic probabilistic inference over permutations," *The Journal of Machine Learning Research*, vol. 10, pp. 997–1070, 2009.

[25] E. M. Loiola, N. M. M. de Abreu, P. O. Boaventura-Netto, P. Hahn, and T. Querido, "A survey for the quadratic assignment problem," *European Journal of Operational Research*, vol. 176, no. 2, pp. 657–690, 2007.

[26] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell System Technical Journal*, vol. 38, no. 3, pp. 611–656, 1959.

[27] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Linear regression with an unknown permutation: Statistical and computational limits," *arXiv preprint arXiv:1608.02902*, 2016.

[28] 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.

[29] K. Yoshihara, "Simple proofs for the strong converse theorems in some channels," in *Kodai Mathematical Seminar Reports*, vol. 16. Dept. of Mathematics, Tokyo Institute of Technology, 1964, pp. 213–222.

APPENDIX A
AUXILIARY LEMMAS

Proofs of the following two lemmas can be found in the full version [27]. For a permutation π on k objects, let G_π denote the corresponding undirected incidence graph, i.e., $V(G_\pi) = [k]$, and $(i,j) \in E(G_\pi)$ iff $j = \pi(i)$ or $i = \pi(j)$.

Lemma 4. *Let π be a permutation on $k \geq 3$ objects such that $\mathsf{d_H}(\pi, I) = k$. Then the vertices of G_π can be partitioned into three sets V_1, V_2, V_3 such that each is an independent set, and $|V_1|, |V_2|, |V_3| \geq \lfloor \frac{k}{3} \rfloor \geq \frac{k}{5}$.*

Lemma 5. *Let Z_ℓ denote a χ^2 random variable with ℓ degrees of freedom. Then its CDF for any $0 \leq p \leq \ell$ is upper bounded as*

$$\begin{aligned}\Pr\{Z_\ell \leq p\} &\leq \left(\frac{p}{\ell}\exp\left(1-\frac{p}{\ell}\right)\right)^{\ell/2} \\ &= \exp\left(-\frac{\ell}{2}\left[\log\frac{\ell}{p} + \frac{p}{\ell} - 1\right]\right). \end{aligned} \tag{19}$$