Denoising Linear Models with Permuted Data

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

***Abstract*—We consider the multivariate linear regression model with shuffled data and additive noise, which arises in various correspondence estimation and matching problems. We focus on the denoising problem and characterize the minimax error rate up to logarithmic factors. We also analyze the performance of two versions of a computationally efficient estimator that are consistent for a large range of input parameters. Finally, we provide an exact algorithm for the noiseless problem and demonstrate its performance on an image point-cloud matching task. Our analysis also extends to datasets with missing data.**

I. Introduction

The linear model is a ubiquitous and well-studied tool for predicting responses y_i from covariates a_i using n samples of data $\{a_i, y_i\}_{i=1}^n$. In this paper, we consider the multivariate version of the model with vector-valued response variables $y_i \in \mathbb{R}^m$, and covariates $a_i \in \mathbb{R}^d$. However, our input consists of shuffled or permuted data $\{a_i, y_{\pi_i}\}_{i=1}^n$, where π represents an unknown permutation. In other words, stacking up each data point as a row of a matrix, we consider the model

$$Y = \Pi^* A X^* + W, \tag{1}$$

where $Y \in \mathbb{R}^{n\times m}$ is the matrix of responses, Π^* is an unknown $n \times n$ permutation matrix, $A \in \mathbb{R}^{n\times d}$ is the matrix of covariates, $X^* \in \mathbb{R}^{d\times m}$ is an unknown matrix of parameters, and W is the additive observation noise[1]. When $m = 1$, this reduces to the vector linear regression model with an unknown permutation, given by

$$y = \Pi^* A x^* + w, \tag{2}$$

which we refer to as the shuffled vector model.

More precisely, we analyze the multivariate model (1) with a fixed design matrix A, and Gaussian[2] noise $W_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. We evaluate an estimator $(\widehat{\Pi}, \widehat{X})$ based on its "denoising" capability, which we capture using the normalized prediction error $\frac{1}{nm}\|\widehat{\Pi} A \widehat{X} - \Pi^* A X^*\|_F^2$. Our primary objective in this paper is to characterize the prediction error in a minimax sense, and we analyze the quantity

$$\inf_{\substack{\widehat{\Pi}\in\mathcal{P}_n \\ \widehat{X}\in\mathbb{R}^{d\times m}}} \sup_{\substack{\Pi^*\in\mathcal{P}_n \\ X^*\in\mathbb{R}^{d\times m}}} \mathbb{E}\left[\frac{1}{nm}\|\widehat{\Pi} A \widehat{X} - \Pi^* A X^*\|_F^2\right], \tag{3}$$

where the expectation is taken over the noise W, and any randomness in the estimator $(\widehat{\Pi}, \widehat{X})$. We also provide algorithms having small worst-case prediction error.

The observation model (1) arises in multiple applications, which are discussed in detail for the shuffled vector model (2) in our earlier work [1]. In this paper, we focus on two applications relevant to the multivariate setting, which we use as running examples throughout the paper.

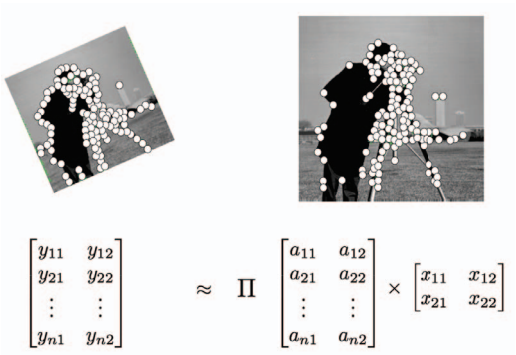

Fig. 1. Example of pose and correspondence estimation for 2D images. The image coordinates are related by an unknown resizing and rotation X. The unknown permutation represents the correspondence between keypoints (white circles) obtained via corner-detection. The matrices Y and A represent coordinates of all keypoints, and approximately obey the relation (1) because all the keypoints detected in the two images are not the same.

The first is the problem of pose and correspondence estimation in images [2], closely related to point-cloud matching in graphics [3]. In this application, we are given two images of a similar object, with the coordinates of one image arising from an unknown linear transformation of the coordinates of the second. In order to determine the linear transformation, keypoints are detected in each of the images individually and then matched. An example is provided in Figure 1. We emphasize that in practice, the keypoint detection algorithm also returns features that help in finding the matching permutation Π^*, but our goal here is to analyze whether there are procedures that are robust to such features being missing or corrupted.

The second application is that of header-free communication in large communication networks [1]. Here, an underlying matrix parameter X^* is measured by multiple sensors, each of which takes a noisy linear observation of the form $a_i^\top X^* + w_i^\top$. In very large networks for Internet of Things applications, for example, it is often seen that the bandwidth between a sensor and fusion center is mainly dominated by a header containing identity information, i.e., a bitstring that identifies sensor i to the fusion center [4]. One possible solution to this problem is header-free communication, which corresponds to introducing the unknown permutation matrix as in our model. If we are still able to achieve similar statistical performance without these headers, then such an approach is clearly preferable from a bandwidth standpoint.

In both the aforementioned applications, estimators with small minimax prediction error are of interest. In the pose and correspondence estimation problem, this amounts to obtaining near-identical keypoint locations on both images; in the sensor network example, we are interested in obtaining a set of noise-

This work was supported in part by NSF grants CCF-1528132 and CCF-0939370.

[1]We refer to the setting $W = 0$ *a.s.* as the noiseless case.

[2]Our results also extend to the case of i.i.d. sub-Gaussian noise.

free linear functions of the input signal. It is important to note that depending on the application, multiple regimes of the parameters n, d and m are of interest. Therefore, in this paper, we focus on capturing the dependence of denoising error rates on all of these parameters.

Our work contributes to the growing bodies of literature on regression problems with unknown permutations and fits within a broader framework of row-space perturbation problems like blind deconvolution [5], phase retrieval [6], and dictionary learning [7]. It is also related to problems of signal recovery from unions of subspaces [8]. Regression problems with unknown permutations have been considered in the context of statistical seriation and univariate isotonic matrix recovery [9], and non-parametric ranking from pairwise comparisons [10], which involves bivariate isotonic matrix recovery. Moreover, the prediction error is used to evaluate estimators in both these applications.

Specializing to our setting, the shuffled vector model (2) was first considered in the context of compressive sensing with a sensor permutation [11]. The first theoretical results were provided by Unnikrishnan et al. [12], who studied the conditions needed to recover an adversarially chosen x^* in the noiseless model with a random design matrix A. Also in the random design setting, our own previous work [1] focused on the complementary problem of recovering Π^* in the noisy model, and showed necessary and sufficient conditions on the SNR under which exact and approximate recovery were possible. An efficient algorithm to compute the maximum likelihood estimate was also provided for the case $d = 1$.

A. Notation

We use $\mathcal{P}_n$ to denote the set of permutation matrices. Let I_d denote the identity matrix of dimension d. We use $\|M\|_F$ and $\|M\|_*$ to denote the Frobenius and nuclear norms of a matrix M, and c, c_1, c_2 to denote universal constants that may change from line to line.

B. Contributions

First, we characterize the minimax prediction error of multivariate linear model with an unknown permutation up to a logarithmic factor, by analyzing the maximum likelihood estimator. Since the maximum likelihood estimate is NP-hard to compute in general [1], we then propose a computationally efficient estimator based on singular value thresholding and sharply characterize its performance, showing that it achieves vanishing prediction error over a restricted range of parameters. We also propose a variant of this estimator that achieves the same error rates, but with the advantage that it does not require the noise variance to be known. Third, we propose an efficient spectral algorithm for the noiseless problem that is exact provided certain natural conditions are met. We demonstrate this algorithm on an image point cloud matching task. Finally, we extend our results to a richer class of models that allows for duplicates and missing data in the dataset.

In the next section, we collect our main theorems and discuss their consequences. We sketch the proof of the minimax lower bound of Theorem 1 in Section III; the remaining proofs can be found in the full version [13].

II. Main results

In this section, we discuss our main results under four headings – minimax rates, polynomial time estimators, efficient procedures for the noiseless problem, and an extension of the model (1) that allows for duplicates.

A. Minimax rates of prediction

The noise W is i.i.d. Gaussian, so the maximum likelihood estimate (MLE) of the parameters (Π^*, X^*) is given by

$$(\widehat{\Pi}_{\mathsf{ML}}, \widehat{X}_{\mathsf{ML}}) = \arg \min_{\substack{\Pi \in \mathcal{P}_n \\ X \in \mathbb{R}^{d \times m}}} \|Y - \Pi A X\|_F^2. \tag{4}$$

We upper bound the prediction error achieved by the maximum likelihood estimator, which shows a distinct dependence on both unknown parameters Π^* and X^*. In general, however, we cannot prove a matching lower bound that captures both of these dependences for all matrices A. As an extreme example, consider the matrix A with identical rows, in which the unknown permutation Π^* plays no role in the observations, and so the denoising error should have no dependence on it.

Consequently, we derive lower bounds that apply provided the matrix A lies in a restricted class, in order to define which we require some additional notation. For a vector v, let v^s denote the vector sorted in decreasing order, and let $\mathbb{B}_{2,n}(1)$ denote the n-dimensional ℓ_2-ball of unit radius centered at 0. Define the matrix class

$$\mathcal{A}(\gamma, \xi) = \Big\{ A \in \mathbb{R}^{n \times d} \mid \exists a \in \mathsf{range}(A) \cap \mathbb{B}_{2,n}(1) \text{ with } a^s_{\lfloor \gamma n \rfloor} \geq a^s_{\lfloor \gamma n \rfloor + 1} + \xi \Big\}.$$

This defines matrices that are not "flat", in that there is some vector in their range obeying the (γ, ξ)-separation condition defined above. Loosely speaking, flat matrices that do not obey such a separation condition comprise almost identical rows. It can be verified that a matrix A with i.i.d. sub-Gaussian entries lies in the class $\mathcal{A}(C_1, C_2/\sqrt{n})$ with high probability for fixed constants C_1, C_2. We are now ready to state the theorem.

Theorem 1. *For any matrix A, and for all parameters $X^* \in \mathbb{R}^{d \times m}$ and $\Pi^* \in \mathcal{P}_n$, we have*

$$\frac{\|\widehat{\Pi}_{\mathsf{ML}} A \widehat{X}_{\mathsf{ML}} - \Pi^* A X^*\|_F^2}{nm} \leq c_1 \sigma^2 \left(\frac{\mathsf{rank}(A)}{n} + \frac{1}{m} \min\{\log n, m\} \right), \tag{5a}$$

with probability greater than $1 - e^{-c(n \log n + m\, \mathsf{rank}(A))}$.

Furthermore, if the matrix $A \in \mathcal{A}(C_1, C_2/\sqrt{n})$, then for any estimator $(\widehat{\Pi}, \widehat{X})$, we have

$$\sup_{\substack{\Pi^* \in \mathcal{P}_n \\ X^* \in \mathbb{R}^{d \times m}}} \mathbb{E}\left[\frac{\|\widehat{\Pi} A \widehat{X} - \Pi^* A X^*\|_F^2}{nm} \right] \geq c_2 \sigma^2 \left(\frac{\mathsf{rank}(A)}{n} + \frac{1}{m} \right). \tag{5b}$$

In the statement of Theorem 1, the constant c_2 depends on the value of the pair (C_1, C_2), but is independent of other problem parameters.

Theorem 1 characterizes the minimax rate up to a factor that is at most logarithmic in n. It shows that the MLE is

minimax optimal for prediction error up to logarithmic factors for all matrices that are not too flat. The bounds have the following interpretation, similar to the results of Flammarion et al. [9] on prediction error for unimodal columns. The first term corresponds to a rate achieved even if the estimator knows the true permutation Π^*; the second term quantifies the price paid for the combinatorial choice among $n!$ permutations. As a result, we see that if $m \gg \log n$, then the permutation does not play much of a role in the problem, and the rates resemble those of standard linear regression. Such a general behaviour is expected, since a large m means that we get multiple observations with the same unknown permutation, and this should allow us to estimate $\widehat{\Pi}$ better.

Clearly, a flat matrix is not influenced by the unknown permutation, and so the second term of the upper bound need not apply. As we demonstrate in the proof, it is likely that the flatness of A can also be incorporated in order to prove a tighter upper bound in this case, but we choose to state the upper bound as holding uniformly for all matrices A, with the loss of a logarithmic factor. It is also worth mentioning that the logarithmic factor in the second term is shown to be nearly tight for the problem of unimodal matrix estimation with an unknown permutation [9], suggesting that a similar factor may also appear in a tight version of our lower bound (5b). For the specific case where $m = 1$ however, which corresponds to the shuffled vector model (2), our bounds are tight up to constant factors, and summarized by the following corollary.

Corollary 1. *If $m = 1$ and $A \in \mathcal{A}(C_1, C_2/\sqrt{n})$, then*

$$c_2\sigma^2 \le \inf_{\substack{\widehat{\Pi}\in\mathcal{P}_n \\ \widehat{x}\in\mathbb{R}^d}} \sup_{\substack{\Pi^*\in\mathcal{P}_n \\ x^*\in\mathbb{R}^d}} \mathbb{E}\left[\frac{1}{n}\|\widehat{\Pi}A\widehat{x} - \Pi^*Ax^*\|_2^2\right] \le c_1\sigma^2.$$

In other words, the normalized minimax prediction error for the shuffled vector model does not decay with the parameters n or d, and so no estimator achieves consistent prediction for every parameter choice (Π^*, X^*). Again, this is a consequence of the fact that we do not get independent observations with the permutation staying fixed unlike when m is large, and herein lies the difficulty of the problem.

Both Theorem 1 and Corollary 1 provide non-adaptive minimax bounds. An interesting question is whether the least squares estimator is minimax optimal over finer classes of Π^* and X^*, i.e., whether it is adaptive in some interesting way. One would expect that the estimator adapts to the number of distinct entries in the matrix AX^*, similarly to the problem of monotone parameter recovery [9].

B. Polynomial time estimators

We now analyze a polynomial time estimator of the quantity Π^*AX^*, given by a singular value thresholding operation. In particular, given a matrix M having the singular value decomposition $M = \sum_{i=1}^r \sigma_i u_i v_i^\top$, its singular value thresholded version at level λ is given by $T_\lambda(M) = \sum_{i=1}^r \sigma_i \mathbb{I}(\sigma_i \ge \lambda) u_i v_i^\top$, where $\mathbb{I}(\cdot)$ is the indicator function of its argument.

The singular value thresholding (SVT) operation serves the purpose of denoising the observation matrix, and has been analyzed in the context of more general matrix estimation problems, e.g., Cai et al. [14] and Chatterjee [15].

Theorem 2. *For any choice of parameters Π^* and X^*, the SVT estimate with $\lambda = 1.1\sigma(\sqrt{n} + \sqrt{m})$ satisfies*

$$\frac{1}{nm}\|T_\lambda(Y) - \Pi^*AX^*\|_F^2 \le c_1\sigma^2\,\mathsf{rank}(A)\left(\frac{1}{n} + \frac{1}{m}\right) \quad (6a)$$

with probability greater than $1 - e^{-cnm}$.

Furthermore, provided $\mathsf{rank}(A) \le m$, there exist parameters Π_0 and X_0 (depending on the matrix A) such that for any value of the threshold λ, we have

$$\frac{1}{nm}\|T_\lambda(Y) - \Pi_0AX_0\|_F^2 \ge c_2\sigma^2\,\mathsf{rank}(A)\left(\frac{1}{n} + \frac{1}{m}\right), \quad (6b)$$

with probability greater than $1 - e^{-cnm}$.

Comparing inequalities (5b) (which holds for any denoised matrix, not just those having the form $\widehat{\Pi}A\widehat{X}$) and (6b), we see that the SVT estimator, while computationally efficient, may be statistically sub-optimal. However, it is consistent in the case where $\mathsf{rank}(A)$ is sufficiently small compared to m and n, and minimax optimal when $\mathsf{rank}(A)$ is a constant. Intuitively, the rate it attains is a result of treating the full matrix Π^*A as unknown, and so it is likely that better, efficient estimators exist that take the knowledge of A into account.

A concern is that the SVT estimator is required to know the noise variance σ^2. This can be taken care of via the square-root LASSO trick [16], which ensures a self-normalization that obviates the necessity for a noise-dependent threshold level. In particular, we define the estimate

$$\widehat{Y}_{\mathsf{sr}}(\lambda) = \arg\min_{Y'} \|Y - Y'\|_F + \lambda\|Y'\|_*. \quad (7)$$

Theorem 3. *If $\mathsf{rank}(A)\left(\frac{1}{n} + \frac{1}{m}\right) \le 1/20$, then for any choice of parameters Π^* and X^*, the square-root LASSO estimate (7) with $\lambda = 2.1\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)$ satisfies*

$$\frac{1}{nm}\|\widehat{Y}_{\mathsf{sr}}(\lambda) - \Pi^*AX^*\|_F^2 \le c_1\sigma^2\,\mathsf{rank}(A)\left(\frac{1}{n} + \frac{1}{m}\right)$$

with probability greater than $1 - 2e^{-cnm}$.

We prove Theorem 3 in the full version [13] for completeness. However, it should be noted that the square-root LASSO has been analyzed for matrix completion problems [17], and our proof follows similar lines. Klopp [17, Theorem 8] addresses the case of randomly observed matrix entries and requires these entries to be bounded, but Theorem 3 for the fully observed but noisy case has no such restriction. The condition $\mathsf{rank}(A)\left(\frac{1}{n} + \frac{1}{m}\right) \le 1/20$ does not significantly affect the claim, since our bounds no longer guarantee consistency of the estimate $\widehat{Y}_{\mathsf{sr}}(\lambda)$ when this condition is violated.

While the optimization problem (7) can be solved efficiently, there may be cases when the noise is (sub-)Gaussian of known variance for which the SVT estimate can be computed more quickly. Hence, the SVT estimator is usually preferred in cases where the noise statistics are known.

C. Exact algorithm for the noiseless case

For the noiseless model, the only efficient algorithm known up to now is for the special case $d = m = 1$ [1]. We provide the following spectral generalization of this special case, which

returns the exact parameters Π^* and X^* provided certain natural conditions are met. In order to define the conditions required for the theorem, we require a few definitions. The (left) *leverage score* vector $\ell(M)$ of a matrix M having the reduced singular value decomposition $M = U_M \Sigma_M V_M^\top$ is defined by the relation $\ell(M) = \mathrm{diag}(U_M U_M^\top)$, where $\mathrm{diag}(S)$ is the vector of diagonal entries of a square matrix S. We now introduce the LEVSORT algorithm:

(i) Compute the leverage scores $\ell(Y)$ and $\ell(A)$.

(ii) Return the permutation $\widehat{\Pi}_{\mathsf{lev}} \in \arg\min_\Pi \|\ell(Y) - \widehat{\Pi}_{\mathsf{lev}}\, \ell(A)\|_2^2$, and the matrix $\widehat{X}_{\mathsf{lev}} = \left(\widehat{\Pi}_{\mathsf{lev}} A\right)^\dagger Y$, where $M^\dagger$ denotes the Moore-Penrose pseudoinverse of M.

Note that this algorithm runs in polynomial time, since it involves only spectral computations and a matching step that can be computed in time $O(n \log n)$ (see [1, Theorem 4]).

Theorem 4. *Consider an instantiation of the noiseless model with* $\mathsf{rank}(A) \le \mathsf{rank}(X^*)$*, and such* $\ell(A)$ *and* $\ell(Y)$ *both have all distinct entries. Then the* LEVSORT *algorithm recovers the parameters* (Π^*, X^*) *exactly.*

The LEVSORT algorithm is a generalization of our own algorithm [1] for scalar X^*, to the matrix setting. However, instead of a simple sorting algorithm, we now require an additional spectral component. While showing the necessity of the condition $\mathsf{rank}(A) \le \mathsf{rank}(X^*)$ is still open, an efficient algorithm that does not impose any conditions is unlikely to exist due to the general problem being NP-hard [1]. Note that the condition includes as a special case all problems in which the matrices A and X^* are full rank, with $d \le m$.

In particular, the pose and correspondence estimation problem for 2D point clouds satisfies the conditions of Theorem 4 under some natural assumptions. We have $d = m = 2$ for all such problems, and $\mathsf{rank}(X^*) = 2$ unless the linear transformation is degenerate. Furthermore, unless the keypoints are generated adversarially, the leverage scores of the matrices A and Y are all distinct. Thus, assuming that the noiseless version of model (1) exactly describes the keypoints detected in the two images, we are guaranteed to find both the pose and the correspondence exactly.

In Figure 2, we demonstrate the guarantee of Theorem 4 on two image correspondence tasks when the keypoints detected in the two images are identical and the transformation between coordinates is linear.

D. Extensions: Dealing with duplicates and missing data

The results of Sections II-A and II-B also hold when the set of perturbations to the rows of the matrix A is allowed to be larger than just the set of permutation matrices $\mathcal{P}_n$. In particular, defining the set of "clustering matrices" $\mathcal{C}_n$ as

$$\mathcal{C}_n = \{D \in \{0,1\}^{n \times n} \ : D\mathbf{1} = \mathbf{1}\},$$

we consider an observation model of the form

$$Y = D^* A X^* + W, \tag{8}$$

where the matrices A, X^*, and W are as before, and $D^* \in \mathcal{C}_n$ now represents a clustering matrix. Such a clustering condition

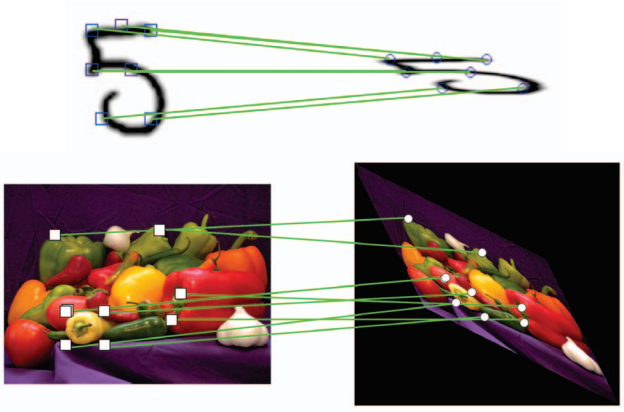

Fig. 2. Synthetic experiment illustrating exact pose and correspondence estimation by the LEVSORT algorithm. The right images are obtained via a linear tranformation of the coordinates of the respective left images, and keypoints are generated according to the noiseless model (1); keypoints are the same in the right and left image.

ensures stochasticity of the matrix D^* (not double stochasticity, as in the permutation model), and corresponds to the case where multiple responses may come from the same covariate, and some of the data may be permuted. Such a model is likely to better fit data from image correspondence problems when the keypoints detected in the two images are quite different. Also, such a formulation bears a superficial resemblance to the k-means clustering problem with Gaussian data [18].

As it turns out, Theorems 1, 2 and 3 also hold for this model, with minor modifications to the proofs. Defining the analogous MLE for this model as

$$\left(\widehat{D}_{\mathsf{ML}}, \widehat{X}_{\mathsf{ML}}\right) = \arg \min_{\substack{D \in \mathcal{C}_n \\ X \in \mathbb{R}^{d \times m}}} \|Y - DAX\|_F^2,$$

we have the following theorem.

Theorem 5. (a) *For any matrix* A*, and for all parameters* $D^* \in \mathcal{C}_n$ *and* $X^* \in \mathbb{R}^{d \times m}$*, we have*

$$\frac{\|\widehat{D}_{\mathsf{ML}} A \widehat{X}_{\mathsf{ML}} - D^* A X^*\|_F^2}{nm} \le c_1 \sigma^2 \left(\frac{\mathsf{rank}(A)}{n} + \frac{1}{m} \min\{\log n, m\}\right),$$

with probability greater than $1 - e^{-c(n \log n + m\, \mathsf{rank}(A))}$.

(b) *For any choice of parameters* D^* *and* X^**, the SVT estimate with* $\lambda = 1.1\sigma(\sqrt{n} + \sqrt{m})$ *satisfies*

$$\frac{1}{nm} \|T_\lambda(Y) - D^* A X^*\|_F^2 \le c_1 \sigma^2\, \mathsf{rank}(A) \left(\frac{1}{n} + \frac{1}{m}\right)$$

with probability greater than $1 - e^{-cnm}$.

(c) *If* $\mathsf{rank}(A)\left(\frac{1}{n} + \frac{1}{m}\right) \le 1/20$*, then for any choice of parameters* D^* *and* X^**, the square-root LASSO estimate* (7) *with* $\lambda = 2.1\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right)$ *satisfies*

$$\frac{1}{nm} \|\widehat{Y}_{\mathsf{sr}}(\lambda) - D^* A X^*\|_F^2 \le c_1 \sigma^2\, \mathsf{rank}(A) \left(\frac{1}{n} + \frac{1}{m}\right)$$

with probability greater than $1 - 2e^{-cnm}$.

Clearly, the lower bounds (5b) and (6b) hold immediately for the model (8) as a result of the inclusion $\mathcal{P}_n \subset \mathcal{C}_n$.

We conclude with a short proof sketch of claim (5b).

III. Proof Sketch of Lower Bound of Theorem 1

In this section, we sketch the proof of the minimax lower bound (5b). Full proofs of all our theorems may be found in the full version of the paper [13].

The bound (5b) follows from a packing set construction and Fano's inequality, which is a standard template used to prove minimax lower bounds [19, Chapter 15]. Suppose we wish to estimate a parameter θ over an indexed class of distributions $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ in the square of a (pseudo-)metric ρ. We refer to a subset $\{\theta^1, \theta^2, \ldots, \theta^M\}$ as a (δ, ϵ)-packing set if

$$\min_{i,j \in [M], i \neq j} \rho(\theta^i, \theta^j) \geq \delta \text{ and } \frac{1}{\binom{M}{2}} \sum_{i,j \in [M]} D(\mathbb{P}_{\theta^i} \| \mathbb{P}_{\theta^j}) \leq \epsilon.$$

Lemma 1 (Fano lower bound)**.** *If we can construct a (δ, ϵ)-packing set of cardinality M, then*

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta} \mathbb{E}\left[\rho(\widehat{\theta}, \theta^*)^2\right] \geq \frac{\delta^2}{2}\left(1 - \frac{\epsilon + \log 2}{\log M}\right).$$

The lower bound (5b) follows from the following claims.

$$\sup_{X^*} \mathbb{E}\left[\frac{\|\widehat{\Pi} A \widehat{X} - \Pi^* A X^*\|_F^2}{nm}\right] \geq \frac{c\sigma^2 \, \mathsf{rank}(A)}{n} \text{ for all } A, \text{ and} \tag{9a}$$

$$\sup_{\Pi^*} \mathbb{E}\left[\frac{\|\widehat{\Pi} A \widehat{X} - \Pi^* A X^*\|_F^2}{nm}\right] \geq \frac{c'\sigma^2}{m} \text{ if } A \in \mathcal{A}(C_1, C_2/\sqrt{n}). \tag{9b}$$

Claim (9a) follows from minimax lower bounds for linear regression [19]. Since we are operating in the matrix setting, we include the proof in the full version [13] for completeness.

We now prove claim (9b) for matrices in a smaller class than $\mathcal{A}(C_1, C_2/\sqrt{n})$; see the full version for an extension of this proof for the class $\mathcal{A}(C_1, C_2/\sqrt{n})$. We let $\mathbf{1}_n^p$ denote the n-dimensional vector having 1 in its first p coordinates and 0 in the remaining coordinates.

Now consider the class of matrices that have $\mathbf{1}_n^p$ in their range. By multiplying with δ and stacking m of these vectors up as columns, we have a matrix $\widetilde{Y}^1 \in \mathbb{R}^{n \times m}$ whose first p rows are identically δ and the rest are identically zero. Define the Hamming distance between two binary vectors $\mathsf{d_H}(u, v) = \#\{i : u_i \neq v_i\}$. We require the following lemma.

Lemma 2. *There exists a set of binary n-vectors $\{v_i\}_{i=1}^M$, each of Hamming weight p and satisfying $\mathsf{d_H}(v_i, v_j) \geq h$, having cardinality $M = \frac{\binom{n}{p}}{\sum_{i=1}^{\lfloor \frac{h-1}{2} \rfloor} \binom{n-p}{i}\binom{p}{i}}$.*

The lemma is proved in the full version [13].

Proof of claim (9b). Applying Lemma 2 and a rescaling argument, we see that there is a packing set $\{\Pi_i \widetilde{Y}^1\}_{i=1}^M$ obeying

$$\frac{1}{\sqrt{nm}}\|\Pi_i \widetilde{Y}^1\|_F = \delta\sqrt{\frac{p}{n}} \text{ for } i \in [M], \text{ and}$$
$$\frac{1}{\sqrt{nm}}\|\Pi_i \widetilde{Y}^1 - \Pi_j \widetilde{Y}^1\|_F \geq \delta\sqrt{\frac{h}{n}} \text{ for } i \neq j \in [M].$$

Fixing some constant $\gamma \in (0, 1)$ and choosing $p = \gamma n$ and $h = \frac{n}{2}\min\{\gamma, 1 - \gamma\}$, it can be verified that we obtain a packing set of size $M \geq e^{c'\gamma \log(1/\gamma) n}$. We now have observation i distributed as $\mathbb{P}_i = \mathcal{N}(\Pi_i \widetilde{Y}^1, \sigma^2 I_{nm})$, and so

$$D(\mathbb{P}_i \| \mathbb{P}_j) = \frac{1}{2\sigma^2}\|\Pi_i \widetilde{Y}^1 - \Pi_j \widetilde{Y}^1\|_F^2 \leq c\frac{\delta^2 \gamma nm}{\sigma^2}.$$

Finally, substituting into the Fano bound of Lemma 1 yields

$$\inf_{\substack{\widehat{\Pi} \in \mathcal{P}_n \\ \widehat{X} \in \mathbb{R}^{d \times m}}} \sup_{\substack{\Pi^* \in \mathcal{P}_n \\ X^* \in \mathbb{R}^{d \times m}}} \mathbb{E}\left[\frac{1}{nm}\|\widehat{\Pi} A \widehat{X} - \Pi^* A X^*\|_F^2\right] \geq \frac{\delta^2}{2}\left(1 - \frac{\frac{c\delta^2\gamma nm}{\sigma^2} + \log 2}{c'\gamma \log(1/\gamma) n}\right).$$

Setting $\delta^2 = c(\gamma)\frac{\sigma^2}{m}$ for a constant $c(\gamma)$ depending only on γ completes the proof provided the vector $\mathbf{1}_n^p \in \mathsf{range}(A)$ for $p = \gamma n$ with $\gamma \in (0, 1)$. □

References

[1] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Linear regression with an unknown permutation: Statistical and computational limits," in *Proceedings of the 54th Allerton Conference on Communication, Control, and Computing*, 2016.

[2] M. Marques, M. Stošić, and J. Costeira, "Subspace matching: Unique solution to point matching with geometric constraints," in *IEEE ICCV*. IEEE, 2009, pp. 1288–1294.

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

[4] 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.

[5] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Problems*, vol. 31, no. 11, p. 115002, 2015.

[6] E. J. Candés, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[7] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, 2011.

[8] 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.

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

[10] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright, "Stochastically transitive models for pairwise comparisons: Statistical and computational issues," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 934–959, 2017.

[11] 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.

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

[13] A. Pananjady, M. J. Wainwright, and T. A. Courtade, "Denoising linear models with permuted data," *arXiv preprint arXiv:1704.07461*, 2017.

[14] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

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

[16] A. Belloni, V. Chernozhukov, and L. Wang, "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, vol. 98, no. 4, pp. 791–806, 2011.

[17] O. Klopp, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[18] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward, "Relax, no need to round: Integrality of clustering formulations," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 2015, pp. 191–200.

[19] M. J. Wainwright, "High-dimensional statistics: A non-asymptotic viewpoint," *in preparation. University of California, Berkeley*, 2015.