Journal of Machine Learning Research 25 (2024) 1-38 Submitted 7/22; Revised 1/24; Published 2/24

Unlabeled Principal Component Analysis and Matrix Completion

Yunzhen Yao YUNZHEN.YAO@EPFL.CH
School of Computer and Communication Sciences
EPFL
CH-1015 Lausanne, Switzerland

Liangzu Peng LPENN@SEAS.UPENN.EDU

Manolis C. Tsakiris MANOLIS@AMSS.AC.CN
Key Laboratory for Mathematics Mechanization
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing, 100190, China

Editor: Michael Mahoney

Abstract

We introduce robust principal component analysis from a data matrix in which the entries of its columns have been corrupted by permutations, termed Unlabeled Principal Component Analysis (UPCA). Using algebraic geometry, we establish that UPCA is a well-defined algebraic problem since we prove that the only matrices of minimal rank that agree with the given data are row-permutations of the ground-truth matrix, arising as the unique solutions of a polynomial system of equations. Further, we propose an efficient two-stage algorithmic pipeline for UPCA suitable for the practically relevant case where only a fraction of the data have been permuted. Stage-I employs outlier-robust PCA methods to estimate the ground-truth column-space. Equipped with the column-space, Stage-II applies recent methods for unlabeled sensing to restore the permuted data. Allowing for missing entries on top of permutations in UPCA leads to the problem of unlabeled matrix completion, for which we derive theory and algorithms of similar flavor. Experiments on synthetic data, face images, educational and medical records reveal the potential of our algorithms for applications such as data privatization and record linkage.

Keywords: robust principal component analysis, matrix completion, record linkage, data re-identification, algebraic geometry

1. Introduction

In principal component analysis, a cornerstone of machine learning and data science, one is given a data matrix $\tilde{X}$, assumed to be a corrupted version of a ground-truth data matrix $X^* = [x_1^* \cdots x_n^*] \in \mathbb{R}^{m \times n}$, typically but not necessarily assumed to have low rank, and the objective is to estimate X^* or the column-space $S^* \subset \mathbb{R}^m$ of X^*. The most common types of corruptions that have attracted interest in modern studies are additive sparse perturbations (Candès et al., 2011; Zhang and Yang, 2018), outlier data points that lie away from S^*

0. A short version of this work has been published in NeurIPS 2021 (Yao et al., 2021a).

©2024 Yunzhen Yao, Liangzu Peng, Manolis C. Tsakiris.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/22-0816.html.

(Xu et al., 2012; Vaswani et al., 2018), missing entries, also known as low-rank —or even high-rank (Eriksson et al., 2012; Ongie et al., 2017, 2021)— matrix completion (Candès and Recht, 2009; Ganti et al., 2015; Balzano et al., 2018; Eftekhari et al., 2019; Bertsimas and Li, 2020), and entry-wise non-linear corruption (El Karoui, 2010; Yao et al., 2021b; Guionnet et al., 2023; Mergny et al., 2024).

Recently, permutations have been emerging as another type of data corruption, typically set in the context of linear regression, where the correspondences between the input and the output data have been partially distorted or are even entirely unavailable (Unnikrishnan et al., 2015, 2018; Hsu et al., 2017; Slawski and Ben-David, 2019; Slawski et al., 2020; Zhang and Li, 2020; Marano and Willett, 2020; Wang et al., 2020; Tsakiris et al., 2020; Mazumder and Wang, 2023; Peng et al., 2022; Onaran and Villar, 2022; Azadkia and Balabdaoui, 2022). There, one is given a point x^* of a linear subspace S^*, but only up to a permutation of its coordinates, say $\tilde{x} = \Pi^* x^*$ with Π^* an unknown permutation, and the goal is to find x^* from the data $\tilde{x}, S^*$. An alternative formulation for this problem is that given a matrix $A \in \mathbb{R}^{m\times r}$, which can be regarded as a basis of the linear subspace S^*, and a response vector $\tilde{x} = \Pi^* A c^*$ shuffled by an unknown permutation Π^*, the goal is to find Π^* and the regression coefficients c^*. This *Unlabeled Sensing* (Unnikrishnan et al., 2015, 2018) problem has many potential applications, e.g., record linkage (Slawski and Ben-David, 2019; Slawski et al., 2020), visual (Santa Cruz et al., 2017, 2018) or textual (Brown et al., 1990; Schmaltz et al., 2016; Shen et al., 2017) permutation learning, matching problems in neuroscience (Nejatbakhsh and Varol, 2021) and biology (Abid and Zou, 2018; Ma et al., 2021; Xie et al., 2021), and DNA-based data storage (Shomorony and Heckel, 2021; Weinberger and Merhav, 2022; Lenz et al., 2022; Ravi et al., 2022).

While methods for unlabeled sensing rely on knowledge of the source subspace S^*, this is not always known in practice. On the other hand, data of the form $\tilde{X} = [\tilde{x}_1, \dots, \tilde{x}_n] \in \mathbb{R}^{m\times n}$ with $\tilde{x}_j = \Pi_j^* x_j^*$ an unknown permutation of an unknown point $x_j^* \in S^*$, are often available, thus raising the question of whether S^* can be estimated from $\tilde{X}$. An important example of this situation is record linkage (Fellegi and Sunter, 1969; Muralidhar, 2017; Antoni and Schnell, 2019; Nikolaenko et al., 2013; Ong et al., 2014; Ranbaduge et al., 2021), where the objective is to integrate data from independent sources, $\tilde{x}_1, \dots, \tilde{x}_n \in \mathbb{R}^m$, for subsequent data analysis. Since the entries of different records $\tilde{x}_i$'s are collected separately, the data matrix $\tilde{X}$ is *unlabeled* in the sense that, the entries of its i-th row do not necessarily correspond to the same entity. Such kind of unlabeled data $\tilde{X}$ also arise in the context of data privatization, where the data provider anonymizes the original data X^* by permuting each column of X^* prior to release (Domingo-Ferrer and Muralidhar, 2016; He et al., 2012; Nikolaenko et al., 2013). In such a way, the marginal distribution of each corresponding original attribute is preserved, and hence so are the statistical properties such as the mean, the median, and the variance of the data. Data re-identification is a concern, since companies with privacy policies, health care providers, and financial institutions may release the collected data after anonymization. Understanding the fundamental limits of re-identifying the original data X^* from the released data $\tilde{X}$ is essential for striking a balance between data privacy and data preservation (Abowd, 2019; Narayanan and Shmatikov, 2008). Applications with unlabeled data also arise in the multiple-image correspondence problem (Zeng et al., 2012; Ji et al., 2014) in image processing and computer vision. Oliveira et al. (2005) showed that estimating

the correspondence of points across a sequence of images of a single rigid body motion, can be expressed as a rank-minimization problem in terms of partial permutation matrices.

1.1 Related Work

In this section, we briefly review some existing work on three problems that interconnect in this paper; that is, unlabeled sensing, robust principal component analysis with outliers, and matrix completion with outliers. Beyond them, we also mention the recent and related works of Breiding et al. (2018, 2023) and Tachella et al. (2023), that combine flavors from data science, inverse problems, and algebraic models.

1.1.1 Unlabeled Sensing

There is a large literature on application-specific problems that involve lack of correspondences, e.g. in computer vision or statistics; here we just review four recent methods for unlabeled sensing (Unnikrishnan et al., 2018) that will be used in this paper. In unlabeled sensing one is given a subspace $S^* \subset \mathbb{R}^m$ of dimension r and a point $\tilde{x}$ which is some permuted version of a point $x^* \in S^*$ and the goal is to recover x^* from S^* and $\tilde{x}$. A critical distinction among methods in the literature is the sparsity level α of the permutation, that is the ratio of coordinates that are moved by the permutation.

The case of *dense* permutations ($\alpha = 1$) is extremely challenging, with existing methods only able to handle small ranks r. We consider two methods known to perform best in this regime. The algebraic-geometric method called AIEM in Tsakiris et al. (2020) has linear complexity in m, and instead concentrates its effort on solving a polynomial system of r equations in r variables to produce an initialization for an expectation maximization algorithm. Currently, this method is efficient for $r \leq 5$ and intractable otherwise. A very different method is CCV-Min of Peng and Tsakiris (2020), which proceeds via branch-and-bound together with concave minimization and can handle $r \leq 8$, though intractable otherwise. For a picture regarding the computational complexity of existing unlabeled sensing methods, we refer to the discussion in Peng and Tsakiris (2020).

For *sparse* permutations (small α) we review two methods (Slawski and Ben-David, 2019; Slawski et al., 2021). The ℓ_1-RR algorithm of Slawski and Ben-David (2019) applies an ℓ_1 robust linear regression relaxation and it works when $\alpha \leq 0.5$. Another approach is the Pseudo-Likelihood method (PL) of Slawski et al. (2021), which fits a two-component mixture density for each entry of $\tilde{x}$, one accounting for fixed data and the other for permuted data. The fitting is done via a combination of hypothesis testing, reweighted least-squares, and alternating minimization; while this method works well for $\alpha \leq 0.7$, it is sensitive to the particular basis of S^* used to generate $\tilde{x}$.

The unlabeled sensing problem has in fact been explored, at least theoretically, towards greater generality (Unnikrishnan et al., 2015, 2018; Dokmanić, 2019; Tsakiris and Peng, 2019; Peng and Tsakiris, 2021; Tsakiris, 2023a). In one such extended setting, already present in (Unnikrishnan et al., 2015, 2018), we are only given a subset of coordinates of $\tilde{x}$ (and the subspace S^*) and we aim to recover x^*. We call this problem *unlabeled sensing with missing entries*. It is a more challenging problem for which very few algorithms exist; e.g. see Elhami et al. (2017); Tsakiris and Peng (2019). Tsakiris and Peng (2019) proposes two algorithms: Algorithm-A is based on a combination of branch-and-bound and a dynamic

programming strategy; Algorithm-B is a RANSAC-style method based also on dynamic programming computation. Both algorithms perform well, if the subspace dimension r is sufficiently small (e.g., ≤ 3) and if one is given sufficiently many entries of $\tilde{x}$.

1.1.2 Robust PCA with Outliers

PCA methods with robustness to outliers will also play a role in this paper. Among a large literature we review four state-of-the-art methods inspired by sparse (You et al., 2017), cosparse (Tsakiris and Vidal, 2018b) and low-rank (Xu et al., 2012) (Rahmani and Atia, 2017) representations. In that context, $\tilde{X}$ can be partitioned into inlier points that lie in an unknown r-dimensional subspace $S^* \subset \mathbb{R}^m$ and outlier points that lie away from S^*; the goal is to recover S^* from $\tilde{X}$.

A successor of Soltanolkotabi and Candés (2012), the convex method of You et al. (2017), which we refer to as Self-Expr, solves a self-expressive elastic net problem so that each $\tilde{x}_j$ is expressed as an ℓ_2-regularized sparse linear combination of the other points. Inlier points need approximately r other inliers for their self-expression as opposed to about m points for outliers. The self-expressive coefficients are used to define transition probabilities of a random walk on the self-representation graph and the average of the t-step transition probability distributions for $t = 1, \ldots, T$ is used as a score for inliers vs outliers, with higher scores expected for the former. Then $\hat{S}$ is taken to be the subspace spanned by the r top $\tilde{x}_j$'s.

Dual Principal Component Pursuit (DPCP) of Tsakiris and Vidal (2015, 2018b) solves a non-smooth non-convex problem for an orthonormal basis B^* of the orthogonal complement of S^*. In contrast to other robust-PCA methods, particularly those based on convex optimization, DPCP was shown in Zhu et al. (2018); Ding et al. (2021) to tolerate as many outliers as the square of the number of inliers, under a relative spherically uniform distribution assumption on inliers and outliers. This assumption is certainly not true for outliers obtained by permuting the coordinates of inliers, but we will experimentally see that an even stronger property holds for the case of UPCA (Figure 5).

The now classical outlier pursuit method of Xu et al. (2012), which we refer to as OP, decomposes via convex optimization $\tilde{X}$ into the sum of a low-rank matrix, representing the inliers, and a column-sparse matrix, representing the outliers. $\hat{S}$ is obtained as the rth principal component subspace of the low-rank part. Finally, the Coherence Pursuit (CoP) method of Rahmani and Atia (2017) is based on the following simple but effective principle: with $\tilde{X}_{-j}$ the matrix $\tilde{X}$ with column j removed, for each $\tilde{x}_j$ one computes its coherence $\tilde{X}_{-j}^\top \tilde{x}_j$ with the rest of the points. As it turns out, inliers tend to have coherences of higher ℓ_2-norm than outliers, and the r top $\tilde{x}_j$'s are taken to span $\hat{S}$.

1.1.3 Matrix Completion with Outliers

As mentioned in the introduction, matrix completion—or robust PCA with missing entries—has been a well-studied problem, with numerous developed theories (Candès and Recht, 2009; Candès and Plan, 2010; Singer and Cucuringu, 2010; Eriksson et al., 2012; Balcan et al., 2019; Tsakiris, 2023b) and algorithms (Cai et al., 2010; Keshavan et al., 2010; Balzano et al., 2010; Majumdar and Ward, 2011; Tanner and Wei, 2013; Bertsimas and Li, 2020); see, e.g., Davenport and Romberg (2016); Vaswani and Narayanamurthy (2018) for a survey.

Closely related to this paper is a more general setting of matrix completion, where the given data matrix $\tilde{\mathsf{X}}$ not only has some entries missing but also some of its columns are outliers. This setup is more challenging, and research on it is relatively scarce. Chen et al. (2011, 2015) considered this problem, and proposed a convex program that minimizes a combination of a nuclear norm with an $\ell_{1,2}$ norm over a matrix of variables. Their method, which we call MCO, is shown to succeed for sufficiently many inliers and observed entries. We will make use of MCO later, to solve our *unlabeled matrix completion* problem.

1.2 Contributions

In this paper, we consider the recovery of X^* from its unlabeled version $\tilde{\mathsf{X}}$, which we term *Unlabeled Principal Component Analysis* (UPCA). We take one step further and generalize UPCA into *Unlabeled Matrix Completion* (UMC), where we now need to recover X^* from only a subset of entries of $\tilde{\mathsf{X}}$. We make contributions in the following three aspects:

1. Theoretical contributions (Section 2)

   (a) We establish that as long as $\mathsf{r} := \operatorname{rank}(\mathsf{X}^*) < \min\{\mathfrak{m}, \mathfrak{n}\}$ and X^* is *generic* (see Definition 1), then up to a permutation of its rows, X^* is the only matrix of rank less than or equal to r that is compatible with $\tilde{\mathsf{X}}$. This asserts that UPCA is a well-posed problem, since the inherent ambiguity of whether $\tilde{\mathsf{X}}$ comes from X^* or a row-permuted version of X^* is in most cases practically harmless (Sections 2.1.1 and 2.1.3).

   (b) We establish that in this basic formulation, UPCA is a purely algebraic problem, by exhibiting a polynomial system of equations parametrized by $\tilde{\mathsf{X}}$, whose solutions are all the row-permutations of X^*; solving the UPCA problem amounts to obtaining one such solution (Section 2.1.4).

   (c) We furthermore generalize our UPCA theorems for UMC, thereby obtaining results of similar "information-theoretical" flavor for the scenario with permuted incomplete data (Section 2.2).

2. Algorithmic contributions (Section 3)

   (a) Inasmuch as solving the polynomial system of UPCA is in principle NP-hard, we introduce an efficient algorithmic pipeline, Algorithm 1, for the practically relevant case where a significant part of the data have undergone the same *dominant permutation*, while the rest of the points have been permuted arbitrarily (see Section 2.1.5); in the case of record linkage this would correspond to one of the records having much larger size than the others. The first stage of the pipeline employs PCA methods with robustness to outliers (Xu et al., 2012; Soltanolkotabi and Candés, 2012; Rahmani and Atia, 2017; You et al., 2017; Tsakiris and Vidal, 2018b; Zhu et al., 2018; Lerman and Maunu, 2018) to produce an estimate $\hat{\mathsf{S}}$ of S^* from $\tilde{\mathsf{X}}$; the second stage of the pipeline uses unlabeled sensing methods (Slawski and Ben-David, 2019; Slawski et al., 2021; Tsakiris et al., 2020; Peng and Tsakiris, 2020) to furnish an estimate $\hat{\mathsf{X}}$ of X^* from $\hat{\mathsf{S}}$ and $\tilde{\mathsf{X}}$ (See Algorithm 1). Moreover, we introduce a simple but efficient algorithm for unlabeled sensing, Algorithm 2, based on least-squares with recursive filtration.

(b) Our algorithmic development for UMC is parallel to that of UPCA. We start with the dominant permutation assumption and introduce a two-stage algorithmic pipeline (Algorithm 3). The first stage detects and completes inlier columns, and then estimates the subspace S^* by matrix completion with column outliers (recall Section 1.1.3). The second stage estimates the data matrix X^* by solving the problem of unlabeled sensing on the projected coordinates for each column.

3. Experimental evaluation (Section 4)

(a) We assess our algorithmic pipeline for UPCA and our proposed unlabeled sensing algorithm on synthetic data (Section 4.1.1 and 4.1.2), face images (Section 4.1.3), educational and medical records (Section 4.1.4), with encouraging results.

(b) We also perform experiments for the proposed UMC pipeline (Section 4.2).

2. Theoretical Foundations

In this section, we formulate and study two problems, unlabeled principal component analysis (UPCA) and unlabeled matrix completion (UMC). The goal of UPCA is to recover a ground-truth rank-deficient matrix X^* from its unlabeled version $\tilde{X}$. The goal of UMC is to also recover X^* from $\tilde{X}$, but now $\tilde{X}$ is a partial observation of a permuted version of X^*. See Figure 1 for an intuitive understanding of the setup; we will formalize the settings soon.

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{pmatrix} \qquad \begin{pmatrix} x_{31} & x_{22} & x_{23} & x_{44} \\ x_{11} & x_{32} & x_{43} & x_{34} \\ x_{21} & x_{42} & x_{33} & x_{14} \\ x_{41} & x_{12} & x_{13} & x_{24} \end{pmatrix} \qquad \begin{pmatrix} x_{31} & * & x_{23} & x_{44} \\ * & x_{32} & x_{43} & * \\ x_{11} & x_{22} & * & x_{14} \\ x_{41} & * & x_{13} & x_{34} \end{pmatrix}$$

(a) Ground-truth X^* (b) UPCA data matrix $\tilde{X}$ (c) UMC data matrix $\tilde{X}$

Figure 1: (1a): Ground-truth matrix X^*. (1b): Data matrix $\tilde{X}$ for UPCA, obtained by shuffling each column of X^* via some unknown permutation. (1c): Data matrix $\tilde{X}$ for UMC, obtained via removing some entries (indicated by $*$) and shuffling every column of X^*. In both UPCA and UMC, we need to recover X^* from data $\tilde{X}$.

2.1 Unlabeled Principal Component Analysis

In this section, we first introduce the problem formulation of the unlabeled principal component analysis problem, and then we give necessary preliminaries on algebraic geometry, followed by the result that UPCA is well-posed. Next, we show UPCA is a purely algebraic problem in this basic formulation. Finally, we discuss a special case where many data have undergone the same dominant permutation.

2.1.1 Problem Formulation

Let us denote by $\mathcal{P}_m$ the set of all permutations of coordinates of $\mathbb{R}^m$. We let $X^* = [x_1^* \cdots x_n^*] \in \mathbb{R}^{m\times n}$ be our ground-truth data matrix with rank $r < \min\{m, n\}$ and column

space $S^* = \mathcal{C}(X^*)$, and we suppose that the available data are

$$\tilde{X} = [\tilde{x}_1 \cdots \tilde{x}_n] = [\Pi_1^* x_1^* \cdots \Pi_n^* x_n^*] \in \mathbb{R}^{m \times n}, \tag{1}$$

where each $\Pi_j^* \in \mathcal{P}_m$ is an unknown permutation. Let $\mathcal{P}_m^n = \prod_{i \in [n]} \mathcal{P}_m$ be n ordered copies of $\mathcal{P}_m$, where $[n] = \{1, \ldots, n\}$. For $\underline{\pi} = (\Pi_1, \ldots, \Pi_n) \in \mathcal{P}_m^n$ we set $\underline{\pi}(\tilde{X}) = [\Pi_1 \tilde{x}_1 \cdots \Pi_n \tilde{x}_n]$. We pose Unlabeled Principal Component Analysis (UPCA) as the following rank minimization problem:

$$\min_{\underline{\pi} \in \mathcal{P}_m^n} \operatorname{rank} \underline{\pi}(\tilde{X}) \tag{2}$$

First, note that (2) never has a unique solution, because if $\underline{\pi} = (\Pi_1, \ldots, \Pi_n)$ is a solution, then so is $\underline{\pi}' = (\Pi\Pi_1, \ldots, \Pi\Pi_n)$ for every permutation $\Pi \in \mathcal{P}_m$. This reveals an inherent ambiguity of UPCA: it is only possible to recover X^* from $\tilde{X}$ up to a permutation ΠX^* of its rows. On the other hand, this is rather harmless in many situations, since ΠX^* is the same dataset as X^* except that the row-features appear now in some different order. Thus, our hope in formulating (2) is that the only solutions are of the form $\underline{\pi} = (\Pi\Pi_1^{*\top}, \ldots, \Pi\Pi_n^{*\top})$ with Π ranging across $\mathcal{P}_m$ and Π_j^* as in (1). However, without any other assumptions on the data X^*, there could in principle be additional undesired permutations that also give $\operatorname{rank} X^*$, or even worse, the minimum rank in (2) could be lower than $r = \operatorname{rank} X^*$. Our results show that for *generic* enough data, such pathological situations do not occur, and the only solutions to (2) are the ones associated with row-permutations of X^*.

2.1.2 Elements of Algebraic Geometry

Before stating our results, we make the notion of *generic* precise using some basic algebraic geometry (Cox et al., 2013; Harris, 2013). Let $Z = (z_{ij})$ be an $m \times n$ matrix of variables z_{ij} and $\mathbb{R}[Z] = \mathbb{R}\big[z_{ij} : i \in [m], j \in [n]\big]$ the ring of polynomials in the z_{ij}'s with real coefficients. An *algebraic variety* of $\mathbb{R}^{m \times n}$ is the set of solutions of a polynomial system of equations in $\mathbb{R}[Z]$. In particular, the set of $(r+1) \times (r+1)$ determinants of Z are polynomials in z_{ij}'s of degree $r+1$ and define the algebraic variety

$$\mathcal{M}_r = \{\, X \in \mathbb{R}^{m \times n} | \operatorname{rank} X \leq r \,\},$$

since $\operatorname{rank} X \leq r$ if and only if all $(r+1) \times (r+1)$ determinants of X are zero.

The algebraic variety $\mathcal{M}_r$ admits a topology, called *Zariski topology*, which makes it convenient to work with. The closed sets in this topology are the *algebraic subvarieties* of $\mathcal{M}_r$. These are sets of matrices of rank $\leq r$, which in addition satisfy certain other polynomial equations in $\mathbb{R}[Z]$. For example, the set of matrices of rank at most $r-1$ is a proper closed subset of $\mathcal{M}_r$, because in addition to the equations defining $\mathcal{M}_r$, it is further defined by requiring all $r \times r$ determinants to be zero. *Open sets* in $\mathcal{M}_r$ are defined as complements of closed sets, or equivalently they are defined by requiring that certain sets of polynomials are not all simultaneously zero. For example, the set of matrices of rank exactly equal to r is a proper open subset of $\mathcal{M}_r$ defined by the non-simultaneous vanishing of all $r \times r$ determinants of Z; a matrix has rank r if and only if all $(r+1) \times (r+1)$ determinants are zero and least one $r \times r$ determinant is non-zero. Now, the algebraic variety $\mathcal{M}_r$ is *irreducible* in the sense that it can not be described as the union of two proper algebraic

subvarieties of it (Kleiman and Landolfi, 1971). A consequence of this is that non-empty open sets of $\mathcal{M}_r$ have the very important property of being topologically dense. This means that given a non-empty open set $\mathcal{U} \subset \mathcal{M}_r$ and a point $X \in \mathcal{M}_r$, every neighborhood of X intersects $\mathcal{U}$. It follows that under any non-degenerate continuous probability measure on $\mathcal{M}_r$, a non-empty Zariski-open set of $\mathcal{M}_r$ has measure 1. For example, the set of matrices in $\mathcal{M}_r$ of rank r is non-empty and open, and thus it is dense. Hence a randomly sampled matrix in $\mathcal{M}_r$ under a continuous probability measure will have rank r with probability 1. We refer to such a fact by saying that a generic matrix in $\mathcal{M}_r$ has rank r. More generally:

Definition 1 *We say that a* generic *matrix in $\mathcal{M}_r$ satisfies a property, if the property is true for every matrix in a non-empty open subset of $\mathcal{M}_r$.*

2.1.3 UPCA is a Well-Posed Problem

Our first theoretical result is as follows.

Theorem 2 *For X^* a generic matrix in $\mathcal{M}_r$, we have that* $\operatorname{rank} \underline{\pi}(\tilde{X}) \geq r$ *for any $\underline{\pi} \in \mathcal{P}_m^n$, with equality if and only if $\underline{\pi}(\tilde{X}) = \Pi X^*$ for some $\Pi \in \mathcal{P}_m$.*

Theorem 2 says that for $X^* \in \mathcal{M}_r$ generic, and up to a permutation of the coordinates of $\mathbb{R}^m$, S^* is the unique r-dimensional subspace that explains the data $\tilde{X}$ in the UPCA sense, and $r = \operatorname{rank} X^*$ is the minimum objective in (2).

2.1.4 UPCA is an Algebraic Problem

How can one go about solving the discrete optimization problem (2)? In general, brute force selection of the Π_j's has complexity $\mathcal{O}\big((m!)^n\big)$, which is out of the question. On the other hand, problem (2) has a rich algebraic structure, which allows us to show that X^*, up to a permutation of its rows, is the unique solution to a polynomial system of equations.

To begin with, for each $j \in [n]$ and each $\ell \in [m]$, we define the following column-symmetric polynomials of $\mathbb{R}[Z]$:

$$\bar{p}_{\ell,j}(Z) := \sum_{i \in [m]} z_{ij}^{\ell}, \qquad p_{\ell,j}(Z) := \bar{p}_{\ell,j}(Z) - \bar{p}_{\ell,j}(\tilde{X})$$

Note that $\bar{p}_{\ell,j}\big(\underline{\pi}(Z)\big) = \bar{p}_{\ell,j}(Z)$ for any $\underline{\pi} \in \mathcal{P}_m^n$ and thus $\bar{p}_{\ell,j}(\tilde{X}) = \bar{p}_{\ell,j}(X^*)$. Now let us think of $X \in \mathcal{M}_r$ as a product of two matrices of size $m \times r$ and $r \times n$, and let us define another polynomial ring with variables associated to these two factors. For $i = r+1, \ldots, m$, and $k \in [r]$ and $j \in [n]$, we let b_{ik}, c_{kj} be a new set of variables over $\mathbb{R}$. Organize the b_{ik}'s to occupy the $(m-r) \times r$ bottom block of an $m \times r$ matrix B whose top $r \times r$ block is the identity matrix of size r, and the c_{kj}'s into a $r \times n$ matrix $C = (c_{kj})$. For $i \in [m]$, we write $b_i^\top$ for the i-th row of B; for $j \in [n]$, we write c_j for the j-th column of C. With $\tilde{x}_{ij}, x_{ij}^*$ the i-th coordinate of $\tilde{x}_j, x_j^*$ respectively, we obtain polynomials $q_{\ell,j}$ for $\ell \in [m]$, $j \in [n]$ of

$\mathbb{R}[B, C]$ by substituting $z_{ij} \mapsto b_i^\top c_j$ in the $p_{\ell,j}(Z)$'s above:

$$\begin{aligned} q_{\ell,j}(B, C) &:= \bar{p}_{\ell,j}(BC) - \bar{p}_{\ell,j}(\tilde{X}) \\ &= \sum_{i\in[m]} (b_i^\top c_j)^\ell - \sum_{i\in[m]} \tilde{x}_{ij}^\ell \\ &= \sum_{i\in[m]} (b_i^\top c_j)^\ell - \sum_{i\in[m]} {x_{ij}^*}^\ell \end{aligned}$$

The set of common roots of all $q_{\ell,j}$'s is an algebraic variety $\mathcal{Y}_{X^*}$ that depends only on X^*:

$$\mathcal{Y}_{X^*} = \big\{(B', C') \in \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n} \,|\, q_{\ell,j}(B', C') = 0,\ \forall \ell \in [m],\ \forall j \in [n];\ \ B'_{[r],[r]} = I_r\big\}$$

Here, $B'_{[r],[r]} = I_r$ signifies that the top $r \times r$ block of $B' \in \mathbb{R}^{m\times r}$ is the identity matrix. Then, with $\Pi \in \mathcal{P}_m$, if the column-space $\mathcal{C}(\Pi X^*)$ of ΠX^* does not drop dimension upon projection onto the first r coordinates, then there exists a unique basis B^*_Π of $\mathcal{C}(\Pi X^*)$ with the identity matrix occurring at the top $r \times r$ block. In that case, there is a unique factorization $\Pi X^* = B^*_\Pi C^*_\Pi$ and the point (B^*_Π, C^*_Π) lies in the variety $\mathcal{Y}_{X^*}$ because

$$\begin{aligned} q_{\ell,j}(B^*_\Pi, C^*_\Pi) &= \bar{p}_{\ell,j}(B^*_\Pi C^*_\Pi) - \bar{p}_{\ell,j}(\tilde{X}) \\ &= \bar{p}_{\ell,j}(\Pi X^*) - \bar{p}_{\ell,j}(X^*) \\ &= \bar{p}_{\ell,j}(X^*) - \bar{p}_{\ell,j}(X^*) = 0. \end{aligned}$$

Our second result says that if X^* is generic, then all points of $\mathcal{Y}_{X^*}$ are of this type. That is, they correspond to factorizations $B^*_\Pi C^*_\Pi$ of ΠX^* as Π varies across all permutations:

Theorem 3 *For a generic matrix X^* in $\mathcal{M}_r$ we have*

$$\mathcal{Y}_{X^*} = \big\{(B^*_\Pi, C^*_\Pi) \in \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n} \,|\, \Pi \in \mathcal{P}_m;\ B^*_{\Pi,[r],[r]} = I_r;\ \Pi X^* = B^*_\Pi C^*_\Pi\big\}$$

Thanks to Theorem 3 we have the following important conceptual finding. Assuming X^* is generic, to obtain X^* up to some permutation of its rows from $\tilde{X}$, one only needs to compute an arbitrary root (B', C') of the polynomial system of equations

$$q_{\ell,j}(B, C) = 0,\ \forall \ell \in [m],\ \forall j \in [n] \tag{3}$$

and multiply its factors to get $B'C'$. Developing a polynomial system solver for UPCA would involve two main challenges: attaining robustness to noise and scalability. We leave such an endeavor to future research.

2.1.5 UPCA with Dominant Permutations

In this section we consider a special case of interest, where part of the data have undergone the same *dominant* permutation (see Figure 2). To make this precise, we define the multiplicity $\mu(\Pi)$ of a permutation $\Pi \in \mathcal{P}_m$ to be the number of times that Π appears as $\Pi = \Pi^*_j$ in (1) with j ranging in $[n]$. Figure 2c shows an example for the case $\mu(\Pi^*_1) = 3$ and $\mu(\Pi^*_3) = 1$. In fact, given the inherent ambiguity of UPCA discussed above, we may as well take this dominant permutation to be the identity matrix I_m of size $m \times m$. We have:

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{pmatrix}$$

(a) Ground-truth X^*

$$\begin{pmatrix} x_{31} & x_{22} & x_{23} & x_{44} \\ x_{11} & x_{32} & x_{43} & x_{34} \\ x_{21} & x_{42} & x_{33} & x_{14} \\ x_{41} & x_{12} & x_{13} & x_{24} \end{pmatrix}$$

(b) UPCA data matrix $\tilde{X}$

$$\begin{pmatrix} x_{41} & x_{42} & \mathbf{x_{23}} & x_{44} \\ x_{11} & x_{12} & \mathbf{x_{43}} & x_{14} \\ x_{21} & x_{22} & \mathbf{x_{33}} & x_{24} \\ x_{31} & x_{32} & \mathbf{x_{13}} & x_{34} \end{pmatrix}$$

(c) UPCA data matrix with a dominant permutation

Figure 2: (2a): Ground-truth matrix X^*. (2b): Data matrix $\tilde{X}$ for UPCA, obtained by shuffling each column of X^* via some unknown permutation. (2c): Data matrix $\tilde{X}$ for UPCA with a dominant permutation, obtained via shuffling some columns (columns 1, 2, 4 in the figure) by the same permutation and shuffling others arbitrarily (column 3 highlighted in bold).

Theorem 4 *Suppose that* $\mu(I_m) > \max\{\mu(\Pi), r\}$ *for any other* $\Pi \neq I_m$. *Then for a generic* $X^* \in \mathcal{M}_r$, *we have that* S^* *is the unique solution to the following consensus maximization problem*

$$\max_{\dim S \leq r} \ \#\{\tilde{x}_j \,|\, \tilde{x}_j \in S\,;\, j \in [n]\}, \tag{4}$$

where # *denotes the cardinality of a set, and the maximization is taken over all subspaces* $S \subset \mathbb{R}^m$ *of dimension* $\leq r$.

Theorem 4 says that for sufficiently generic ground-truth data X^*, the given data $\tilde{X}$ admit a natural partition into a set of inliers and outliers with respect to the linear subspace S^*:

$$\tilde{X}_{\text{in}} := \{\tilde{x}_j \,|\, \tilde{x}_j \in S^*\}, \quad \tilde{X}_{\text{out}} := \{\tilde{x}_j \,|\, \tilde{x}_j \notin S^*\}$$

Of course we do not know what the partition into inliers and outliers is, because we do not know what S^* is. But the presence of this geometric structure is enough for PCA methods with robustness to outliers to operate on $\tilde{X}$ in order to estimate S^*. Section 3.1 proceeds algorithmically building on this insight.

2.2 Unlabeled Matrix Completion

A generalization of UPCA with practical significance is to consider PCA from data corrupted by both permutations and missing entries. To proceed we need some extra notations.

With ω_j a subset of $[m]$ we let $P_{\omega_j} \in \mathbb{R}^{m\times m}$ be the matrix representing the projection of $\mathbb{R}^m$ onto the coordinates contained in ω_j, that is P_{ω_j} is a diagonal matrix with the kth diagonal element non-zero and equal to 1 if and only if $k \in \omega_j$. With ω_j as above for every $j \in [n]$, we write $\Omega = \bigcup_{j\in[n]} \omega_j \times \{j\} \subset [m] \times [n]$ and $\underline{p}_\Omega = (P_{\omega_1}, \ldots, P_{\omega_n})$. Let $\mathbb{R}^\Omega$ be the subspace of $\mathbb{R}^{m\times n}$ of all matrices that have zeros in the complement of Ω. The association $X = [x_1 \cdots x_n] \mapsto \underline{p}_\Omega(X) = [P_{\omega_1}x_1 \cdots P_{\omega_n}x_n]$ induces a map

$$\underline{p}_\Omega : \mathcal{M}_r \longrightarrow \mathbb{R}^\Omega$$

With this notation, in ordinary bounded-rank matrix completion (of which low-rank matrix completion is a special case) one is given a partially observed matrix $\underline{p}_\Omega(X^*)$ and the

objective is to compute an at most rank-r completion, that is an element of the *fiber*

$$\underline{p}_\Omega^{-1}\big(\underline{p}_\Omega(X^*)\big) = \big\{X \in \mathcal{M}_r \,|\, \underline{p}_\Omega(X) = \underline{p}_\Omega(X^*)\big\}$$

A big question is to characterize the observation patterns Ω for which $\underline{p}_\Omega(X^*)$ is generically finitely completable, in the sense that there exists a dense open set $\mathcal{U}$ of $\mathcal{M}_r$ such that for every $X^* \in \mathcal{U}$ the fiber $\underline{p}_\Omega^{-1}\big(\underline{p}_\Omega(X^*)\big)$ is a finite set. Even harder is the characterization of the Ω's that are generically uniquely completable, i.e. the fiber consists only of X^*. Both of these questions remain open in their generality, while several authors have made progress from different points of view, including rigidity theory (Singer and Cucuringu, 2010), algebraic combinatorics (Király and Tomioka, 2012; Király et al., 2015), tropical geometry (Bernstein, 2017) and algebraic geometry (Tsakiris, 2023b, 2024).

Next, we let $\mathcal{P}_{\omega_j}$ be the permutations $\Pi \in \mathcal{P}_m$ that permute only the coordinates in ω_j and set $\mathcal{P}_\Omega = \prod_{j\in[n]} \mathcal{P}_{\omega_j}$. With $\underline{\pi}_\Omega = (\Pi_1, \dots, \Pi_n) \in \mathcal{P}_\Omega$ the association $X = [x_1 \cdots x_n] \mapsto \underline{\pi}_\Omega(X) = [\Pi_1 x_1 \cdots \Pi_n x_n]$ induces a map

$$\underline{\pi}_\Omega : \mathbb{R}^\Omega \longrightarrow \mathbb{R}^\Omega$$

Now suppose that the available data matrix $\tilde{X}$ is of the form $\tilde{X} = \underline{\tilde{\pi}}_\Omega \circ \underline{p}_\Omega(X^*)$ for some $\underline{\pi}^*_\Omega \in \mathcal{P}_\Omega$. Then the problem of *unlabeled matrix completion* can be posed as finding an element X in the fiber $\big(\underline{\pi}_\Omega \circ \underline{p}_\Omega\big)^{-1}(\tilde{X})$ of some map

$$\mathcal{M}_r \xrightarrow{\underline{p}_\Omega} \mathbb{R}^\Omega \xrightarrow{\underline{\pi}_\Omega} \mathbb{R}^\Omega$$

Assuming $\tilde{X}$ is generic, $\underline{p}_\Omega$ can be determined by inspection of the missing-pattern of $\tilde{X}$, while $\underline{\tilde{\pi}}_\Omega$ is unknown, as in unlabeled-PCA. Also, for a fixed $\underline{\pi}_\Omega \in \mathcal{P}_{m,\Omega}$ there is no a priori guarantee that $\tilde{X}$ is in the image of the map $\underline{\pi}_\Omega \circ \underline{p}_\Omega$, while as $\underline{\pi}_\Omega$ varies in $\mathcal{P}_{m,\Omega}$ more than one $\underline{\pi}_\Omega \circ \underline{p}_\Omega$'s may reach $\tilde{X}$ via possibly infinitely many X's in $\mathcal{M}_r$.

For this model, we will obtain theoretical recovery guarantees in Sections 2.2.1 and 2.2.2. Under the dominant permutation hypothesis, we will have theoretical assertions in Section 2.2.3, which further leads us to an algorithm in Section 3.2 and experimental analysis in Section 4.2.

2.2.1 Finite Recovery for UMC

In what follows we describe conditions under which finitely many $X \in \mathcal{M}_r$ explain the data $\tilde{X}$ for UMC. We exploit recent results of Tsakiris (2023b, 2024), where a family of generically finitely completable Ω's was studied. The following definition (Sturmfels and Zelevinsky, 1993) is needed for the description of the family.

Definition 5 *An (r, m)-SLMF (Support of a Linkage Matching Field) is a set*

$$\Phi = \bigcup_{j\in[m-r]} \varphi_j \times \{j\} \subset [m] \times [m-r]$$

with the φ_j's subsets of $[m]$ of cardinality $r+1$, satisfying

$$\#\bigcup_{j\in\mathcal{T}} \varphi_j \geq \#\mathcal{T} + r, \forall \mathcal{T} \subseteq [m-r].$$

We have the following finiteness result for unlabeled matrix completion:

Theorem 6 *Suppose $\Omega \subset [m] \times [n]$ satisfies the following two conditions. First, $\#\omega_j \geq r$ for every $j \in [n]$. Second, there exists a partition $[n] = \bigcup_{\nu \in [r]} \mathcal{J}_\nu$ of $[n]$ into r subsets $\mathcal{J}_\nu$, such that for every $\nu \in [r]$ there exist $m - r$ subsets $\varphi_j^\nu \in \bigcup_{k \in \mathcal{J}_\nu} \Omega_k$ with $j \in [m-r]$ such that $\Phi_\nu = \bigcup_{j \in [m-r]} \varphi_j^\nu \times \{j\}$ is an (r, m)-SLMF. For $\tilde{X} = \underline{\pi}_\Omega^* \circ \underline{p}_\Omega(X^*)$, where X^* is a generic matrix in $\mathcal{M}_r$, and $\underline{\pi}_\Omega^* \in \mathcal{P}_{m,\Omega}$, the following set of unlabeled completions is finite:*

$$\bigcup_{\underline{\pi}_\Omega \in \mathcal{P}_{m,\Omega}} \left(\underline{\pi}_\Omega \circ \underline{p}_\Omega\right)^{-1}(\tilde{X}) \tag{5}$$

The set (5) can be thought of as the set of all at most rank-r unlabeled completions of $\tilde{X}$. They can be computed, at least on a conceptual level, in a similar fashion as in UPCA by symmetric polynomials, this time supported on Ω.

2.2.2 UMC is an Algebraic Problem

We extend Theorem 3 to reveal an algebraic structure of the UMC problem:

Theorem 7 *Suppose Ω satisfies the hypothesis of Theorem 6. Then there is a Zariski-open dense set $\mathcal{U}$ in $\mathcal{M}_r$, such that for every $X^* \in \mathcal{M}_r$ the unlabeled completions (5) of $\tilde{X}$ are of the form $B'C'$, with identity in the top $r \times r$ block of B', and (B', C') ranging among the finitely many roots of the polynomial system*

$$\sum_{i \in \omega_j} (b_i^\top c_j)^\ell - \sum_{i \in \omega_j} \tilde{x}_{ij}^\ell = 0, \quad j \in [n],\, \ell \in [\#\omega_j]$$

In particular, for every root (B', C') there is $\underline{\pi}_\Omega \in \mathcal{P}_{m,\Omega}$ with $\underline{\pi}_\Omega \circ \underline{p}_\Omega(B'C') = \underline{p}_\Omega(X^*)$.

Remark 8 *By inspecting the proof of Theorem 6 one sees that the effect of the permutations manifests itself only through the fact that $\mathcal{P}_{m,\Omega}$ is a finite group of automorphisms of $\mathbb{R}^\Omega$. Hence, the proof and the statement of Theorem 6 remain unchanged if one replaces $\mathcal{P}_{m,\Omega}$ by any finite group of automorphisms of $\mathbb{R}^\Omega$. What will change in Theorem 7, is that one now needs to use polynomials that are invariant to the action of the specific group. Indeed, for permutations these are the symmetric polynomials.*

2.2.3 UMC with Dominant Permutations

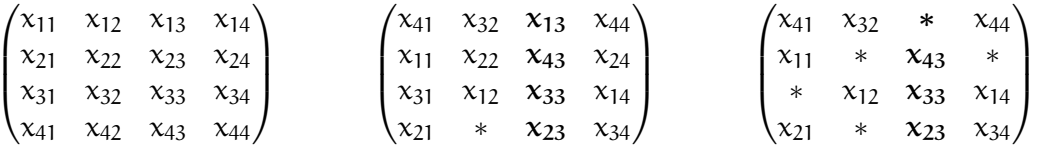
In Section 2.1.5 we discussed UPCA under the dominant permutation assumption. In that scenario, inliers are the columns of $\tilde{X}$ that lie in the (shuffled) ground-truth subspace ΠS^*, while outliers arise as the columns that are shuffled by permutations other than Π and thus driven away from ΠS^* (Figure 3b); namely, inliers and outliers in UPCA are partitioned as per

$$\tilde{X}_{\text{in}} := \{\tilde{x}_j \,|\, \Pi_j^* = \Pi\}, \quad \tilde{X}_{\text{out}} := \{\tilde{x}_j \,|\, \Pi_j^* \neq \Pi\}. \tag{6}$$

We now extend that scenario to the UMC setting.

Generalizing (6), we can define the partition for the UMC data $\tilde{X}$:

$$\tilde{X}_{\text{in}} := \{\tilde{x}_j \,|\, P_{\omega_j}\Pi_j^* = P_{\omega_j}\Pi\}, \quad \tilde{X}_{\text{out}} := \{\tilde{x}_j \,|\, P_{\omega_j}\Pi_j^* \neq P_{\omega_j}\Pi\} \tag{7}$$



(a) Ground-truth X^* (b) UPCA data matrix with a dominant permutation (c) UMC data matrix $\tilde{X}$ with a dominant permutation

Figure 3: Similar to Figure 1, yet the difference is as follows. In Figure 3b and 3c, columns $1, 2, 4$ of $\tilde{X}$ have been shuffled by the same permutation (i.e., the dominant permutation); column 3 (highlighted in bold) is shuffled by a different permutation and thus treated as an outlier.

Given the inherent ambiguity of UPCA, we can assume the dominant permutation is the identity I_m as in Section 2.1.5 (this is equivalent to replacing the ground-truth subpsace S^* by $\Pi^* S^*$, a harmless assumption for theoretical purposes, and often for practical ones as well). Then, the dominant identity permutation assumption entails sufficiently many j's for which $\Pi_j^* = I_m$, and also that $\tilde{X}_{in}$ contains sufficiently many data points $\tilde{x}_j$'s that would span S^* if correctly completed; we can thus naturally regard every point of $\tilde{X}_{in}$ as an inlier. However, pathological scenarios would arise if completing any point of $\tilde{X}_{out}$ in whatever way yielded a point in S^*. Fortunately, this pathological situation can in general be ruled out:

Proposition 9 *For a generic $X^* \in \mathcal{M}_r$, and for $\tilde{x} = P_\omega \Pi^* x^*$ a column in $\tilde{X}$ with $\#\omega \geq r+1$ and $P_\omega \Pi^* \neq P_\omega I_m$, any completion of $\tilde{x}$ is away from S^*, i.e. $P_\omega y \neq \tilde{x}$, $\forall y \in S^*$.*

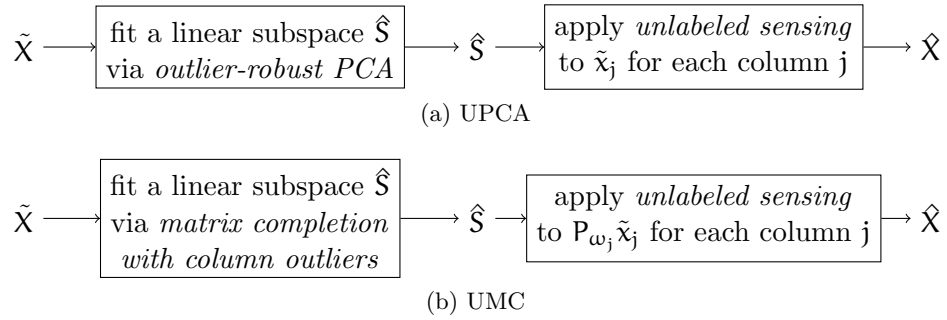
For a generic matrix $X^* \in \mathcal{M}_r$ with $\#\omega_j \geq r+1$ $(\forall j)$, it is now safe to treat every column of $\tilde{X}_{out}$ as an outlier, and indeed (7) gives a well-defined partition of inliers $\tilde{X}_{in}$ and outliers $\tilde{X}_{out}$. This extends the insight of UPCA with the dominant identity permutation assumption, and makes it possible to estimate S^* via matrix completion methods that are robust to outliers. Precise algorithmic solutions leveraging such insights fall right into Section 3.2.

3. Algorithms

In this section, we study the problems of UPCA and UMC under the dominant identity permutation assumption. We propose two-stage algorithmic pipelines for both problems:

- For UPCA, the first stage computes a subspace $\hat{S}$ from $\tilde{X}$ via outlier-robust PCA methods. The second stage applies unlabeled sensing methods to $\tilde{X}$ (and $\hat{S}$) in a column-wise manner. Figure 4a gives a diagram, and Section 3.1 gives full details.

- The UMC pipeline parallels and extends that of UPCA. The first stage computes $\hat{S}$ from $\tilde{X}$ via matrix completion with column outliers. The second stage first takes $P_{\omega_j}\hat{S}$ and $P_{\omega_j}\tilde{x}_j$ as inputs, and outputs an estimate $P_{\omega_j}\hat{x}_j$ via solving the problem of unlabeled sensing with missing entries. After completing the missing entries in $P_{\omega_j}\hat{x}_j$ by a least-square method, we get the estimate $\hat{x}_j$. See Figure 4b and Section 3.2.



(a) UPCA

(b) UMC

Figure 4: The proposed algorithmic pipelines for UPCA and UMC.

3.1 Two-Stage Algorithmic Pipeline for UPCA

We saw in the previous section that the UPCA problem (2) is well-defined (Theorem 2) and in principle solvable by a polynomial system of equations (Theorem 3). However, this polynomial system is at the moment intractable to solve even for moderate dimensions. On the other hand, Theorem 4 suggests the following practical two-stage algorithmic pipeline for the case where there is a dominant permutation, which we will take to be the identity.

Here we proposed a two-stage method, summarized in Algorithm 1.

3.1.1 Stage-I of UPCA

The existence of a dominant identity permutation enables estimating the underlying subspace, which is the task of Stage-I in the proposed algorithmic pipeline. Hence, at Stage-I a PCA method with robustness to outliers is employed to produce an estimate $\hat{S}$ of S^* from $\tilde{X}$. Such robust PCA methods include OP (Xu et al., 2012), Self-Repr (Soltanolkotabi and Candés, 2012; You et al., 2017), CoP (Rahmani and Atia, 2017), and DPCP (Tsakiris and Vidal, 2018b; Zhu et al., 2018; Lerman and Maunu, 2018), as mentioned in Section 1.1.2.

3.1.2 Stage-II of UPCA

Once equipped with a robust estimate of S^*, the aim of Stage-II is to estimate X^*, which can be achieved by employing methods for unlabeled sensing. These methods take a point $\tilde{x}_j$ of $\tilde{X}$, identified as an outlier with respect to the subspace $\hat{S}$, and return an estimate $\hat{x}_j$ of x_j^* by (directly or indirectly) solving the problem

$$\min_{\Pi \in \mathcal{P}_m,\, \hat{x}_j \in \hat{S}} \|\tilde{x}_j - \Pi \hat{x}_j\|_2 \tag{8}$$

Hence, at Stage-II of the pipeline, one feeds $\hat{S}$ and $\tilde{X}$ to an unlabeled sensing method (Slawski and Ben-David, 2019; Slawski et al., 2021; Tsakiris et al., 2020; Peng and Tsakiris, 2020; Mazumder and Wang, 2023; Onaran and Villar, 2022), which operates point by point, returning for every $\tilde{x}_j$ an estimate $\hat{x}_j$. Here one may choose to threshold the $\tilde{x}_j$'s based on their distance to $\hat{S}$ and apply unlabeled sensing on the outliers only. Alternatively, if extra computational power is available for dispensing with choosing a threshold, one may apply unlabeled sensing on every $\tilde{x}_j$; we follow this approach in the experiments for UPCA.

Algorithm 1 Two-stage Algorithmic Pipeline for UPCA

Input: observed data matrix $\tilde{X}$, rank r
estimate $\hat{S}$ of $S^* \leftarrow$ outlier-robust PCA on $\tilde{X}$ ▷ Stage-I
for $j = 1, \dots, n$ **do** ▷ Stage-II
    estimate $\hat{x}_j$ of $x_j^* \leftarrow$ unlabeled sensing (8) on $(\tilde{x}_j, \hat{S})$
end for
return estimate $\hat{X} = [\hat{x}_1, \dots, \hat{x}_n]$ of X^*

3.1.3 A New Method For Unlabeled Sensing: LSRF

Inasmuch as there are very few scalable unlabeled sensing methods, we here propose a simple but comparatively efficient alternative in sparsely permuted cases, named *Least-Squares with Recursive Filtration* (LSRF), see Algorithm 2. This method is parameter-free and alternates between ordinary least-squares and a dimensionality reduction step that removes the coordinate of the ambient space on which the residual error attains its maximal value, until r coordinates are left.

LSRF is designed to solve unlabeled sensing with sparse permutations, empirically succeeding for up to 60%-70% permuted entries. The complexity is $\mathcal{O}(m^2r^2)$, and can be further reduced to $\mathcal{O}((\frac{mr}{\kappa})^2)$ by removing κ rows in A in each iteration when m is large. For comparing computational complexity with other algorithms, which often have hyper-parameters, AIEM is exponential in r and linear in m; each iteration of PL has complexity at least $\mathcal{O}(mr^3)$ if done via second-order optimization; ℓ_1-RR has complexity at least $\mathcal{O}(mr^2 + k(m + r^2))$ if done via sub-gradient descent, where k is the number of iterations.

Algorithm 2 Unlabeled Sensing via Least-Squares with Recursive Filtration (LSRF)

Input: permuted point $\tilde{x}_j$, basis B^* of subspace S^*
$v^{(0)} \leftarrow \tilde{x}_j, A^{(0)} \leftarrow B^*$
for $k = 1, \dots, m - r$ **do**
    $c \leftarrow A^{(k-1)\dagger} v^{(k-1)}$
    $i' \leftarrow \operatorname{argmax}_i |v_i^{(k-1)} - A_i^{(k-1)} c|$
    remove the i'th entry of $v^{(k-1)}$ to get $v^{(k)}$
    remove the i'th row of $A^{(k-1)}$ to get $A^{(k)}$
end for
return estimate $\hat{x}_j = A^{(m-r)} A^{(m-r)\dagger} v^{(m-r)}$ for x_j^*

3.2 Two-Stage Algorithmic Pipeline for UMC

As in UPCA, we propose a two-stage pipeline for UMC that can be effective under the dominant permutation assumption. We detail our algorithmic pipeline next.

3.2.1 Stage-I of UMC

Stage-I of UMC parallels that of UPCA, with the same goal of estimating the ground-truth subspace S^*, yet with the additional challenge that the data matrix now has missing entries (Figure 1). Recall that, under the dominant permutation assumption, we can treat the columns permuted by the dominant permutation as inliers and others outliers (Figure 3). As such, Stage-I amounts to solving the problem of matrix completion with column outliers (reviewed in Section 1.1.3). To do so, a direct solution is employing the convex program of Chen et al. (2015), called MCO, which estimates S^* in a way that is robust to column outliers and missing entries.

However, MCO comes with two issues that might hinder its accuracy. First, it aims to complete inliers and detect outliers *simultaneously*; doing so can be very challenging and thus error-prone. Second, its *convex* program can not leverage the inherent non-convexity of the problem. To alleviate these issues, we build upon existing *non-convex* outlier-robust PCA procedures and propose an alternative to MCO. The alternative proposal detects inliers, completes inliers, and estimates the ground-truth subspace S^* *in cascade*:

1. (*Detect Inliers*) We first complete all missing entries of $\tilde{X}$ by the value 0, thereby obtaining a complete matrix $\tilde{X}_0$; such matrix $\tilde{X}_0$ is sometimes called *zero-filled* data matrix (Yang et al., 2015; Tsakiris and Vidal, 2018a). Similarly to (7), $\tilde{X}_0$ can be partitioned into *zero-filled inliers* and *zero-filled outliers*. Then we proceed as if the zero-filled inliers were correct completions of UMC inliers $\tilde{X}_{\text{in}}$ that span the ground-truth space S^*, and run existing outlier-robust PCA methods on $\tilde{X}_0$. Since the missing entries of $\tilde{X}$ are heuristically filled by zeros, such a subspace estimate might be inaccurate, away from S^*, and thus inadequate for the subsequent recovery task. On the other hand, if we are further given an *inlier threshold* as a hyper-parameter, then we can obtain an estimated partition of inliers and outliers. In particular, an incomplete point $\tilde{x}_j$ is classified as an inlier if the distance between its zero-filled version and the estimated subspace is smaller than the inlier threshold; otherwise it is an outlier. Hence, we declare a set of inliers as determined by the partition, and will use these estimated inliers for the sequel.
2. (*Complete Inliers*) The detected inliers $\tilde{x}_j$'s in the previous step are in fact incomplete, and at this point, we will no longer rely on their zero-filled versions; instead, we aim to find their authentic completions (x_j^*)'s, which are expected to span S^* if the detected inliers are indeed inliers in the sense of (7). In other words, we are now confronted with a low-rank matrix completion task. Therefore, step 2 is to complete the detected inliers using standard matrix completion algorithms.
3. (*Estimate* S^*) Finally, given an estimate for the ground-truth rank r, we can now simply perform a singular value decomposition on the matrix of completed inliers, and obtain the final estimate $\hat{S}$ of S^*.

The above routine in cascade can be upgraded into a block coordinate descent method (Peng and Vidal, 2023): Alternate among inlier detection and completion and subspace estimation. Moreover, convergence guarantees of (Peng and Vidal, 2023) might be applied here. That said, we do not pursue this idea of block coordinate descent here, as the above routine already shows satisfactory recovery performance (see Section 4.2).

3.2.2 Stage-II of UMC

After Stage-I, we have obtained an estimate $\widehat{S}$ of the ground-truth subspace S^* and completed inliers. Since each outlier $\tilde{x}_j$ is a point x_j^* in S^* except being permuted and having some entries missing, there is a chance of recovering a good estimate $\hat{x}_j$ of x_j from $\widehat{S}$ and $\tilde{x}_j$. In such cases, we can first restore the permutation of each column via solving (8) on observed entries (using the subspace $P_{\omega_j}\widehat{S}$) and then complete the missing entries of that column via a least-squares computation. We summarize our approach in Algorithm 3.

Algorithm 3 Two-stage Algorithmic Pipeline for UMC

Input: observed data matrix $\tilde{X}$, rank r
$\tilde{X}_0 \leftarrow$ fill missing entries in $\tilde{X}$ with zeros
estimate $\widehat{S}$ of $S^* \leftarrow$ matrix completion with column outliers on $\tilde{X}_0$ ▷ Stage-I
$\widehat{B} \leftarrow$ basis of $\widehat{S}$
for $j = 1, \ldots, n$ **do** ▷ Stage-II
    estimate $P_{\omega_j}\widehat{\Pi}_j$ of $P_{\omega_j}\Pi_j^* \leftarrow$ unlabeled sensing (8) on $(P_{\omega_j}\tilde{x}_j, P_{\omega_j}\widehat{S})$
    estimate coefficient $\hat{c}_j$ of x_j^* in basis $\widehat{B} \leftarrow$ least squares on $(P_{\omega_j}\widehat{B}, P_{\omega_j}\widehat{\Pi}_j\tilde{x}_j)$
    $\hat{x}_j \leftarrow \widehat{B} \cdot \hat{c}_j$
end for
return estimate $\widehat{X} = [\hat{x}_1, \ldots, \hat{x}_n]$ of X^*

4. Experimental Evaluation

Here we perform synthetic and real data experiments to evaluate the proposed algorithmic pipelines for UPCA (Section 4.1) and UMC (Section 4.2). We use two metrics for performance evaluation. The first is the largest principal angle $\theta_{\max}(S^*, \widehat{S})$ between the estimated subspace $\widehat{S}$ and ground-truth S^*, and this is used for Stage-I to evaluate subspace learning accuracy. The second metric is the relative estimation error $\frac{\|\widehat{X}-X^*\|_F}{\|X^*\|_F}$ between the estimated data matrix $\widehat{X}$ and the ground-truth X^*, which quantifies the final performance of our algorithmic pipeline. For both metrics, smaller values imply better performance.

4.1 UPCA Experiments

We begin by assessing the performance of Stage-I of the pipeline in Section 4.1.1. This entails understanding how different PCA methods with robustness to outliers behave when the outliers are induced by permutations, as in Theorem 4. Next in section 4.1.2, we evaluate the overall UPCA pipeline of Algorithm 1 on synthetic data with added spherical noise.

4.1.1 Stage-I of UPCA

To understand how different PCA methods with robustness to outliers behave when the outliers are induced by permutations, we access the performance of Stage-I of the pipeline in section 4.1.1. We consider Self-Expr (Self-Expressive elastic net) (You et al., 2017; Soltanolkotabi and Candés, 2012), CoP (Coherence Pursuit) (Rahmani and Atia, 2017), OP

(Outlier Pursuit) (Xu et al., 2012), and DPCP (Dual Principal Component Pursuit) (Tsakiris and Vidal, 2018b; Lerman and Maunu, 2018); these methods are reviewed in Section 1.1.2.

We fix $m = 50$ and $n = 500$. With $\dim S^*$ taking values $r = 1 : 1 : 49$, S^* is sampled uniformly at random from the Grassmannian $\mathrm{Gr}(r, m)$. Then n points x_j^* are sampled uniformly at random from the intersection of S^* with the unit sphere of $\mathbb{R}^m$ to yield X^*. Denote by n_{in} the number of inliers and n_{out} the number of outliers, with $n_{\text{in}} + n_{\text{out}} = n$. We consider outlier ratios $n_{\text{out}}/n = 0.1 : 0.1 : 0.9$. For a fixed outlier ratio, we set $\tilde{\Pi}_j$ to the identity for $j \in [n_{\text{in}}]$ and determine the $\tilde{\Pi}_j$'s for $j > n_{\text{in}}$ as follows. An important parameter in the design of a permutation Π is its sparsity level $\alpha \in [0, 1]$. This is the ratio of coordinates that are moved by Π. To obtain $\tilde{\Pi}_j$, for a fixed α, for each $j > n_{\text{in}}$, we randomly choose αm coordinates and subsequently a random permutation on those coordinates. We consider permutation sparsity levels $\alpha = 1, 0.6, 0.2, 0.1$.

In Self-Repr and CoP, $\hat{S}$ is taken to be the subspace spanned by the top r $\tilde{x}_j$'s with largest inlier scores. We use the Iteratively-Reweighed-Least-Squares method proposed by Tsakiris and Vidal (2017) and Lerman and Maunu (2018) for solving the DPCP problem as it works very well empirically.[1] The output subspace $\hat{S}$ of OP is obtained as the rth principal component subspace of the decomposed low-rank matrix. For Self-Expr we use $\lambda = 0.95$, $\alpha = 10$ and $T = 1000$, see section 5 in You et al. (2017). For DPCP we use $T_{\max} = 1000$, $\epsilon = 10^{-9}$ and $\delta = 10^{-15}$, see Algorithm 2 in Tsakiris and Vidal (2018b). Finally, OP uses $\lambda = 0.5$ and $\tau = 1$ in Algorithm 1 of Xu et al. (2012).

Figure 5 depicts the *outlier-ratio versus rank* phase transitions, where to calibrate the analysis with what we know about these methods from prior work, we have included in the top row of the figure the phase transitions for outliers randomly chosen from the unit sphere. By reading that top row we recall: i) DPCP has overall the best performance across all ranks and all outlier ratios, ii) OP identifies correctly S^* only in the low rank low outlier-ratio regime, as expected from its conceptual formulation, and iii) CoP and Self-Expr, even though low-rank methods in spirit, they have accuracy similar to each other and considerably better than OP. We also note that CoP is the fastest method requiring $0.51\mathrm{sec}$ for the computation of a single phase transition plot for each trial (i.e., average time for running all settings of outlier ratio $0.1 : 01 : 0.9$ and rank $1 : 1 : 49$ once), Self-Expr is the slowest with $752\mathrm{sec}$ and DPCP and OP take $1.31\mathrm{sec}$ and $5.62\mathrm{sec}$, respectively[2].

Now let us look at what happens for permutation-induced outliers. For $\alpha = 1$, where the permutations move all the coordinates of the points they are corrupting, we see that the phase transition plots are practically the same as for random outliers. In other words, obtaining the outliers by randomly permuting all coordinates of inlier points, with different permutations for different outliers, seems to be yielding an outlier set as generic for the task of subspace learning as sampling the outliers randomly from the unit sphere. A second interesting phenomenon is observed when the permutation ratio is decreased to $\alpha = 0.1$. In that regime the methods exhibit two very different trends. On one hand, CoP and Self-Expr appear to break down, which is expected, because as the permutations become more sparse, the outlier points become more coherent with the rest of the data set. On the other hand, the accuracy of DPCP and OP improves for sparser permutations; a justification for this

1. Theoretical advances on Iteratively-Reweighed-Least-Squares have also been recently made; see, e.g., Peng et al. (2022, 2023) and Section 2.4 of Peng and Vidal (2023).
2. Experiments are run on an Intel(R) i7-8700K, 3.7 GHz, 16GB machine.

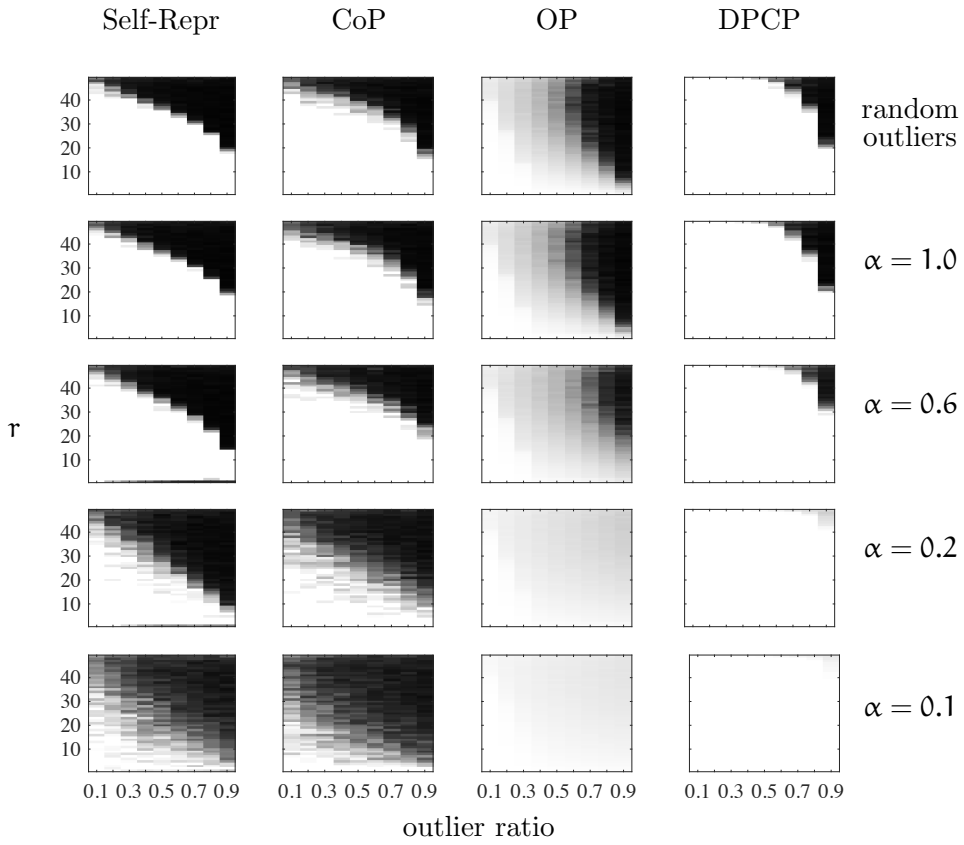


Figure 5: $\theta_{\max}(S^*, \hat{S})$ in UPCA Stage-I: outlier ratio vs. rank phase transitions for various PCA methods with robustness to outliers.

is that both methods get initialized via the SVD of $\tilde{X}$, which yields a subspace closer to S^* for smaller α. For example, the value of principal angles $\theta_{\max}(S^*, \hat{S})$ for Self-Expr, CoP, OP, DPCP, for $\alpha = 0.2$, outlier ratio 0.9 and $r = 49$ are $79°, 83°, 14°, 13°$, respectively. As another example, for $\alpha = 0.1$, outlier ratio 0.7 and $r = 25$ the value of $\theta_{\max}(S^*, \hat{S})$ is $67°, 80°, 7°, (10^{-6})°$, with the methods ordered as above. Overall, DPCP is consistently outperforming the rest of the methods, justifying it as our primary choice in the next section. An interesting research direction is to analyze the theoretical guarantees of these methods for this specific type of outliers.

4.1.2 The Full Pipeline of UPCA

Now we evaluate the UPCA pipeline of Algorithm 1 on synthetic data. We keep $m = 50$ as before, and add spherical noise with to a fixed SNR of 40dB. We get the estimate $\hat{S}$ of S^* via DPCP (Tsakiris and Vidal, 2018b; Lerman and Maunu, 2018) in Stage-I and apply the unlabeled sensing methods (Tsakiris et al., 2020; Peng and Tsakiris, 2020; Slawski and Ben-David, 2019; Slawski et al., 2021) and Algorithm 2 in Stage-II to get $\hat{X}$ from $\hat{S}$ and $\tilde{X}$. We distinguish between dense and sparse permutations.

Dense Permutations. We first consider dense permutations, that is $\alpha = 1$. This is an extremely challenging case, with the difficulty manifesting itself through the fact that existing methods can only handle small ranks r. We consider AIEM and CCV-Min, two state-of-the-art methods mentioned in Section 1.1.1. For AIEM we use a maximum number of 1000 iterations in the alternating minimization of (8). For CCV-Min we use a precision of 0.001, the maximal number of iterations is set to 50, and the maximum depth to 12 for $r = 3$ and 14 for $r = 4, 5$.

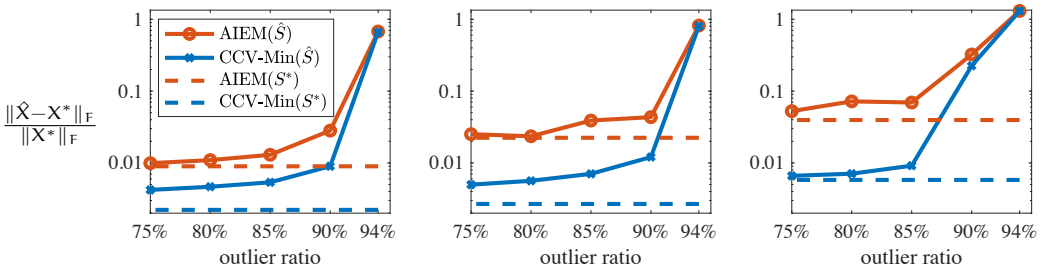
Figure 6 depicts the relative estimation error of $\hat{X}$ for different outlier ratios from 75% (25 inliers) to 94% (6 inliers) and ranks $r = 3, 4, 5$. To assess the overall effect of the quality of $\hat{S}$, we use two versions of AIEM and CCV-Min. The first, denoted by AIEM($\hat{S}$) and CCV-Min($\hat{S}$), uses as input the estimated subspace $\hat{S}$, while the second version, AIEM(S^*) and CCV-Min(S^*), uses the ground-truth subspace S^*. Note that the estimation error of AIEM(S^*)/CCV-Min(S^*) is independent of the outlier ratio. On the other hand, the estimation error of AIEM($\hat{S}$)/CCV-Min($\hat{S}$) depends on the outlier ratio through the computation of $\hat{S}$. Indeed, $\hat{S}$ is expected to be closer to S^* for smaller outlier ratios, as we already know from Figure 5. In particular, for up to 75% outliers the estimation error of AIEM($\hat{S}$)/ CCV-Min($\hat{S}$) coincides with that of AIEM(S^*)/ CCV-Min(S^*), indicating an accurate estimation of S^*. At the other extreme, for 94% outliers both AIEM($\hat{S}$)/ CCV-Min($\hat{S}$) break down, indicating that the estimation of $\hat{S}$ failed. Finally, note that CCV-Min has at least half order of magnitude smaller estimation error than AIEM. This is due to our specific choice of the branch & bound CCV-Min parameters which control the trade-off between accuracy and running time; for example, for $r = 3$ and 75% outliers, AIEM runs in 42msec with 1% error, while CCV-Min needs about 15sec to bound $\hat{X}$ 0.42% away from X^*.

For AIEM, we use the customized Gröbner basis solvers of Tsakiris et al. (2020), developed for $r \leq 4$, which solve the polynomial system in milliseconds, and the maximum number iterations in the alternating minimization procedure is $T_{\max} = 1000$. For $r = 5$, the design of such solvers is an open problem[3], thus we use the generic solver Bertini (Bates et al.), which runs within a few seconds. For $r \geq 6$ though, AIEM remains as of now practically intractable. For CCV-Min the precision is 0.001, $T_{\max} = 50$, and the maximum depth is 12 for $r = 3$ and 14 for $r = 4, 5$. For ℓ_1-RR we use $\lambda = 0.01\sqrt{\log(n)/n}$ in (13) of Slawski and Ben-David (2019).

Sparse Permutations. The methods that we saw in the previous section certainly apply in the special case where only a fraction of the coordinates is permuted. However, they are still subject to the same computational limitations that practically require the rank r to be small ($r \leq 6$ for AIEM and $r \leq 8$ for CCV-Min). On the other hand, the problem of linear regression without correspondences is tractable for a wider range of ranks when the permutations are sparse (small α). This important case arises in applications such as record linkage, where domain specific algorithms are only able to guarantee partially correctly matched data. Here we consider three methods, ℓ_1-RR (Slawski and Ben-David, 2019), PL (Slawski et al., 2021), and our proposed LSRF (Algorithm 2).

Figures 7b-7d show the relative estimation error of UPCA for $\alpha = 0.1 : 0.1 : 0.6$, rank $r = 1 : 1 : 25$ and outlier ratio fixed to 90%, with $\hat{S}$ computed in Stage-I by DPCP (Tsakiris and Vidal, 2018b; Lerman and Maunu, 2018) and $\hat{X}$ computed via ℓ_1-RR, PL, or LSRF

3. The fast solver generator of Larsson et al. (2017) is an improved version of the one used by Tsakiris et al. (2020) for $r = 3, 4$. However, we found that for $r = 5$ it suffers from numerical stability issues.



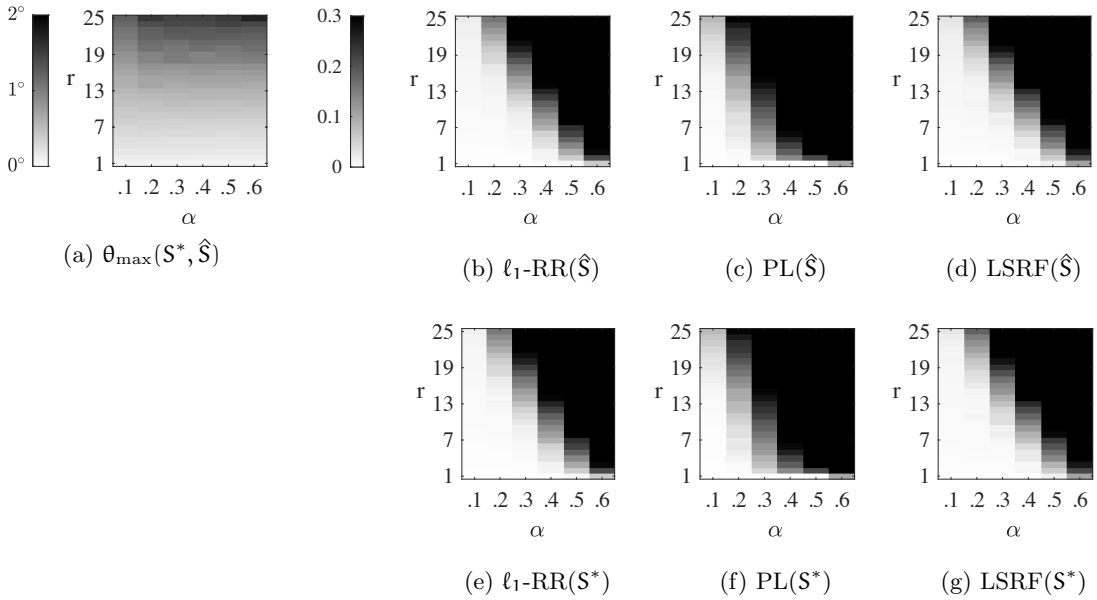
(a) $r = 3$ (b) $r = 4$ (c) $r = 5$

Figure 6: UPCA (Algorithm 1) for dense permutations ($\alpha = 1$) with $\hat{S}$ produced by DPCP (Tsakiris and Vidal, 2018b; Lerman and Maunu, 2018) at Stage-I and $\hat{X}$ produced by AIEM (Tsakiris et al., 2020) or CCV-Min (Peng and Tsakiris, 2020) at Stage-II.

from $\tilde{X}$ and $\hat{S}$ in Stage-II. It is important to note that $r = 25 = m/2$ is the largest rank for which unique recovery of X^* is theoretically possible (Unnikrishnan et al., 2015, 2018; Tsakiris and Peng, 2019; Dokmanić, 2019). Figure 7a shows that $\theta_{\max}(S^*, \hat{S})$ always stays below 2°, indicating the success of DPCP. As before, we also show in Figures 7e-7g the estimation error when S^* is used instead of $\hat{S}$. Evidently, the performance is nearly identical regardless of whether $\hat{S}$ or S^* is used, again justifying the success of Stage-I. Now ℓ_1-RR and Algorithm 2 have similar accuracy, but Algorithm 2 is more efficient than ℓ_1-RR, considering that computing $\hat{X}$ takes 0.3sec seconds for Algorithm 2 and 1.5min for ℓ_1-RR. Even though PL delivers $\hat{X}$ in 1sec, it is not performing as well, which we attribute to its sensitivity on the particular basis of S^* that is used to generate the data; this is not available here since DPCP returns the specific basis of dual principal components.

4.1.3 Experiments on Face Images

In this section we offer a flavor of how the ideas discussed so far apply in a high-dimensional example with real data. We use the well-known database Extended Yale B (Georghiades et al., 2001), which contains fixed-pose face images of distinct individuals, with 64 images per individual under different illumination conditions. It is well-established that the images of each individual approximately span a low-dimensional subspace. It turns out that for our purpose the value $r = \dim S^* = 4$ is good enough, and values higher than r do not bring improvements that human eyes can distinguish. Since each image has size 192×168, the images of each individual can be approximately seen as $n = 64$ points x_j^*, $j \in [64]$ of a 4-dimensional linear subspace S^*, embedded in an ambient space of dimension $m = 32256$. In what follows we only deal with the images of a fixed individual. We consider four permutation types corresponding to fully or partially ($\alpha = 0.4$) permuting image patches of size 16×24 or 48×42, as shown in the second column of Figure 8. To generate a fixed number of $n_{\text{out}} = 16$ outliers only one out of the four permutation types is used for each trial. The original images (inliers) together with the ones that have undergone patch-permutation (outliers) are given without any inlier/outlier labels, and the task is to restore all corrupted images. This is a special case of visual permutation learning, recently considered using deep networks (Santa Cruz et al., 2017, 2018).



(a) $\theta_{\max}(S^*, \hat{S})$ (b) ℓ_1-RR($\hat{S}$) (c) PL($\hat{S}$) (d) LSRF($\hat{S}$)

(e) ℓ_1-RR(S^*) (f) PL(S^*) (g) LSRF(S^*)

Figure 7: Estimation error $\frac{\|X^* - \hat{X}\|_F}{\|X^*\|_F}$ of UPCA (Algorithm 1) for sparse permutations ($\alpha \leq 0.6$) and outlier ratio 90%, with $\hat{S}$ computed by DPCP (Tsakiris and Vidal, 2018b; Lerman and Maunu, 2018) in Stage-I and $\hat{X}$ computed by ℓ_1-RR (Slawski and Ben-David, 2019), PL (Slawski et al., 2021) or Algorithm 2 in Stage-II.

We compute $\hat{S}$ as follows. With $\tilde{X} = U\Sigma V^\top$ the thin SVD of $\tilde{X}$, where $U \in \mathbb{R}^{32256 \times 64}$, DPCP fits a 4-dimensional subspace $\bar{S}$ to the columns of $\bar{X} = U^\top \tilde{X}$, a process which takes about a tenth of a second. Then $\bar{S}$ is embedded back into $\mathbb{R}^{32256}$ via the map $U : \mathbb{R}^{64} \to \mathbb{R}^{32256}$ to yield $\hat{S}$. To compute $\hat{X}$ from $\hat{S}$ and $\tilde{X}$ we use the custom algebraic solver of AIEM as well as ℓ_1-RR, PL, LSRF, with a proximal subgradient implementation of ℓ_1-RR using the toolbox of Beck and Guttmann-Beck (2019).

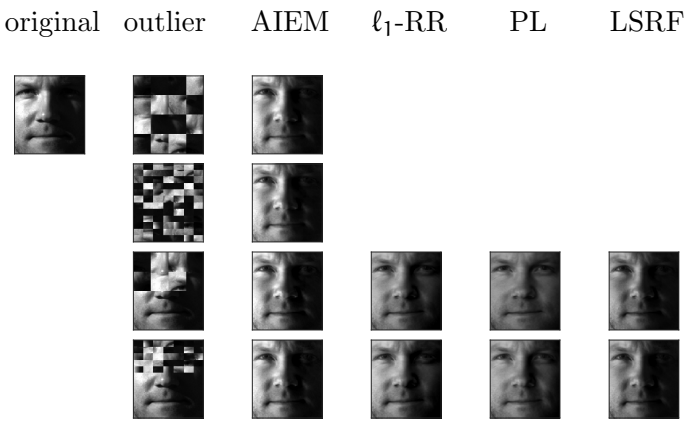
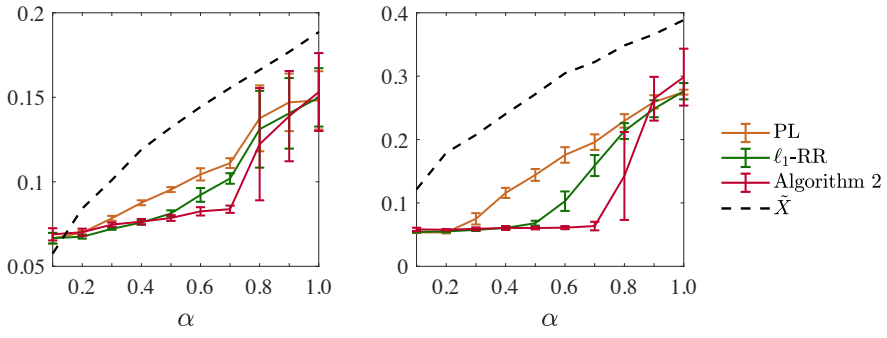


Figure 8: UPCA on the face data set Extended Yale B.

The first column of Figure 8 shows an original image, and the second column shows the corresponding outlier obtained by applying a sample permutation for each of the four different permutation types. Columns three to six give the corresponding point in the output of Algorithm 1 for different unlabeled sensing methods and $\widehat{\mathbf{S}}$ computed by DPCP (Tsakiris and Vidal, 2018b). CCV-Min (Peng and Tsakiris, 2020) is not included as branch-and-bound becomes prohibitively expensive for such large $\mathtt{m}$). Notably, AIEM (Tsakiris et al., 2020) rather satisfactorily restores the original image regardless of permutation type. The performance of the other three methods is shown only for their operational regime, where the given data are corrupted by sparse permutations, and Algorithm 2 most accurately captures the illumination of the original image. Overall, we find these results encouraging, especially if one takes into consideration that the methods are very efficient, requiring only $0.2\mathtt{sec}$ (AIEM), $7\mathtt{sec}$ (ℓ_1-RR), $0.2\mathtt{sec}$ (PL) and $10\mathtt{sec}$ (Algorithm 2), discounting the DPCP step, which costs $0.1\mathtt{sec}$, regardless of permutation type. This is in contrast with existing deep network architectures for visual permutation learning, such as (Santa Cruz et al., 2018), which are based on branch-and-bound and thus have in principle an exponential complexity in the number of permuted patches.

4.1.4 Experiments on Data Re-identification (UPCA)

Finally, we evaluate the UPCA Algorithm 1 for the task of re-identification (section 1) using real educational and medical records and simulated permutations for various sparsity levels α, thus emulating a privacy protection scenario. Both of the data sets that we use contain no personally identifiable information. DPCP (Tsakiris and Vidal, 2018b; Lerman and Maunu, 2018) computes $\widehat{\mathbf{S}}$ in Stage-I and ℓ_1-RR (Slawski and Ben-David, 2019), PL (Slawski et al., 2021) or Algorithm 2 produce $\widehat{\mathbf{X}}$ in Stage-II.



(a) high-school scores (b) breast tumor features

Figure 9: Relative estimation error $\frac{\|\widehat{\mathbf{X}}-\mathbf{X}^*\|_{\mathrm{F}}}{\|\mathbf{X}^*\|_{\mathrm{F}}}$ for UPCA on real data in de-anonymization.

The first data set consists of the test scores of $\mathtt{m} = 707$ high-school students on 6 subjects during two different periods, together with the sum of the score tests for each period, thus $\mathtt{n} = 14$. For 7 out of 14 tests we apply random permutations of the student indices and thus have 50% outliers. With $\mathtt{r} = 3$, the relative estimation errors on the score records are shown in Figure 9a. The black dashed line depicts the relative difference between the observed data $\tilde{\mathbf{X}}$ and the original data $\mathbf{X}^*$, which as expected increases for higher α's. The performance of

ℓ_1-RR, PL and Algorithm 2 is in alignment with our earlier findings in that Algorithm 2 tends to have a superior performance and PL is the least competitive. All these methods apply in principle for sparse permutations and thus their accuracy naturally degrades for large α.

The second data set consists of all the benign cases in Breast Cancer Wisconsin (Diagnostic) (Asuncion and Newman, 2007). It has $m = 357$ patients and $n = 30$ features of a breast mass digitized image for each patient. We randomly permute the patient indices for 15 of the features thus having 50% outliers and set $r = 4$. Figure 9b shows the relative estimation error of $\hat{X}$ for various permutation sparsity levels α, with the unlabeled sensing methods exhibiting the same trend as before. Remarkably, for $\alpha = 0.7$, the UPCA Algorithm 1 incorporating Algorithm 2 in Stage-II reduces the original error of the data $\tilde{X}$ from 32.24% to 6.35% in 0.5sec, as opposed to 15.90% and 19.57% when ℓ_1-RR (Slawski and Ben-David, 2019) or PL (Slawski et al., 2021) are incorporated, respectively.

4.2 UMC Experiments

In this section, we evaluate the proposed two-stage algorithmic pipeline, Algorithm 3, for UMC. Section 4.2.1 tests Stage-I and reports the recovery accuracy of the subspace S^*. Section 4.2.2 presents the performance of the full pipeline in terms of recovering X^*.

The experiments of Sections 4.2.1 and 4.2.2 operate on synthetic data with the following setup. We set the ambient dimension $m = 50$, the overall number of data points $n = 100$, the dimension of the ground-truth subspace $r = 3$, the noise level is 0.01, inliers are associated with the dominant identity permutation, and outliers are shuffled by random dense permutations ($\alpha = 1$).

Finally, in Section 4.2.3 we study again the experiments of data re-identification once shown in Section 4.1.4, this time for the case of UMC.

4.2.1 Stage-I of UMC

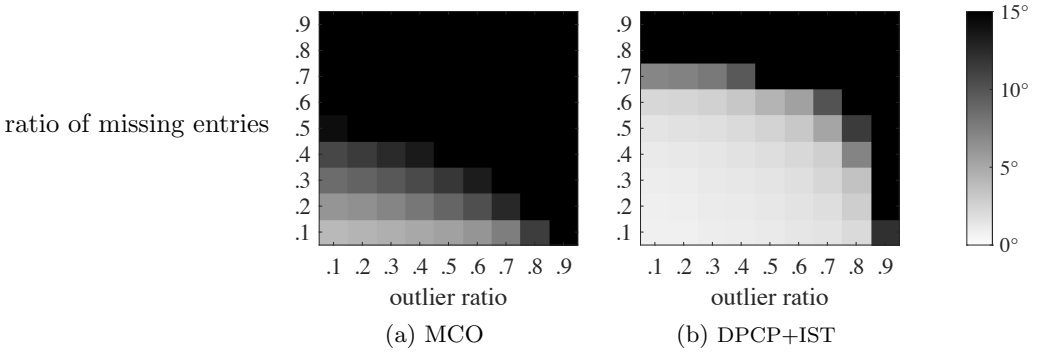
As discussed in Section 3.2, Stage-I amounts to solving the problem of matrix completion with column outliers, and for this, we can use two approaches. One is MCO (Chen et al., 2015), which simultaneously detects the inliers and estimates the ground-truth subspace S^*. The other approach, called DPCP+IST, is an instantiation of our idea in Section 3.2: (1) detect the inliers applying the DPCP method (Tsakiris and Vidal, 2018b) on zero-filled data; (2) complete the detected inliers using a non-convex method based on *iterative soft thresholding* (Majumdar and Ward, 2011), which we call IST; (3) estimate S^* using an SVD on the matrix of completed inliers.

With the output $\hat{S}$ of either of the above two approaches, we report the largest principal angle $\theta_{\max}(S^*, \hat{S})$ in Figure 10 for different ratios of outliers ($0.1 : 0.1 : 0.9$) and missing entries ($0.1 : 0.1 : 0.9$). In particular, via Figure 10 we deliver two messages:

- Both MCO and DPCP+IST find an accurate enough subspace estimate $\hat{S}$ given sufficiently many inliers and observed entries. For example, we have $\theta_{\max}(S^*, \hat{S})$ equal to 2.44° for MCO and 5.51° for DPCP+IST for 10% missing entries and 50% outliers.

- The accuracy of MCO decays more rapidly than DPCP+IST in the presence of more outliers and more missing entries. For example, with 50% outliers and 50% missing

entries, we have $\theta_{\max}(S^*, \hat{S})$ equal to $20.37°$ for MCO and $2.61°$ for DPCP+IST. Moreover, the figure shows DPCP+IST can handle up to 60% missing entries & 60% outliers, or 10% missing entries & 40% outliers, with errors of roughly $5°$ when $m = 50, n = 100$, and $r = 3$.



(a) MCO (b) DPCP+IST

Figure 10: The largest principal angle $\theta_{\max}(S^*, \hat{S})$ for Stage-I of Algorithm 3 on synthetic data with varying ratios of outliers and missing entries.

Overall, at least in the present setting, DPCP+IST appears to be more accurate and more robust than MCO for Stage-I. This is perhaps because DPCP+IST benefits from decoupling the estimation task into several steps, where each step sufficiently leverages the non-convex structure of the problem.

Figure 11 reports the largest principal angle $\theta_{\max}(S^*, \hat{S})$ for $r = 1 : 1 : 20$ when the outlier ratio is 40% and the ratio of missing entries is 20% with $m = 50, n = 100$. In this setting, the performance starts dropping significantly when the rank is greater than 12.

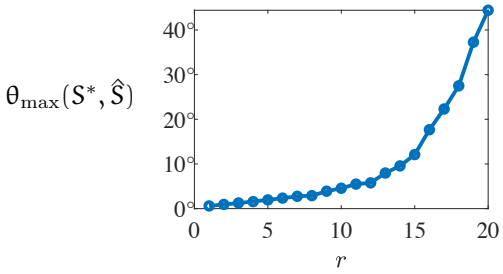


Figure 11: The largest principal angle $\theta_{\max}(S^*, \hat{S})$ for Stage-I of DPCP+IST on synthetic data with varying ranks.

4.2.2 The Full Pipeline of UMC

We now evaluate the whole two-stage pipeline (Algorithm 3) for UMC. We solve Stage-I via DPCP+IST; see Figure 10 and Section 4.2.1. We solve Stage-II via either AIEM or CCV-Min; see Section 1.1.1.

In the experiments, we fix the outlier ratio to 40% with varying ratios of missing entries $0.1 : 0.1 : 0.8$. For better illustration, we report the recovery accuracy for inliers and outliers separately. In particular, with the ground-truth matrix X^*_{in} of inliers and the estimated inlier matrix $\hat{X}_{in}$ given by the algorithm, we report the relative error $\frac{\|\hat{X}_{in}-X^*_{in}\|_F}{\|X^*_{in}\|_F}$ that reflects the recovery accuracy of inliers. The metric $\frac{\|\hat{X}_{out}-X^*_{out}\|_F}{\|X^*_{out}\|_F}$ is defined similarly for outliers.

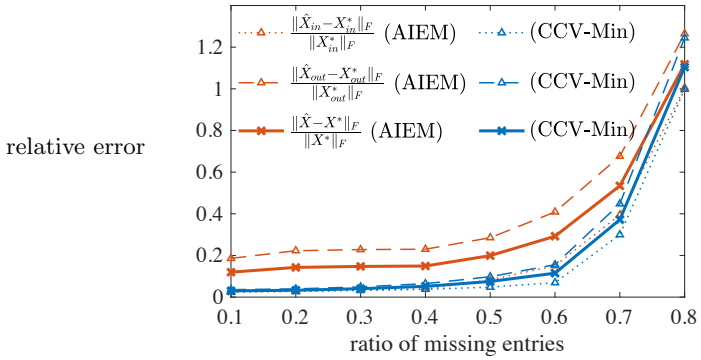
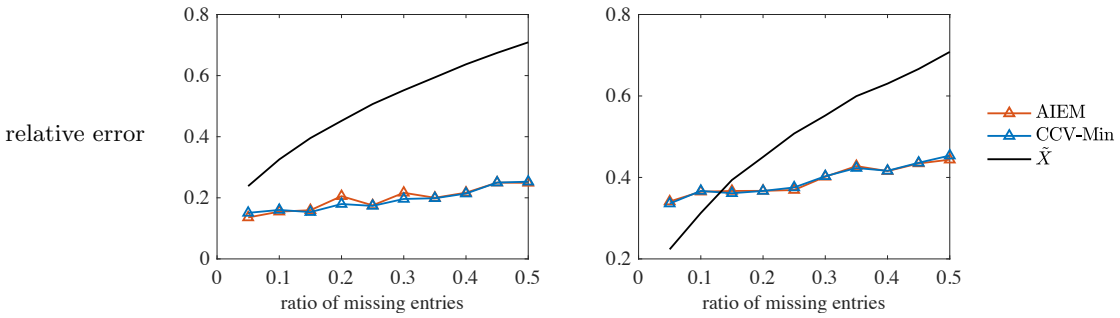


Figure 12: Relative errors for UMC with $m = 50$, $n = 100$, $r = 3$, and outlier ratio 40%.

4.2.3 EXPERIMENTS ON DATA RE-IDENTIFICATION (UMC)

Extending the UPCA experiments of Section 4.1.4, we now evaluate our UMC algorithm on the medical and educational data. We fix the outlier ratio to be 50%, and vary the ratio of missing entries among $0.05 : 0.05 : 0.50$.



(a) high-school scores (b) breast tumor features

Figure 13: Relative estimation errors of the proposed UMC pipeline on real data.

In Figure 13 we present the results. The black curves indicate the distances from the (zero-filled) observed data matrix $\tilde{X}$ to the ground-truth X^*_{in}; the distances grow as the ratio of missing entries increases. The proposed UMC algorithm operates on $\tilde{X}$ to restore X^*, which is intuitively why it gives smaller relative errors as shown by the colored curves. In particular, for 25% missing entries, our algorithm reduces the relative error from 51% (black)

to 17% (red and blue) for the high-school scores (Figure 13a) and from 51% to 37% for the breast tumor features (Figure 13b).

5. Proofs

The proofs of Theorem 2-7 use basic algebraic geometry and we recall the required notions as we go along. An accessible introduction on the subject is Cox et al. (2013), while a more advanced is Harris (2013). We also refer to the works Tsakiris (2023b); Tsakiris et al. (2020) on matrix completion and linear regression without correspondences, whose mathematical analysis is very related. The proof of Theorem 7 is very similar to the proof of Theorem 3 and is omitted.

5.1 Proof of Theorem 2

We first prove the theorem over $\mathbb{C}$, then we transfer the statement over $\mathbb{R}$. We note here that there is nothing special about $\mathbb{R}$ and $\mathbb{C}$ with regards to the problem. Indeed, the same proof applies if one replaces $\mathbb{R}$ with any infinite field $\mathbb{F}$ and $\mathbb{C}$ with the algebraic closure $\bar{\mathbb{F}}$ of $\mathbb{F}$. Set

$$\mathcal{M}_{\mathbb{C}} = \{\, X \in \mathbb{C}^{m\times n} |\, \operatorname{rank}_{\mathbb{C}} X \leq r \,\}$$

and note that since $\mathcal{M}_{\mathbb{C}}$ is irreducible, the intersection of finitely many non-empty open sets in $\mathcal{M}_{\mathbb{C}}$ is itself non-empty and open, and thus dense. Here irreducibility means that $\mathcal{M}_{\mathbb{C}}$ can not be decomposed as the union of two proper subvarieties of $\mathcal{M}_{\mathbb{C}}$.

Lemma 10 *There is an open dense set $\mathcal{U}_1$ in $\mathcal{M}_{\mathbb{C}}$ such that for any $X \in \mathcal{U}_1$ and any $\underline{\pi} = (\Pi_1, \ldots, \Pi_n) \in \prod_{i\in[n]} \mathcal{P}_m$, every $m \times r$ submatrix of $\underline{\pi}(X)$ has rank r.*

Proof First, fix some $\underline{\pi} = (\Pi_1, \ldots, \Pi_n) \in \prod_{i\in[n]} \mathcal{P}_m$ and then some index set $\mathcal{J} = \{j_1, \ldots, j_r\} \subset [n]$. The submatrix $\underline{\pi}(X)_{\mathcal{J}} := [\Pi_{j_1} x_{j_1}, \cdots, \Pi_{j_r} x_{j_r}]$ of $\underline{\pi}(X)$ has rank less than r if and only if all of its $r \times r$ minors are zero. For each subset $\mathcal{I} = \{i_1, \ldots, i_r\} \subset [m]$ we have a polynomial $\det \underline{\pi}(Z)_{\mathcal{I},\mathcal{J}} \in \mathbb{C}[Z]$ where $\underline{\pi}(Z)_{\mathcal{I},\mathcal{J}}$ is the row-submatrix of $\underline{\pi}(Z)_{\mathcal{J}}$ obtained by selecting the rows with index in $\mathcal{I}$. The set of matrices in $\mathbb{C}^{m\times n}$ for which the evaluation of this polynomial is non-zero is an open set, call it $\mathcal{U}_{\underline{\pi},\mathcal{I},\mathcal{J}}$. Then $\underline{\pi}(X)_{\mathcal{J}}$ has rank r if and only if $X \in \mathcal{U}_{\underline{\pi},\mathcal{J}} := \bigcup_{\mathcal{I}} \mathcal{U}_{\underline{\pi},\mathcal{I},\mathcal{J}}$, where $\mathcal{I}$ ranges over all subsets of $[m]$ of cardinality r. As a union of finitely many open sets, $\mathcal{U}_{\underline{\pi},\mathcal{J}}$ is open. Moreover, every $m \times r$ submatrix of $\underline{\pi}(X)$ has rank r if and only if $X \in \mathcal{U}_{\underline{\pi}} := \bigcap_{\mathcal{J}} \mathcal{U}_{\underline{\pi},\mathcal{J}}$, where now $\mathcal{J}$ ranges over all subsets of $[n]$ of cardinality r. $\mathcal{U}_{\underline{\pi}}$ is open because it is the finite intersection of open sets. Finally, every $m \times r$ submatrix of $\underline{\pi}(X)$ has rank r for any $\underline{\pi}$ if and only if X is in the open set $\mathcal{U}_1 := \bigcap_{\underline{\pi}} \mathcal{U}_{\underline{\pi}}$, where the intersection is taken over all $\underline{\pi}$'s.

The proof will be complete once we show that $\mathcal{U}_1$ is non-empty. By what we said above about intersections of finitely many non-empty open sets in an irreducible variety, it is enough to show that each $\mathcal{U}_{\underline{\pi},\mathcal{J}}$ is non-empty. We do this by constructing a specific $X \in \mathcal{U}_{\underline{\pi},\mathcal{J}}$. Recall here that any $\Pi \in \mathcal{P}_m$ is diagonalizable over $\mathbb{C}$ with non-zero eigenvalues. It is an elementary fact in linear algebra that there exists a choice of eigenvector v_k of Π_{j_k} for every $k \in [r]$ such that $v_1, \ldots, v_r$ are linearly independent. Now our X is taken to be the matrix with v_k at column j_k for every $k \in [r]$ and zero everywhere else. Clearly $X \in \mathcal{M}_{\mathbb{C}}$ and moreover

$\underline{\pi}(X)_{\mathcal{J}} = [\Pi_{j_1} x_{j_1} \cdots, \Pi_{j_r} x_{j_r}] = [\Pi_{j_1} v_1 \cdots, \Pi_{j_r} v_r] = [\lambda_1 v_1 \cdots \lambda_r v_r]$, where λ_k is the corresponding eigenvalue of v_k. Since none of the λ_k's is zero, this matrix has rank r, that is $X \in \mathcal{U}_{\underline{\pi},\mathcal{J}}$. ■

Denote by $\mathcal{C}(X)$ the column-space of X and I_m the identity matrix of size $m \times m$. Note also that whenever p is a non-zero polynomial in v variables with coefficients in $\mathbb{C}$, there is always some $\xi \in \mathbb{C}^v$ such that $p(\xi) \neq 0$.

Lemma 11 *There is an open dense set $\mathcal{U}_2$ in $\mathcal{M}_{\mathbb{C}}$ such that for any $X \in \mathcal{U}_2$, we have that $\Pi x_j \notin \mathcal{C}(X)$ for any $\Pi \in \mathcal{P}_m \setminus \{I_m\}$ and any $j \in [n]$.*

Proof $\Pi x_j \notin \mathcal{C}(X)$ if and only if $\text{rank}[X\ \Pi x_j] = r+1$. As in the proof of Lemma 10, this condition is met on an open set $\mathcal{U}_{\Pi,j}$ of $\mathcal{M}_{\mathbb{C}}$ where some $(r+1)\times(r+1)$ determinant of $[X\ \Pi x_j]$ is non-zero. Then the statement of the theorem is true on the open set $\mathcal{U}_2 = \bigcap_{\Pi \in \mathcal{P}_m,\, j\in[n]} \mathcal{U}_{\Pi,j}$. As in the proof of Lemma 10, to show that $\mathcal{U}_2$ is non-empty it suffices to show that each $\mathcal{U}_{\Pi,j}$ is non-empty. We show the existence of an $X \in \mathcal{U}_{\Pi,j}$. Let $Z = (z_{ik})$ be an $m \times r$ matrix of variables over $\mathbb{C}$ and consider the polynomial ring $\mathbb{C}[Z]$. Let us write z_k for the kth column of Z. Since Π is not the identity, there exists some $i \in [m]$ such that z_{i1} is different from the ith element of Πz_1, where z_1 is the first column of Z. Instead, suppose that the variable z_{i1} appears in the i'th coordinate of Πz_1 with $i' \neq i$. Now take any $\mathcal{I} \subset [m]$ with cardinality $r+1$ such that $i, i' \in \mathcal{I}$ and consider $\det[Z\ \Pi z_1]_{\mathcal{I}}$ where $[Z\ \Pi z_1]_{\mathcal{I}}$ is the submatrix of $[Z\ \Pi z_1]$ obtained by selecting the rows with index in $\mathcal{I}$. This is a polynomial of $\mathbb{C}[Z]$ that has the form $\pm z_{i1}^2 \det[z_2 \cdots z_r]_{\mathcal{I}\setminus\{i,i'\}} + \cdots$ where the remaining terms do not involve z_{i1}^v for $v > 1$. Since the entries of Z are algebraically independent, $\det[z_2 \cdots z_r]_{\mathcal{I}\setminus\{i,i'\}}$ is a non-zero polynomial. We conclude that $\det[Z\ \Pi z_1]_{\mathcal{I}}$ is also a non-zero polynomial. Hence there exists some $Z' \in \mathbb{C}^{m\times r}$ such that $\det[Z'\ \Pi z_1']_{\mathcal{I}} \neq 0$. Now define X by setting $x_j = z_1'$, $x_{j_k} = z_k'$, $k \in [r]$ for any choice of j_k's distinct from j, and zeros everywhere else. By construction $X \in \mathcal{U}_{\Pi,j}$. ■

Let $f : \mathbb{C}^{m\times r} \times \mathbb{C}^{r\times n} \to \mathcal{M}_{\mathbb{C}}$ be the surjective map given by $f(B', C') = B'C'$.

Lemma 12 *There is an open dense set $\mathcal{U}_3$ in $\mathcal{M}_{\mathbb{C}}$ such that for any $X \in \mathcal{U}_3$, we have that for any $j \in [n]$, any $\mathcal{J} = \{j_1, \ldots, j_r\} \subset [n]$ with $j \notin \mathcal{J}$ and any $\Pi_1, \ldots, \Pi_r \in \mathcal{P}_m$ not all identities, it holds that* $\text{rank}[x_j\ \Pi_1 x_{j_1} \cdots \Pi_r x_{j_r}] = r+1$.

Proof With $j, \mathcal{J}$ and Π_k's fixed, the set $\mathcal{U}_{j,\mathcal{J},\Pi_1,\ldots,\Pi_r}$ of X's in $\mathcal{M}_{\mathbb{C}}$ for which the rank of $[x_j\ \Pi_1 x_{j_1} \cdots \Pi_r x_{j_r}]$ is $r+1$, is open. Indeed, this is defined by the non-simultaneous vanishing of all $(r+1)\times(r+1)$ minors of $[z_j\ \Pi_1 z_{j_1} \cdots \Pi_r z_{j_r}]$, where z_k is the kth column of the matrix of variables Z from the proof of Lemma 11. We note that these are polynomials in Z with integer coefficients. Set $\mathcal{U}_3 = \bigcap_{j,\mathcal{J},\Pi_1,\ldots,\Pi_r} \mathcal{U}_{j,\mathcal{J},\Pi_1,\ldots,\Pi_r}$ where the intersection is taken over all choices of $j, \mathcal{J}, \Pi_1, \ldots, \Pi_r$ as in the statement of the lemma. As in the proof of Lemma 10, the set $\mathcal{U}_3$ is open and to show that it is non-empty is suffices to show that each $\mathcal{U}_{j,\mathcal{J},\Pi_1,\ldots,\Pi_r}$ is non-empty.

Let $\mathcal{U}_1, \mathcal{U}_2$ be the open sets of Lemmas 10 and 11. Since $\mathcal{M}_{\mathbb{C}}$ is irreducible and $\mathcal{U}_1, \mathcal{U}_2$ are open and non-empty, we have that $\mathcal{U}_1 \cap \mathcal{U}_2$ is non-empty. Since f is surjective, $f^{-1}(\mathcal{U}_1 \cap \mathcal{U}_2)$ is also non-empty. Take any $(B', C') \in f^{-1}(\mathcal{U}_1 \cap \mathcal{U}_2)$. By definition, the

rank of $[\Pi_1 B' c'_{j_1} \cdots \Pi_r B' c'_{j_r}]$ is r. By hypothesis, there is some $k \in [r]$ such that Π_k is not the identity and thus again by definition we have $\operatorname{rank}[B'\ \Pi_k B' c'_{j_k}] = r+1$. Consequently, $\Pi_k B' c'_{j_k} \notin \mathcal{C}(B')$ and so the two r-dimensional subspaces $\mathcal{C}(B')$ and $\mathcal{C}\big([\Pi_1 B' c'_{j_1} \cdots \Pi_r B' c'_{j_r}]\big)$ are distinct. Thus there exists some $c'' \in \mathbb{C}^r$ such that $B'c'' \notin \mathcal{C}\big([\Pi_1 B' c'_{j_1} \cdots \Pi_r B' c'_{j_r}]\big)$. Define $C'' \in \mathbb{C}^{r\times n}$ by setting $c''_\nu = c'_\nu$ for every $\nu \neq j$ and $c''_j = c''$. Then by construction $B'C'' \in \mathcal{U}_{j,\mathcal{J},\Pi_1,\ldots,\Pi_r}$. ■

Take $X^* = [x_1^* \cdots x_n^*] \in \mathcal{U}_3$ and let $\tilde{X} = [\tilde{\Pi}_1 x_1^* \cdots \tilde{\Pi}_n x_n^*]$. Now $\operatorname{rank} \tilde{X} = \operatorname{rank} \tilde{\Pi}_1^{-1}\tilde{X} = \operatorname{rank}[x_1^*\ \tilde{\Pi}_1^{-1}\tilde{\Pi}_2 x_2^* \cdots \tilde{\Pi}_1^{-1}\tilde{\Pi}_n x_n^*]$. If there is some $k \geq 2$ such that $\tilde{\Pi}_1 \neq \tilde{\Pi}_k$, by Lemma 12 any $m \times (r+1)$ submatrix of $\tilde{\Pi}_1^{-1}\tilde{X}$ that contains columns 1 and k will have rank $r+1$. On the other hand, when all $\tilde{\Pi}_k$'s are equal for $k \in [n]$, the rank of $\tilde{X}$ is r by Lemma 10. This concludes the proof of the theorem over $\mathbb{C}$ with the claimed open set being $\mathcal{U}_3$, which we denote in the sequel by $\mathcal{U}_{\mathbb{C}}$.

Set $\mathcal{M}_{\mathbb{R}} = \{\, X \in \mathbb{R}^{m\times n} | \operatorname{rank}_{\mathbb{R}} X \leq r \,\}$. There is an inclusion of sets $i : \mathcal{M}_{\mathbb{R}} \hookrightarrow \mathcal{M}_{\mathbb{C}}$ where for $X \in \mathcal{M}_{\mathbb{R}}$ we view $i(X)$ as the complex matrix associated to X. The reason for this inclusion is that if the columns of X generate an r-dimensional subspace over $\mathbb{R}$, then they generate an r-dimensional subspace over $\mathbb{C}$. To finish the proof, it suffices to show the existence of a non-empty open set $\mathcal{U}_{\mathbb{R}}$ in $\mathcal{M}_{\mathbb{R}}$ such that $i(\mathcal{U}_{\mathbb{R}}) \subset \mathcal{U}_{\mathbb{C}}$. This comes from two key ingredients. The first one is the observation that the polynomials that induce $\mathcal{U}_{\mathbb{C}}$, i.e. the polynomials of $\mathbb{C}[Z]$ whose non-simultaneous vanishing indicates membership of a point $X \in \mathcal{M}_{\mathbb{C}}$ in $\mathcal{U}_{\mathbb{C}}$, they have integer and thus real coefficients. This can be seen by inspecting the proof of Lemma 12. Call the set of these polynomials $\mathfrak{p}_{\mathcal{U}} \subset \mathbb{Z}[Z]$. For the second ingredient, let $\mathfrak{p}_{\mathcal{M}} \subset \mathbb{Z}[Z]$ be the set of all $(r+1)\times(r+1)$ minors of the matrix of variables Z. It is a matter of linear algebra that $\mathcal{M}_{\mathbb{R}}$ and $\mathcal{M}_{\mathbb{C}}$ are the common roots of the polynomial system $\mathfrak{p}_{\mathcal{M}}$ over $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ respectively. What is instead a difficult theorem in commutative algebra is that the following algebraic converse is true; see Section 2.6 in Tsakiris (2023b): a polynomial $q \in \mathbb{R}[Z]$ vanishes on every point of $\mathcal{M}_{\mathbb{R}}$ if and only if it is a polynomial combination of elements of $\mathfrak{p}_{\mathcal{M}}$, that is if and only if $q = \sum_{p\in\mathfrak{p}_{\mathcal{M}}} c_p\, p$ for some c_p's in $\mathbb{R}[Z]$. This statement also holds true if we replace $\mathbb{R}$ with $\mathbb{C}$. Now the set $\mathcal{U}_{\mathbb{C}}$ consists of those points of $\mathcal{M}_{\mathbb{C}}$ that are roots of the polynomial system $\mathfrak{p}_{\mathcal{M}}$ but not of $\mathfrak{p}_{\mathcal{U}}$. Since $\mathcal{U}_{\mathbb{C}}$ is non-empty, not all polynomials in $\mathfrak{p}_{\mathcal{U}}$ are polynomial combinations of $\mathfrak{p}_{\mathcal{M}}$. But then, by what we just said, not all points of $\mathcal{M}_{\mathbb{R}}$ are common roots of $\mathfrak{p}_{\mathcal{U}}$. This means that the open set of $\mathcal{M}_{\mathbb{R}}$ defined by the non-simultaneous vanishing of all polynonials in $\mathfrak{p}_{\mathcal{U}}$ is non-empty. This open set is the claimed $\mathcal{U}$.

5.2 Proof of Theorem 3 and Theorem 7

Let $\mathcal{U}_1$ be the open set of Theorem 1. Let $\mathcal{U}_2$ be the set of X's for which $\mathcal{C}(X)$ does not drop dimension under projection onto any r coordinates. This set is open in $\mathcal{M}$ because $X \in \mathcal{U}_2$ if and only if for any $\mathcal{I} \subset [m]$ of cardinality r not all $r\times r$ minors of $X_{\mathcal{I}}$ are zero, $X_{\mathcal{I}}$ being the row-submatrix of X obtained by selecting the rows with index in $\mathcal{I}$. Set $\mathcal{U} = \mathcal{U}_1 \cap \mathcal{U}_2$. Then for any $X^* \in \mathcal{U}$ and any $\Pi \in \mathcal{P}_m$ there is a unique factorization $\Pi X^* = B^*_\Pi C^*_\Pi$ with the top $r\times r$ block of $B^*_\Pi \in \mathbb{R}^{m\times r}$ being the identity. Since $\bar{p}_{\ell,j}(\tilde{X}) = \bar{p}_{\ell,j}(X^*) = \bar{p}_{\ell,j}(\Pi X^*) = \bar{p}_{\ell,j}(B^*_\Pi C^*_\Pi)$ we

have that $(B^*_\Pi, C^*_\Pi) \in \mathcal{Y}_{X^*}$ for every $\Pi \in \mathcal{P}_m$. For the reverse direction we recall a basic fact (see Lemma 2 of Song et al. (2018)):

Lemma 13 *Fix any* $j \in [n]$. *Suppose that* $\xi_1, \xi_2 \in \mathbb{R}^m$ *are such that* $\bar{p}_{\ell,j}(\xi_1) = \bar{p}_{\ell,j}(\xi_2)$ *for every* $\ell \in [m]$. *Then* $\xi_1 = \Pi\xi_2$ *for some* $\Pi \in \mathcal{P}_m$.

Now let $(B', C') \in \mathcal{Y}_{X^*}$ and write c'_j for the jth column of C'. For a fixed $j \in [n]$ the equations $q_{\ell,j}(B', C') = 0$ are equivalent to $\bar{p}_{\ell,j}(B'c'_j) = \bar{p}_{\ell,j}(x^*_j)$ for every $\ell \in [m]$. By Lemma 13 there must exist some $\Pi_j \in \mathcal{P}_m$ such that $B'c'_j = \Pi_j x^*_j$. This is true for every $j \in [n]$ so that $B'C' = [\Pi_1 x^*_1 \cdots \Pi_n x^*_n]$. This implies that $\operatorname{rank}[\Pi_1 x^*_1 \cdots \Pi_n x^*_n] = r$. Since $X^* \in \mathcal{U}$, Theorem 1 gives that all Π's must be the same permutation $\Pi \in \mathcal{P}_m$, so that $B'C' = \Pi X^*$. Since by construction for any $(B'', C'') \in \mathcal{Y}_{X^*}$ the top $r \times r$ block of B'' is the identity, we have that $B' = B^*_\Pi$ and thus necessarily $C' = C^*_\Pi$.

The proof of Theorem 7 is very similar to the proof of Theorem 3 and is omitted.

5.3 Proof of Theorem 4

First notice that S^* contains exactly $\mu(I_m)$ columns of $\tilde{X}$ according to Lemma 11. Now we suppose $\tilde{x}_{j_1}, \ldots, \tilde{x}_{j_{\mu(I_m)}}$ are $\mu(I_m) \geq r+1$ points in $\tilde{X}$ such that not all $\Pi_{j_1}, \ldots, \Pi_{j_{\mu(I_m)}}$ are the identity I_m. Since $\mu(I_m) > \mu(\Pi)$ for any other $\Pi \neq I_m$, it is impossible that $\Pi_{j_1} = \cdots = \Pi_{j_{\mu(I_m)}}$. According to Theorem 1, the points $\tilde{x}_{j_1}, \ldots, \tilde{x}_{\mu(I_m)}$ span a subspace of dimension at least $r+1$.

5.4 Proof of Theorem 6

As before, we first consider the problem over $\mathbb{C}$. The transfer to $\mathbb{R}$ follows the same argument as in the proof of Theorem 1 in Tsakiris (2023b) and is omitted. We use the same letters $\underline{p}_\Omega : \mathcal{M}_\mathbb{C} \to \mathbb{C}^\Omega$ and $\underline{\pi}_\Omega : \mathbb{C}^\Omega \to \mathbb{C}^\Omega$ to indicate the same maps between the corresponding spaces over $\mathbb{C}$.

The map $\underline{p}_\Omega$ is defined by $x_{ij} \mapsto x_{ij}$ if $(i,j) \in \Omega$ and $x_{ij} \mapsto 0$ if $(i,j) \in \Omega^c$. These are polynomial functions in the x_{ij}'s so that $\underline{p}_\Omega$ is a *morphism* of irreducible algebraic varieties. In particular, $\underline{p}_\Omega$ is continuous in the Zariski topology, and thus inverse images of open sets are open. Now, under the hypothesis on Ω, it was shown in Tsakiris (2023b, 2024) that Ω is generically finitely completable. This is equivalent to the existence of a dense open set U_0 in $\mathcal{M}_\mathbb{C}$ such that for every $X^* \in U_0$ the fiber $\underline{p}^{-1}_\Omega(X^*)$ is a finite set. It is also equivalent to saying that $\underline{p}_\Omega$ is dominant, in the sense that the image $\underline{p}_\Omega(\mathcal{M}_\mathbb{C})$ of $\underline{p}_\Omega$ is a dense set in $\mathbb{C}^\Omega$, that is the closure of $\underline{p}_\Omega(\mathcal{M}_\mathbb{C})$ is $\mathbb{C}^\Omega$.

Lemma 14 *The image* $\underline{p}_\Omega(U_0)$ *of* U_0 *under* $\underline{p}_\Omega$ *contains a non-empty open set of* $\mathbb{C}^\Omega$.

Proof A locally closed set is the intersection of a closed set with an open set. A constructible set is the finite union of locally closed sets. Chevalley's theorem says that a morphism of algebraic varieties takes a constructible set to a constructible set. Since U_0 is open, it is constructible, and thus $\underline{p}_\Omega(U_0)$ is also constructible. Hence, there exists a positive integer s, closed sets $Y_k, k \in [s]$ and non-empty open sets $V_k, k \in [s]$ of $\mathcal{M}_\mathbb{C}$ such that $\underline{p}_\Omega(U_0) = \bigcup_{k\in[s]} Y_k \cap V_k$. For the sake of contradiction, suppose that $\underline{p}_\Omega(U_0)$ does not contain any non-empty open set. Then necessarily all Y_k's are proper closed

sets, otherwise some V_k is contained in $\underline{p}_\Omega(U_0)$. Hence $\underline{p}_\Omega(U_0)$ is contained in the closed set $Y = \bigcup_{k\in[s]} Y_k$. This is a proper closed set because $\mathcal{M}_{\mathbb{C}}$ is irreducible. But then the closure of $\underline{p}_\Omega(U_0)$ is contained in Y, which contradicts the fact that $\underline{p}_\Omega(U_0)$ is dense in $\mathbb{C}^\Omega$. ■

By Lemma 14 $\underline{p}_\Omega(U_0)$ contains a non-empty open set V_0. Now each $\underline{\pi}_\Omega \in \mathcal{P}_\Omega$ is a bijective polynomial function on $\mathbb{C}^\Omega$. Hence $\underline{\pi}_\Omega : \mathbb{C}^\Omega \to \mathbb{C}^\Omega$ is a homeomorphism of topological spaces, so that $\underline{\pi}_\Omega(V_0)$ is a non-empty open set of $\mathbb{C}^\Omega$. Define $V = \bigcap_{\underline{\pi}_\Omega\in\mathcal{P}_\Omega} \underline{\pi}_\Omega(V_0)$. It is a non-empty open set of $\mathbb{C}^\Omega$.

Lemma 15 *For any $\underline{\pi}'_\Omega \in \mathcal{P}_\Omega$ we have that $\underline{\pi}'_\Omega(V) = V$.*

Proof If $f : S \to T$ is a one-to-one (injective) function of sets and S_1, S_2 are subsets of S, then we always have $f(S_1 \cap S_2) = f(S_1) \cap f(S_2)$. Hence $\underline{\pi}'_\Omega(V) = \bigcap_{\underline{\pi}_\Omega\in\mathcal{P}_\Omega} \underline{\pi}'_\Omega \circ \underline{\pi}_\Omega(V_0)$. But $\mathcal{P}_\Omega$ is a group under composition of functions so that the coset $\underline{\pi}'_\Omega\mathcal{P}_\Omega = \{\underline{\pi}'_\Omega \circ \underline{\pi}_\Omega \,|\, \underline{\pi}_\Omega \in \mathcal{P}_\Omega\}$ is equal to $\mathcal{P}_\Omega$. ■

As noted earlier, $\underline{p}_\Omega$ is continuous and so the set $U = \underline{p}_\Omega^{-1}(V)$ is open. Moreover, it is non-empty since V is a subset of V_0 which is a subset of the image of $\underline{p}_\Omega$. Hence U is dense in $\mathcal{M}_{\mathbb{C}}$. Suppose that $X^* \in U$ and set $\tilde{X} = \tilde{\underline{\pi}}_\Omega \underline{p}_\Omega(X^*)$ for some $\tilde{\underline{\pi}}_\Omega \in \mathcal{P}_\Omega$. It is enough to show that as $\underline{\pi}_\Omega$ ranges in $\mathcal{P}_\Omega$ there are only finitely many $X \in \mathcal{M}_{\mathbb{C}}$ such that $\underline{\pi}_\Omega \underline{p}_\Omega(X) = \tilde{X}$. This equation can be written as $\underline{p}_\Omega(X) = \underline{\pi}'_\Omega \underline{p}_\Omega(X^*)$ with $\underline{\pi}'_\Omega = \underline{\pi}_\Omega^{-1} \circ \tilde{\underline{\pi}}_\Omega$. Since $X^* \in U$, $\underline{p}_\Omega(X^*) \in V$. Thus $\underline{\pi}'_\Omega \underline{p}_\Omega(X^*) \in \underline{\pi}'_\Omega(V)$. By Lemma 15 we have $\underline{\pi}'_\Omega(V) = V$ and so $\underline{\pi}'_\Omega \underline{p}_\Omega(X^*) \in V$. But $V \subset V_0$ so that $\underline{\pi}'_\Omega \underline{p}_\Omega(X^*) \in V_0$. Since V_0 is a subset of $\underline{p}_\Omega(U_0)$, there is some $X_0 \in U_0$ such that $\underline{\pi}'_\Omega \underline{p}_\Omega(X^*) = \underline{p}_\Omega(X_0)$. By definition of U_0 the fiber $\underline{p}_\Omega^{-1}\big(\underline{p}_\Omega(X_0)\big)$ is a finite set. But $\underline{p}_\Omega^{-1}\big(\underline{p}_\Omega(X_0)\big) = \underline{p}_\Omega^{-1}\big(\underline{\pi}'_\Omega \underline{p}_\Omega(X^*)\big)$ so that there are finitely many X's in $\mathcal{M}$ that map under $\underline{p}_\Omega$ to $\underline{\pi}'_\Omega \underline{p}_\Omega(X^*)$. As there are finitely many choices for $\underline{\pi}'_\Omega$, we are done.

Acknowledgments

The last author acknowledges the support of the National Key R&D Program of China (2023YFA1009402) and of the CAS Project for Young Scientists in Basic Research (YSBR-034).

References

Abubakar Abid and James Zou. A stochastic expectation-maximization approach to shuffled linear regression. In *Annual Allerton Conference on Communication, Control, and Computing*, 2018.

John M Abowd. Stepping-up: The Census Bureau tries to be a good data steward in the 21st century. 2019.

Manfred Antoni and Rainer Schnell. The past, present and future of the german record linkage center. *Jahrbücher für Nationalökonomie und Statistik*, 239(2):319–331, 2019.

Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

Mona Azadkia and Fadoua Balabdaoui. Linear regression with unmatched data: a deconvolution perspective. *arXiv preprint arXiv:2207.06320*, 2022.

Maria-Florina Balcan, Zhao Liang, Yingyu Song, David P Woodruff, and Hongyang Zhang. Non-convex matrix completion and related problems via strong duality. *Journal of Machine Learning Research*, 20(102):1–56, 2019.

Laura Balzano, Robert Nowak, and Benjamin Recht. Online identification and tracking of subspaces from highly incomplete information. In *Annual Allerton Conference on Communication, Control, and Computing*, 2010.

Laura Balzano, Yuejie Chi, and Yue M Lu. Streaming PCA and subspace tracking: The missing data case. *Proceedings of the IEEE*, 106(8):1293–1310, 2018.

Daniel J Bates, Jonathan D Hauenstein, Andrew J Sommese, and Charles W Wampler. Bertini: Software for numerical algebraic geometry.

Amir Beck and Nili Guttmann-Beck. FOM–a MATLAB toolbox of first-order methods for solving convex optimization problems. *Optimization Methods and Software*, 34(1):172–193, 2019.

Daniel Irving Bernstein. Completion of tree metrics and rank 2 matrices. *Linear Algebra and its Applications*, 533:1–13, 2017.

Dimitris Bertsimas and Michael Lingzhi Li. Fast exact matrix completion: A unified optimization framework for matrix completion. *Journal of Machine Learning Research*, 21 (1):9399–9441, 2020.

Paul Breiding, Sara Kališnik, Bernd Sturmfels, and Madeleine Weinstein. Learning algebraic varieties from samples. *Revista Matemática Complutense*, 31(3):545–593, 2018.

Paul Breiding, Fulvio Gesmundo, Mateusz Michałek, and Nick Vannieuwenhoven. Algebraic compressed sensing. *Applied and Computational Harmonic Analysis*, 65:374–406, 2023.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Emmanuel J Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

Yudong Chen, Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust matrix completion and corrupted columns. In *International Conference on Machine Learning*, 2011.

Yudong Chen, Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Matrix completion with column manipulation: Near-optimal sample-robustness-rank tradeoffs. *IEEE Transactions on Information Theory*, 62(1):503–526, 2015.

David Cox, John Little, and Donal OShea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer Science & Business Media, 2013.

Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4): 608–622, 2016.

Tianyu Ding, Zhihui Zhu, René Vidal, and Daniel P Robinson. Dual principal component pursuit for robust subspace learning: Theory and algorithms for a holistic approach. In *International Conference on Machine Learning*, 2021.

Ivan Dokmanić. Permutations unlabeled beyond sampling unknown. *IEEE Signal Processing Letters*, 26(6):823–827, 2019.

Josep Domingo-Ferrer and Krishnamurty Muralidhar. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences*, 337:11–24, 2016.

Armin Eftekhari, Gregory Ongie, Laura Balzano, and Michael B Wakin. Streaming principal component analysis from incomplete data. *Journal of Machine Learning Research*, 20: 86–1, 2019.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

Golnoosh Elhami, Adam Scholefield, Benjamin Bejar Haro, and Martin Vetterli. Unlabeled sensing: Reconstruction algorithm and theoretical guarantees. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

Brian Eriksson, Laura Balzano, and Robert Nowak. High-rank matrix completion. In *Artificial Intelligence and Statistics*, 2012.

Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

Ravi Sastry Ganti, Laura Balzano, and Rebecca Willett. Matrix completion under monotonic single index models. In *Advances in Neural Information Processing Systems*, 2015.

Athinodoros S Georghiades, Peter N Belhumeur, and David J Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

Alice Guionnet, Justin Ko, Florent Krzakala, Pierre Mergny, and Lenka Zdeborová. Spectral phase transitions in non-linear wigner spiked models. *arXiv preprint arXiv:2310.14055*, 2023.

Joe Harris. *Algebraic geometry: a first course*, volume 133. Springer Science & Business Media, 2013.

Xianmang He, Yanghua Xiao, Yujia Li, Qing Wang, Wei Wang, and Baile Shi. Permutation anonymization: Improving anatomy for privacy preservation in data publication. In *New Frontiers in Applied Data Mining: PAKDD 2011 International Workshops*, 2012.

Daniel J Hsu, Kevin Shi, and Xiaorui Sun. Linear regression without correspondence. In *Advances in Neural Information Processing Systems*, 2017.

Pan Ji, Hongdong Li, Mathieu Salzmann, and Yuchao Dai. Robust motion segmentation with unknown correspondences. In *European Conference on Computer Vision*, 2014.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

Franz Király and Ryota Tomioka. A combinatorial algebraic approach for the identifiability of low-rank matrix completion. In *International Conference on Machine Learning*, 2012.

Franz J Király, Louis Theran, and Ryota Tomioka. The algebraic combinatorial approach for low-rank matrix completion. *Journal of Machine Learning Research*, 16:1391–1436, 2015.

Steven L Kleiman and John Landolfi. Geometry and deformation of special schubert varieties. *Compositio Mathematica*, 23(4):407–434, 1971.

Viktor Larsson, Kalle Astrom, and Magnus Oskarsson. Efficient solvers for minimal problems by syzygy-based reduction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Andreas Lenz, Paul H Siegel, Antonia Wachter-Zeh, and Eitan Yaakobi. The noisy drawing channel: Reliable data storage in DNA sequences. *IEEE Transactions on Information Theory*, 69(5):2757–2778, 2022.

Gilad Lerman and Tyler Maunu. Fast, robust and non-convex subspace recovery. *Information and Inference: A Journal of the IMA*, 7(2):277–336, 2018.

Rong Ma, T Tony Cai, and Hongzhe Li. Optimal permutation recovery in permuted monotone matrix model. *Journal of the American Statistical Association*, 116(535):1358–1372, 2021.

Angshul Majumdar and Rabab Kreidieh Ward. Some empirical advances in matrix completion. *Signal Processing*, 91(5):1334–1338, 2011.

Stefano Marano and Peter Willett. Making decisions by unlabeled bits. *IEEE Transactions on Signal Processing*, 68:2935–2947, 2020.

Rahul Mazumder and Haoyue Wang. Linear regression with partially mismatched data: local search with theoretical guarantees. *Mathematical Programming*, 197(2):1265–1303, 2023.

Pierre Mergny, Justin Ko, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of non-linear low-rank matrix estimation. *arXiv preprint arXiv:2403.04234*, 2024.

Krishnamurty Muralidhar. Record re-identification of swapped numerical microdata. *Journal of Information Privacy and Security*, 13(1):34–45, 2017.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, 2008.

Amin Nejatbakhsh and Erdem Varol. Neuron matching in c. elegans with robust approximate linear regression without correspondence. In *Winter Conference on Applications of Computer Vision*, 2021.

Valeria Nikolaenko, Udi Weinsberg, Stratis Ioannidis, Marc Joye, Dan Boneh, and Nina Taft. Privacy-preserving ridge regression on hundreds of millions of records. In *IEEE Symposium on Security and Privacy*, 2013.

Ricardo Oliveira, Joao Costeira, and Joao Xavier. Optimal point correspondence through the use of rank constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

Efe Onaran and Soledad Villar. Shuffled linear regression through graduated convex relaxation. *arXiv preprint arXiv:2209.15608*, 2022.

Toan C Ong, Michael V Mannino, Lisa M Schilling, and Michael G Kahn. Improving record linkage performance in the presence of missing linkage data. *Journal of Biomedical Informatics*, 52:43–54, 2014.

Greg Ongie, Rebecca Willett, Robert D Nowak, and Laura Balzano. Algebraic variety models for high-rank matrix completion. In *International Conference on Machine Learning*, 2017.

Greg Ongie, Daniel Pimentel-Alarcón, Laura Balzano, Rebecca Willett, and Robert D Nowak. Tensor methods for nonlinear matrix completion. *SIAM Journal on Mathematics of Data Science*, 3(1):253–279, 2021.

Liangzu Peng and Manolis C Tsakiris. Linear regression without correspondences via concave minimization. *IEEE Signal Processing Letters*, 27:1580–1584, 2020.

Liangzu Peng and Manolis C Tsakiris. Homomorphic sensing of subspace arrangements. *Applied and Computational Harmonic Analysis*, 55:466–485, 2021.

Liangzu Peng and René Vidal. Block coordinate descent on smooth manifolds: Convergence theory and twenty-one examples. In *Conference on Parsimony and Learning*, 2023.

Liangzu Peng, Christian Kümmerle, and René Vidal. Global linear and local superlinear convergence of IRLS for non-smooth robust regression. In *Advances in Neural Information Processing Systems*, 2022.

Liangzu Peng, Christian Kümmerle, and René Vidal. On the convergence of IRLS and its variants in outlier-robust estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

Mostafa Rahmani and George K Atia. Coherence pursuit: Fast, simple, and robust principal component analysis. *IEEE Transactions on Signal Processing*, 65(23):6260–6275, 2017.

Thilina Ranbaduge, Peter Christen, and Rainer Schnell. Large scale record linkage in the presence of missing data. *arXiv preprint arXiv:2104.09677*, 2021.

Aditya Narayan Ravi, Alireza Vahid, and Ilan Shomorony. Coded shotgun sequencing. *IEEE Journal on Selected Areas in Information Theory*, 3(1):147–159, 2022.

Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. Deeppermnet: Visual permutation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. Visual permutation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12): 3100–3114, 2018.

Allen Schmaltz, Alexander M Rush, and Stuart M Shieber. Word ordering without syntax. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, 2017.

Ilan Shomorony and Reinhard Heckel. DNA-based storage: Models and fundamental limits. *IEEE Transactions on Information Theory*, 67(6):3675–3689, 2021.

Amit Singer and Mihai Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1621–1641, 2010.

Martin Slawski and Emanuel Ben-David. Linear regression with sparsely permuted data. *Electronic Journal of Statistics*, 13:1–36, 2019.

Martin Slawski, Emanuel Ben-David, and Ping Li. Two-stage approach to multivariate linear regression with sparsely mismatched data. *Journal of Machine Learning Research*, 21(1):8422–8463, 2020.

Martin Slawski, Guoqing Diao, and Emanuel Ben-David. A pseudo-likelihood approach to linear regression with partially shuffled data. *Journal of Computational and Graphical Statistics*, 30(4):991–1003, 2021.

Mahdi Soltanolkotabi and Emmanuel J Candés. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

Xuming Song, Hayoung Choi, and Yuanming Shi. Permuted linear model for header-free communication via symmetric polynomials. In *IEEE International Symposium on Information Theory*, 2018.

Bernd Sturmfels and Andrei Zelevinsky. Maximal minors and their leading terms. *Advances in mathematics*, 98(1):65–112, 1993.

Julián Tachella, Dongdong Chen, and Mike Davies. Sensing theorems for unsupervised learning in linear inverse problems. *Journal of Machine Learning Research*, pages 1–45, 2023.

Jared Tanner and Ke Wei. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

Manolis Tsakiris. Results on the algebraic matroid of the determinantal variety. *Transactions of the American Mathematical Society*, 377(01):731–751, 2024.

Manolis Tsakiris and Liangzu Peng. Homomorphic sensing. In *International Conference on Machine Learning*, 2019.

Manolis Tsakiris and René Vidal. Theoretical analysis of sparse subspace clustering with missing entries. In *International Conference on Machine Learning*, 2018a.

Manolis C Tsakiris. Determinantal conditions for homomorphic sensing. *Linear Algebra and its Applications*, 656:210–223, 2023a.

Manolis C Tsakiris. Low-rank matrix completion theory via Plücker coordinates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b.

Manolis C Tsakiris and René Vidal. Dual principal component pursuit. In *IEEE International Conference on Computer Vision Workshop*, 2015.

Manolis C Tsakiris and René Vidal. Hyperplane clustering via dual principal component pursuit. In *International Conference on Machine Learning*, 2017.

Manolis C Tsakiris and René Vidal. Dual principal component pursuit. *Journal of Machine Learning Research*, 19(18):1–50, 2018b.

Manolis C Tsakiris, Liangzu Peng, Aldo Conca, Laurent Kneip, Yuanming Shi, and Hayoung Choi. An algebraic-geometric approach for linear regression without correspondences. *IEEE Transactions on Information Theory*, 66(8):5130–5144, 2020.

Jayakrishnan Unnikrishnan, Saeid Haghighatshoar, and Martin Vetterli. Unlabeled sensing: Solving a linear system with unordered measurements. In *Annual Allerton Conference on Communication, Control, and Computing*, 2015.

Jayakrishnan Unnikrishnan, Saeid Haghighatshoar, and Martin Vetterli. Unlabeled sensing with random linear measurements. *IEEE Transactions on Information Theory*, 64(5): 3237–3253, 2018.

Namrata Vaswani and Praneeth Narayanamurthy. Static and dynamic robust PCA and matrix completion: A review. *Proceedings of the IEEE*, 106(8):1359–1379, 2018.

Namrata Vaswani, Thierry Bouwmans, Sajid Javed, and Praneeth Narayanamurthy. Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery. *IEEE Signal Processing Magazine*, 35(4):32–55, 2018.

Guanyu Wang, Stefano Marano, Jiang Zhu, and Zhiwei Xu. Target localization by unlabeled range measurements. *IEEE Transactions on Signal Processing*, 68:6607–6620, 2020.

Nir Weinberger and Neri Merhav. The DNA storage channel: Capacity and error probability bounds. *IEEE Transactions on Information Theory*, 68(9):5657–5700, 2022.

Yujia Xie, Yixiu Mao, Simiao Zuo, Hongteng Xu, Xiaojing Ye, Tuo Zhao, and Hongyuan Zha. A hypergradient approach to robust regression without correspondence. In *International Conference on Learning Representations*, 2021.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

Congyuan Yang, Daniel Robinson, and René Vidal. Sparse subspace clustering with missing entries. In *International Conference on Machine Learning*, 2015.

Yunzhen Yao, Liangzu Peng, and Manolis Tsakiris. Unlabeled principal component analysis. *Advances in Neural Information Processing Systems*, 2021a.

Yunzhen Yao, Liangzu Peng, and Manolis C Tsakiris. Unsigned matrix completion. In *International Symposium on Information Theory*, 2021b.

Chong You, Daniel P Robinson, and René Vidal. Provable self-representation based outlier detection in a union of subspaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Zinan Zeng, Tsung-Han Chan, Kui Jia, and Dong Xu. Finding correspondence from multiple images via sparse and low-rank decomposition. In *European Conference on Computer Vision*, 2012.

Hang Zhang and Ping Li. Optimal estimator for unlabeled linear regression. In *International Conference on Machine Learning*, 2020.

Teng Zhang and Yi Yang. Robust PCA by manifold optimization. *Journal of Machine Learning Research*, 19(1):3101–3139, 2018.

Zhihui Zhu, Yifan Wang, Daniel Robinson, Daniel Naiman, René Vidal, and Manolis Tsakiris. Dual principal component pursuit: Improved analysis and efficient algorithms. In *Advances in Neural Information Processing Systems*, 2018.