Low-Rank Matrix Recovery with Unknown Correspondence

Zhiwei Tang\textsuperscript{1} \quad Tsung-Hui Chang\textsuperscript{1} \quad Xiaojing Ye\textsuperscript{2} \quad Hongyuan Zha\textsuperscript{1}

\textsuperscript{1}The Chinese University of Hong Kong, Shenzhen  \quad \textsuperscript{2}Georgia State University

Abstract

We study a matrix recovery problem with unknown correspondence: given the observation matrix $M_o = [A, \hat{P}B]$, where $\hat{P}$ is an unknown permutation matrix, we aim to recover the underlying matrix $M = [A, B]$. Such problem commonly arises in many applications where heterogeneous data are utilized and the correspondence among them are unknown, e.g., due to data mishandling or privacy concern. We show that, in some applications, it is possible to recover $M$ via solving a nuclear norm minimization problem. Moreover, under a proper low-rank condition on $M$, we derive a non-asymptotic error bound for the recovery of $M$. We propose an algorithm, $M^3O$ (Matrix recovery via Min-Max Optimization) which recasts this combinatorial problem as a continuous minimization problem and solves it by proximal gradient with a Max-Oracle. $M^3O$ can also be applied to a more general scenario where we have missing entries in $M_o$ and multiple groups of data with distinct unknown correspondence.

Experiments on simulated data, the MovieLens 100K dataset and Yale B database show that $M^3O$ achieves state-of-the-art performance over several baselines and can recover the ground-truth correspondence with high accuracy. The code is provided in \url{https://github.com/TZW1998/MRUC}.

1 INTRODUCTION

In the era of big data, one usually needs to utilize data gathered from multiple disparate platforms when accomplishing a specific task. However, the correspondence among the data samples from these different sources are often unknown or noisy, due to either missing identity information or privacy reasons.\cite{Unnikrishnan2018, Gruteser2003, Das2018, Chan2018}. Examples include the record linkage problem\cite{Chan2001}, the federated recommender system\cite{Yang2020} and the vertical federated learning\cite{Nock2020}. Consider the simplest scenario, we have two data matrices $A = [a_1, ..., a_n]^\top$, $B = [b_1, ..., b_n]^\top$ with $a_i \in \mathbb{R}^{m_A}$ and $b_i \in \mathbb{R}^{m_B}$, which are from two different platforms (data sources). As discussed above, the correspondence $(a_i, b_i)$ may not be available, and thereby the goal is to recover the underlying correspondence between $a_1, ..., a_n$ and $b_{\pi(1)}, ..., b_{\pi(n)}$, where $\pi(\cdot)$ denotes an unknown permutation. We can translate such problem described above as a matrix recovery problem, i.e., to recover the matrix $M = [A, B]$ from the permutated observation $M_o = [A, \hat{P}B]$, where $\hat{P} \in \mathcal{P}_n$ is an unknown permutation matrix and $\mathcal{P}_n$ denotes the set of all $n \times n$ permutation matrices. We term this problem as Matrix Recovery with Unknown Correspondence (MRUC). Inspired by the classical low-rank model for matrix recovery\cite{Wright2010, Hastie2010, Ma2021} and the image recovery and alignment problem\cite{Zeng2012, Zhou2015}, we want to solve the following rank minimization problem for MRUC,

$$\min_{P \in \mathcal{P}_n} \text{rank}([A, PB_o]).$$  \hspace{1cm} (1)

Applications. The major application of MRUC problem is related to Vertical Federated Learning (VFL)\cite{Kairouz2011}, which aims at learning from feature partitioned data. This work specifically considers Recommender System (RS) in the context of VFL. One classical work on this problem is the multi-domain recommender system considered in\cite{Zhang2012}. Unfortunately, they neglect a crucial issue that data from these diverse platforms (or domains) are not always well aligned for two primary rea-
sons. The first is that the correspondence information could be noisy due to mishandle in data processing. The other is that those platforms may not be allowed to share the true linkage information for preserving privacy. As the first step to address these issues, in this work, we study RS in an extreme setting of VFL, i.e., no correspondence information is provided. Another application is the Visual Permutation Learning problem [Santa Cruz et al., 2017], where one needs to recover the original image from the shuffled pixels. Though less practical, this problem is still interesting to know under what structure in data one can guarantee a successful recovery. Both of the two applications give rise to a challenging extension of the MRUC problem, where we not only need to recover multiple correspondence across different data sources, but also face the difficulty of dealing with the missing values in data matrix.

Unlabeled Sensing. One similarly motivated problem is the Unlabeled Sensing (US) problem considered by [Unnikrishnan et al., 2018; Pananjady et al., 2017a; Tsakiris et al., 2020; Peng and Tsakiris, 2020; Tsakiris et al. and Peng, 2019; Slawski et al., 2021; Xie et al., 2021]. Especially, as discussed in Appendix ??, the MRUC problem is closely related to the Multivariate Unlabeled Sensing (MUS) problem, which has been studied in [Zhang et al., 2019a;b; Zhang and Li, 2020; Slawski et al., 2020b]. Specifically, the MUS is the multivariate linear regression problem with unknown correspondence, i.e., it solves

$$\min_{P \in P_n, W} \| Y - PXW \|_F^2, \quad (2)$$

where $W \in \mathbb{R}^{m_2 \times m_1}$ is the regression coefficient matrix, $Y \in \mathbb{R}^{n \times m_1}$ and $X \in \mathbb{R}^{n \times m_2}$ denotes the output and the permuted input respectively, and $\| \cdot \|_F$ is the matrix Frobenius norm. When $m_1 = 1$, the MUS problem reduces to an US problem. Despite of the similarity to the US problem, we remark that MRUC problem has its own distinct features and, as shown in Section 4, the algorithm for the MUS problem can not be directly and effectively applied, especially when there are multiple unknown correspondence and missing entries to be considered.

Related works. To the best of our knowledge, the concurrent and independent work [Yao et al., 2021] is the only work that also considers the MRUC problem. Theoretically, [Yao et al., 2021] showed that there exists a non-empty open subset $U \subseteq \mathbb{R}^{n \times (m_1 + m_2)}$ such that $\forall M \in U$, solving (1) is bound to recover the original correspondence. However, such results only prove its existence for the subset $U$ and do not provide a concrete characterization. Regarding the algorithm design, [Yao et al., 2021] first learn a robust subspace following the idea of [Slawski et al., 2020b], and then solves problem (1) heuristically as multiple independent US problems using algorithms from [Tsakiris et al., 2020; Peng and Tsakiris, 2020]. However, there are two main drawbacks in their algorithm that largely limit its practical value. First, as discussed in Appendix ?? and Remark 8, it ignores the interaction among the shuffled columns and hence can not recover the permutation correctly. Second, their method can not deal with data with missing values. Another recent paper [Nock et al., 2021] also shares a similar concern with ours on how correspondence information can affect VFL, though in a different context.

Contributions of this work. Our contributions in this work lie in both theoretical and practical aspects. Theoretically, we are the first to rigorously study how the rank of the data matrix is perturbed by the permutation, and show that problem (1) can be used to recover a generic low-rank random matrix almost surely. Besides, we propose a nuclear norm minimization problem as a surrogate for problem (1), and is also the first to study the property of nuclear norm under permutation. Practically, we propose an efficient algorithm M^O that solves the nuclear norm minimization problem, which overcomes the aforementioned two shortcomings in [Yao et al., 2021]. Notably, M^O works very well even for an extremely difficult task, where we need to recover multiple unknown correspondence from the data that are densely permuted and contain missing values. We remark that this is so far a challenging problem unexplored in the existing literature. Based on these findings, we also reach a novel and important observation for VFL: Even without any data linkage information, it is still possible for each participant/platform to benefit from VFL.

Outline. We start with building the theoretical understanding for the problem (1) and its convex relaxation in Section 2. Then, based on the theoretical intuition obtained from Section 2, we develop an efficient algorithm in Section 3 for most complicated scenario. The simulation results are presented in Section 4 and the conclusions are drawn in Section 5.

Notations. Given two matrices $X, Y \in \mathbb{R}^{n \times m}$, we denote $\langle X, Y \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}Y_{ij}$ as the matrix inner product. We denote $X(i)$ as the $i$th row of the matrix $X$ and $X(i, j)$ as the element at the $i$th row and the $j$th column. We denote $1_m \in \mathbb{R}^m$ and $1_{n \times m} \in \mathbb{R}^{n \times m}$ as the all-one vector and matrix, respectively, and $I_n$ be the $n \times n$ identity matrix. For $\alpha \in \mathbb{R}^m, \beta \in \mathbb{R}^n$, we define the operator $\oplus$ as $\alpha \oplus \beta = \alpha_1 + 1_m \beta^T \in \mathbb{R}^{m \times n}$. We denote $\| \cdot \|_*, \| \cdot \|_0$ as the nuclear norm for matrices. For vectors, we denote $\| \cdot \|_0, \| \cdot \|_1$ as the zero norm and 1-norm respectively.

2 MATRIX RECOVERY VIA A LOW-RANK MODEL

In this section, we study the role of low-rank model for recovering row permutation.

How is matrix rank perturbed by row permutation? To rigorously answer this question, we first introduce the notion cycle decomposition of a permutation.
Definition 2.1 (Cycle decomposition of a permutation [Dummit and Foote, 1991]). Let $S$ be a finite set, $\pi(\cdot)$ be a permutation on $S$. A cycle $(a_1, ..., a_n)$ is a permutation sending $a_j$ to $a_{j+1}$ for $1 \leq j \leq n-1$ and $a_n$ to $a_1$. Then a cycle decomposition of $\pi(\cdot)$ is an expression of $\pi(\cdot)$ as a union of several disjoint cycle.

It can be verified that any permutation on a finite set has a unique cycle decomposition [Dummit and Foote, 1991]. Therefore, we can define the cycle number of a permutation $\pi(\cdot)$ as the number of disjoint cycles with length greater than 1, which is denoted as $C(\pi)$. We also define the non-sparsity of a permutation as the Hamming distance between it and the original sequence, i.e., $H(\pi) = \sum_{s \in S} I[\pi(s) \neq s]$. It is obvious that $H(\pi) > C(\pi)$ if $\pi$ is not an identity permutation. As a simple example, we consider the permutation $\pi(\cdot)$ that maps the sequence $(1,2,3,4,5,6)$ to $(3,1,2,5,4,6)$. Now the cycle decomposition for it is $\pi(\cdot) = (132)(45)(6)$, and $C(\pi) = 2$, $H(\pi) = 5$.

We denote the original matrix as $M = [A, B] \in \mathbb{R}^{n \times m}$ with $A \in \mathbb{R}^{n \times m_A}$, $B \in \mathbb{R}^{n \times m_B}$, and $r = \text{rank}(M)$, $r_A = \text{rank}(A)$, $r_B = \text{rank}(B)$. We denote the corresponding permutation as $\pi_P(\cdot)$ for any permutation matrix $P \in \mathcal{P}_n$. The following proposition says that the perturbation effect of a permutation $\pi$ on the rank of $M$ could become stronger, if $\pi$ permutes more rows and contains less cycles.

**Proposition 2.2.** For all $P \in \mathcal{P}_n$, we have

$$\text{rank}([A, PB]) \leq \min\{n, m, r_A + r_B, r + H(\pi_P) - C(\pi_P)\}. \quad (3)$$

Similar result for the case with multiple permutations is summarized in Corollary ?? in Appendix ?? . It turns out that, without any further assumption on $M$, (3) is sharp and cannot be improved. Notably, the upper bound in (3) is attained with probability 1 for a generic low-rank random matrix.

**Definition 2.3.** A probability distribution on $\mathbb{R}$ is called a proper distribution if its density function $p(\cdot)$ is absolutely continuous with respect the Lebesgue measure on $\mathbb{R}$.

**Proposition 2.4.** If the original matrix $M$ is a random matrix with $M = RE$ where $R \in \mathbb{R}^{n \times r}$ and $E \in \mathbb{R}^{r \times m}$ are two random matrices whose entries are i.i.d and follow a proper distribution on $\mathbb{R}$, and $r \leq \min\{\sqrt{n}, m_A, m_B\}$, then $\forall P \in \mathcal{P}_n$, the equality below holds with probability 1.

$$\text{rank}([A, PB]) = \min\{2r, r + H(\pi_P) - C(\pi_P)\} \quad (4)$$

**Discussion on Proposition 2.4.** It is worthwhile to mention that our Proposition 2.4 strengthens the Theorem 1 in [Yao et al., 2021] to some extent. Specifically, [Yao et al., 2021] shows that, with probability 1, the rank of the perturbed matrix will never be lower than that of the original matrix. Compared to them, our result precisely predicts how much the rank will increase after row perturbation. Besides, Proposition 2.4 is especially favorable from the optimization perspective, as now the rank is a monotone function w.r.t the degree of perturbation.

**Convex relaxation for the rank function.** Despite the previous theoretical justification for problem (1), it is non-convex and non-smooth. Another crucial issue is that we often have a noisy observation matrix and it is well known that the rank function is extremely sensitive to the additive noise. In this paper, we assume that the observation matrix is corrupted by i.i.d Gaussian additive noise, i.e.,

$$M_o = [A_o, B_o] = [A, \tilde{P}B] + W, W(i,j) \sim \mathcal{N}(0, \sigma^2),$$

where $\sigma^2$ denotes the variance of the noise. We denote the singular values of a matrix $X \in \mathbb{R}^{n \times m}$ as $\sigma_1^X, ..., \sigma_k^X$ where $k = \min\{n, m\}$. Since $\text{rank}(X) = \|\sigma_1^X, ..., \sigma_k^X\|_0$, from Proposition 2.4 we can view the perturbation effect of a permutation to a low-rank matrix as breaking the sparsity of its singular values, which leads naturally to the nuclear norm minimization problem that has been shown to be robust to additive noise and favor low-rank solution [Wright and Ma, 2021], i.e.,

$$\min_{P \in \mathcal{P}_n} \|A_o, PB_o\|_* = \|\sigma_{M_o}^1, ..., \sigma_{M_o}^k\|_1. \quad (5)$$

**Theoretical justification for the nuclear norm.** Nuclear norm has a long history being used as a convex surrogate for the rank, and it has been theoretically justified for applications like low-rank matrix completion [Candes and Tao, 2010] [Wright and Ma, 2021]. It is also important to see whether the nuclear norm is still a good surrogate for the rank minimization problem (1). In this work, we establish a sufficient condition on $A$ and $B$ under which problem (5) is provably justified for correspondence recovery. We denote $A = \sum_{i=1}^{r_A} \sigma_i^A u_i^A v_i^A^T$, $B = \sum_{i=1}^{r_B} \sigma_i^B u_i^B v_i^B^T$ as the singular values decomposition of $A$ and $B$, where the $\sigma_i^A$ and $\sigma_j^B$ are the non-zero singular values. To derive the worst-case error bound of nuclear norm minimization, we propose the following assumption on $M$.

**Assumption 2.5.** There exists a constant $\epsilon_1 \geq 0$, $\epsilon_2 \geq 0$, $\epsilon_3 \geq 0$ such that

$$|\sigma_i^A - \sigma_j^B| \leq \epsilon_1, \forall i = 1, ..., r, \quad (6)$$

$$\|u_i^A - u_i^B\| \leq \epsilon_2, \forall i = 1, ..., T, \quad (7)$$

$$\min_{u \in U} \min_{i \neq j} |u(i) - u(j)| \geq \epsilon_3 > 0, \quad (8)$$

where we denote $\sigma_i^A = 0$ if $i > r_A$, and similarly for $\sigma_i^B$, $T = \min\{r_A, r_B\}$ and $U = \{u_1^A, ..., u_T^A, u_1^B, ..., u_T^B\}$.

Here we provide some intuition behind these assumptions. Firstly, from the definition of nuclear norm, it can be simply
With these assumptions, we derive the following result,

\[ -Z/N \leq (\| [A, PB] \|_* - \| M \|_*) / \| M \|_* \leq Z/N, \]  

(9)

where \( N = \max\{\| A \|_*, \| B \|_*\} \) and \( Z = \min\{\| A \|_*, \| B \|_*\} \). The inequality (9) indicates that \( A \) and \( B \) should have comparable magnitude, i.e., \( \| A \|_* \approx \| B \|_* \), otherwise the influence of the permutation will be less significant. With this observation, as depicted by [5], we assume that the singular values of \( A \) and \( B \) are comparable. As for (7), we propose it with an aim to capture the intuition that if \( A \) and \( B \) are data from the same group of users, the distance (in SVD sense) between \( A \) and \( B \) should be close, i.e., the matrix \([A, B]\) should be "low-rank". We would like to interpret the constants \( \epsilon_2 \) as a continuous measure for the low-rankness of a matrix, because it indicates that the column space of \( M \) can be approximated by the column space of one of its submatrices. Lastly, it is easy to verify that if there is a \( P \in P_n \) such that \( u_i^P = Pu_i^B \) for all \( i \), then \([A, PB] = [A, B]\). Therefore, we propose (8) to avoid this case.

**Remark 1.** Though these assumptions could be refined, we remark that they are almost sharp. In Appendix ??, we construct a few concrete counterexamples which do not satisfy these assumptions and are impossible to be recovered within meaningful accuracy by nuclear norm minimization problem.

With these assumptions, we derive the following result, which provides high probability bound for the approximation error of [5]. We denote the solution to (5) as \( \hat{\pi} \), and let \( \pi^* \) and \( \hat{\pi} \) be the corresponding permutation to the permutation matrices \( P^* \) and \( \hat{P} \), respectively. We define the difference between the two permutations \( \pi^* \) and \( \hat{\pi} \) as the Hamming distance

\[ d_H(\pi^*, \hat{\pi}) \overset{\text{def}}{=} \sum_{i=1}^{n} I(\pi^*(i) \neq \hat{\pi}(i)). \]

**Proposition 2.6.** Under Assumptions 2.5 if additionally \( \epsilon_1 \leq \frac{D}{2\sqrt{r}},\) \( \epsilon_2 \leq \min\{\frac{1}{3\sqrt{r}} \sqrt{\frac{2L}{N}}, \frac{\sqrt{D}}{2N}\} \), and \( \sigma \leq \frac{D}{\sqrt{2r} \epsilon_2} \), then the following bound

\[ d_H(\pi^*, \hat{\pi}) \leq \frac{2}{c_3} \left( 2 - \left( \sqrt{2D/(D + (\sqrt{2} + 2)\epsilon_1 r + \sqrt{2\epsilon_2 N + 2\sqrt{2DL}\sigma})} - \sqrt{2}\epsilon_2 \right)^2 \right) \]

(10)

holds with probability at least \( 1 - 2 \exp\{-\frac{D}{8\epsilon_2}\} \), where \( L = \max\{n, m\}, D = \| A \|_* + \| B \|_* \).

The proof to all the aforementioned theoretical results are provided in Appendix ??.

**Remark 2.** From Proposition 2.6 we can see that when \( \epsilon_3 > 0 \), and \( \epsilon_1 \rightarrow 0,\) \( \epsilon_2 \rightarrow 0,\) \( \sigma \rightarrow 0 \), the error \( d_H(\pi^*, \hat{\pi}) \) will converge to zero with probability 1. We can also discover that the correspondence can be difficult to recover when: The rank of original matrix \( M \) is high; The magnitude of \( A \) and \( B \) w.r.t rank or nuclear norm are not comparable; The strength of noise is high. Notably, the numerical experiments in Section 4.1 corroborate these findings as well. Due to page limit, we refer detailed discussion and analysis on Proposition 2.6 to Appendix ??.

**Remark 3.** In many applications, we can only observe part of the full data. Therefore, it is worthwhile to investigate whether nuclear norm minimization could work when we can only access a small subset of the entries in \( M_o \). Notably, Figure 1 empirically gives the positive answer and shows that the "monotone relationship of nuclear norm w.r.t numbers of permuted rows" is gracefully degraded when the percentage of observable entries is decreasing. This phenomenon is remarkable since it indicates the original correspondence can be recovered from only part of the full data. The matrices used to generate Figure 1 are the same as those in Section 4.1 and the nuclear norm is computed approximately by first filling the missing entries using SoftImpute algorithm [Mazumder et al., 2010].

### 3 ALGORITHM

In this section, we develop an algorithm for MRUC based on the intuition obtained from Section ?? Moreover, we require that the algorithm can deal with the scenario with missing values, i.e., our observed data is \( P_{o\Omega}(M_o) = P_{\Omega}(M_o) \), where \( P_{\Omega} \) is an operator that selects entries which are in the set of observable indices \( \Omega \). In this scenario, problem [5] can not be directly used since the evaluation of the nuclear norm and optimization of the permutation are coupled together. Inspired by the matrix completion method [Hastie et al., 2015] [Mazumder et al., 2010], we propose to solve an
alternative form of (5) as follows,
\[
\min_{\hat{M} \in \mathbb{R}^{n \times m}} \min_{P \in \mathcal{P}_n} \left\| P_{\Omega}([A, PB_0]) - P_{\Omega}(\hat{M}) \right\|^2_F + \lambda \left\| \hat{M} \right\|_*,
\]
where \( \lambda > 0 \) is the penalty coefficient. We denote that \( \hat{M} = [\hat{M}_A, \hat{M}_B] \) and \( \hat{M}_A, \hat{M}_B \) are the two submatrices with the same dimension as \( A_o \) and \( B_o \) respectively. We can write (11) equivalently as
\[
\min_{\hat{M} \in \mathbb{R}^{n \times m}} \min_{P \in \mathcal{P}_n} \left\| P_{\Omega}(A_o) - P_{\Omega}(\hat{M}_A) \right\|^2_F
+ \langle C(\hat{M}_B), P \rangle + \lambda \left\| \hat{M} \right\|_*,
\]
where \( C(\hat{M}_B) \in \mathbb{R}^{n \times n} \) is the pairing cost matrix with
\[
C(\hat{M}_B)(i, j) = \sum_{(j, j') \in \Omega} \left( \hat{M}_B(i, j') - B_o(j, j') \right)^2,
\]
for \( \forall i, j = 1, ..., n \).

Baseline algorithm. A conventional strategy to handle an optimization problem like (12) is the alternating minimization or the block coordinate descent algorithm [Abid et al., 2017]. Specifically, it executes the following two updates iteratively until it converges.
\[
\hat{M}^{\text{new}} \leftarrow \arg \min_{\hat{M} \in \mathbb{R}^{n \times m}} \left\| P_{\Omega}(A_o, \hat{P}^{\text{old}} B_o) - P_{\Omega}(\hat{M}) \right\|^2_F
+ \lambda \left\| \hat{M} \right\|_*,
\]
\[
\hat{P}^{\text{new}} \leftarrow \arg \min_{P \in \mathcal{P}_n} \langle C(\hat{M}_B^{\text{new}}), P \rangle.
\]
The first update step (13) is a convex optimization problem and can be solved by the proximal gradient algorithm [Mazumder et al., 2010]. The second update step (14) is actually a discrete optimal transport problem which can be solved by the classical Hungarian algorithm with time complexity \( O(n^3) \) [Jonker and Volgenant, 1986]. However, as we will see in the Section 4 this algorithm performs poorly, and it is likely to fall into an undesirable local solution quickly in practice. Specifically, the main reason is that the solution of (14) is often not unique and a small change in \( \hat{M}_B \) would lead to large change of \( \hat{P} \). To address this issue, we propose a novel and efficient algorithm \( \text{M}^3 \text{O} \) algorithm based on the entropic optimal transport [Peyré et al., 2019] and min-max optimization [Jin et al., 2020a].

Smoothing the permutation with entropy regularization. For any \( a \in \mathbb{R}^n \), \( b \in \mathbb{R}^m \), we define
\[
\Pi(a, b) = \{ S \in \mathbb{R}^{n \times m} : S1_m = a, S^T1_n = b, S(i, j) \geq 0 \quad \forall i, j \},
\]
which is also known as the Birkhoff polytope. The famous Birkhoff-von Neumann theorem [Birkhoff, 1946] states that the set of extremal points of \( \Pi(1_n, 1_n) \) is equal to \( \mathcal{P}_n \). Inspired by [Xie et al., 2021] and the interior point method for linear programming [Bertsekas, 1997], in order to smooth the optimization process of the baseline algorithm, we relax \( P \) from being an exact permutation matrix, i.e., to keep \( P \) staying inside the Birkhoff polytope \( \Pi(1_n, 1_n) \). That is, we propose to replace the combinatorial problem (14) with the following continuous optimization problem
\[
\min_{P \in \Pi(1_n, 1_n)} \langle C(\hat{M}_B), P \rangle + \epsilon \mathcal{H}(P),
\]
where \( \mathcal{H}(P) \defeq \sum_{i,j} P(i, j)(\log(P(i, j)) - 1) \) is the matrix negative entropy and \( \epsilon > 0 \) is the regularization coefficient. Notably, (15) is also known as the Entropic Optimal Transport (EOT) problem [Peyré et al., 2019], which is a strongly convex optimization problem and can be solved roughly in the \( O(n^2) \) complexity per iteration by the Sinkhorn algorithm. Specifically, the Sinkhorn algorithm solves the dual problem of (15).

\[
\max_{\alpha, \beta \in \mathbb{R}^n} W_\epsilon(\hat{M}_B, \alpha, \beta) \defeq \langle 1_n, \alpha \rangle + \langle 1_n, \beta \rangle - \epsilon \left( 1_{n \times n}, \exp \left( \frac{\alpha + \beta - C(\hat{M}_B)}{\epsilon} \right) \right).
\]
which reduces the variables dimension from \( n^2 \) to \( 2n \) and is thus greatly favorable in the high dimension scenario. By substituting the inner minimization problem of (14) with (15), we end up with solving the following unconstrained min-max optimization problem
\[
\min_{\hat{M}} \max_{\alpha, \beta \in \mathbb{R}^n} \left\| A - \hat{M}_A \right\|^2_F + W_\epsilon(\hat{M}_B, \alpha, \beta) + \lambda \left\| \hat{M} \right\|_*
\]
Follows the idea of [Jin et al., 2020a], we consider to adopt a proximal gradient algorithm with a Max-Oracle for (17). Specifically, we employ the Sinkhorn algorithm [Peyré et al., 2019] as the Max-Oracle to retrieve an \( \epsilon \)-good Nash equilibrium in Algorithm 1 where \( \text{prox}_{\lambda \| \cdot \|_*}(\cdot) \) is the proximal operator of nuclear norm, \( \rho_k \) is the gradient stepsize and
\[
F_\epsilon(\hat{M}_A, \alpha, \beta) \defeq \left\| A - \hat{M}_A \right\|^2_F + W_\epsilon(\hat{M}_B, \alpha, \beta).
\]
The convergence property of \( \text{M}^3 \text{O} \) can be obtained by following [Jin et al., 2020a], which shows that, with a decaying stepsize, \( \text{M}^3 \text{O} \) is bound to converge to an \( \epsilon \)-good Nash equilibrium within \( O(\epsilon^{-2}) \) iterations.

Remark 4. A recent work [Xie et al., 2020] proposes a decaying strategy for the entropy regularization coefficient \( \epsilon \) in (15) so that the optimal solutions of (14) and (15) do
\textbf{Algorithm 1} \textit{M}^3\textit{O} (Simplified)

\textbf{Input}: tolerance $\varepsilon$, observation $M_o$, initialization $\hat{M}$.

\textbf{repeat}

Run the Sinkhorn algorithm to find $\alpha^*, \beta^*$ such that

$$W_e(\hat{M}_B^k, \alpha^*, \beta^*) > \max_{\alpha, \beta} W_e(\hat{M}_B^k, \alpha, \beta) - \varepsilon;$$

$$\hat{M}^{k+1} \leftarrow \text{prox}_{\lambda\|\cdot\|_1}(\hat{M}^k - \rho_k \nabla M F_e(\hat{M}^k, \alpha^*, \beta^*)).$$

\textbf{until} converged

not deviate too much. Inspired by it, in our practice, we take large $\varepsilon$ in the beginning and gradually shrink it by half whenever the objective value stops improving for $K$ steps.

\textbf{Remark 5}. A useful trick is that we should not take large stepsize $\rho_k$ in the early iterations because the permutation matrix could still be far away from the optimal one. However, a small stepsize would slow to convergence. Heuristically, we propose an adaptive stepsize strategy that decreases $\rho_k$ in the beginning and gradually shrink it by half.

Remark 6. As discussed in Section 1, in many cases we have to deal with the problem that involves multiple correspondence, i.e., we need to recover the matrix $M = [A, B_1, \ldots, B_d]$ from the observation data $\tilde{P}_\Omega(M_o)$, where

$$M_o = [A_o, B_1^d, \ldots, B_d^d] = [A, \tilde{P}_1 B_1, \ldots, \tilde{P}_d B_d] + W;$$

where $\tilde{P}_l \in \mathcal{P}_n$ and $W$ is a noise matrix. We refer such problem as the \textit{d}-correspondence problem. An important observation is that, although the number of possible correspondence increase exponentially as $d$ grows, the complexity of $\text{M}^3\text{O}$ per iteration only linearly increases with $d$ and can be implemented in a fully parallel fashion. Specifically, in this scenario, we solve the problem

$$\min_{\tilde{M}} \min_{P_1, \ldots, P_d} \|P_\Omega(A_o) - P_\Omega(\tilde{M}_A)\|_F^2 + \sum_{l=1}^{d} \{C(\tilde{M}_{B_l}), P_l\}$$

$$+ \epsilon \mathcal{H}(P_l) + \lambda \|\tilde{M}\|_*$

\text{s.t.} $P_l \in \Pi(1_n, 1_n)$, $l = 1, \ldots, d$,

where we denote $\tilde{M} = [\tilde{M}_A, \tilde{M}_{B_1}, \ldots, \tilde{M}_{B_d}]$. Here $\tilde{M}_A$ and $\tilde{M}_{B_l}$ have the same dimension with $A_o$ and $B_l^d$, respectively. One can find that the inner problems for solving $P_l$ are actually decoupled for each $l$, which guarantees an efficient parallel implementation.

\textbf{Remark 7}. Since problem (11) has a similar form to that considered in [Mazumder et al., 2010]. We adopt the same tuning strategy of $\lambda$ as in [Mazumder et al., 2010], which suggests that we should start with large $\lambda$ and gradually decrease it.

We relegate more details about $\text{M}^3\text{O}$ to Appendix ??.

\figure{2}{Performance of various algorithms on a simulated 1-correspondence problem.}

4 EXPERIMENTS

In this section, we evaluate our proposed $\text{M}^3\text{O}$ on both synthetic and real-world datasets, including the MovieLens
100K and the Extended Yale B dataset. We also provide an ablation study for the decaying entropy regularization strategy and the adaptive stepsize strategy proposed in Remarks 4 and 5. In all the experiments, we employ the Soft-Impute algorithm [Mazumder et al., 2010] as a standard algorithm for matrix completion. Extra experiment details and auxiliary results can be found in Appendix ??.

Algorithms. We denote the following algorithms for comparison in all the experiments:

1. Oracle: Running the Soft-Impute algorithm with ground-truth correspondence.
2. Baseline: The Baseline algorithm in (13) and (14).
3. MUS: Since there is currently no existing algorithm directly applicable to the scenario considered by (18), we modify and extend the algorithm in [Zhang and Li, 2020], which is originally proposed for the MUS problem, to deal with the MRUC problem. The details of the adapted algorithm are provided in Appendix ??.

Remark 8. As discussed in [Pananjady et al., 2017a], leveraging the prior knowledge that multiple columns are shuffled by the same permutation is generally helpful for permutation recovery. This is why we only adopt the MUS algorithm in [Zhang and Li, 2020] instead of those US algorithms considered by [Yao et al., 2021] for comparison. For a more serious and experimental discussion, we refer readers to Appendix ??.

4.1 SYNTHETIC DATA

We first investigate the property of our proposed M³O algorithm on the synthetic data.

Data generation. We generate the original data matrix in this form \( M = RE + \eta W \), where \( R \in \mathbb{R}^{n \times r}, E \in \mathbb{R}^{r \times k}, W \in \mathbb{R}^{r \times m}, \) \( n \) and \( \eta > 0 \) indicates the strength of the additive noise. The entries of \( R, E, W \) are all i.i.d sampled from the \( N(0, 1) \). Then we split the data matrix \( M \) by \( M = [A, B_1, ..., B_d] \) where we denote \( A \in \mathbb{R}^{n \times m_A}, B_1 \in \mathbb{R}^{n \times m_1}, ..., B_d \in \mathbb{R}^{n \times m_d} \) to represent data from \( d + 1 \) data sources. The permuted observation matrix \( M_o \) is obtained by first generating \( d \) permutation matrices \( P_1, ..., P_d \) randomly and independently, and then computing \( M_o = [A, P_1 B_1, ..., P_d B_d] \). Finally, we remove \( (1 - |\Omega| \cdot 100\%/(n \cdot m)) \) percent of the entries of \( M_o \) randomly and uniformly, where \( |\Omega| \) indicates the number of observable entries.

Ablation study. We denote the following variants of M³O for the ablation study.

1. M³O-AS-DE: M³O with both Adpative Stepsize and Decaying Entropy regularization.
2. M³O-DE: M³O with Decaying Entropy regularization only. M³O-DE-1 and M³O-DE-2 adopt constant stepsize \( \rho_k = 0.5 \) and \( \rho_k = 0.01 \), respectively.
3. M³O-AS: M³O with Adpative Stepsize only. The entropy coefficient \( \epsilon \) is fixed to 0.0005.

In the following results, we denote \( \pi_l \) as the corresponding permutation to \( P_l \). We initialize \( M \) from Gaussian distribution for the M³O algorithm and its variants. We choose initial \( \epsilon \) as 0.1 and \( K = 100 \) as the default for the decaying entropy regularization, and set \( \omega = 3 \) as the default for the adaptive stepsize. We also report the achieved objective values of (18) for the tested algorithms, except for the MUS algorithm since it has a different objective. We denote \( \hat{\pi} \) as the recovered permutation.

Results. Figure 2 displays the result under the setting \( \eta = 0.1, |\Omega| \cdot 100\%/(n \cdot m) = 80\% \), \( n = m = 100, r = 5, d = 1, m_A = 60 \) and \( m_1 = 40 \). The algorithm M³O-AS-DE achieves the best result, and can recover the ground-truth correspondence. M³O-AS behaves similarly to Baseline and MUS. They all converge to a poor local solution quickly. M³O-DE-1 converges quickly and also falls into a poor local solution due to large stepsize, while M³O-DE-2 adopts a small stepsize and hence suffers from slow convergence. Due to the superiority of M³O-AS-DE over the other variants, in the following results, we refer M³O as M³O-AS-DE for short.

<table>
<thead>
<tr>
<th>( \Omega ) ( 100%/(n \cdot m) )</th>
<th>( \hat{\pi} ) ( d )</th>
<th>( \sum_{i=1}^{d} d_H(\hat{\pi}_l, \pi_l) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(100,40,30,30)</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>(100,20,40,40)</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>(100,45,25,25,25)</td>
<td>3</td>
<td>50</td>
</tr>
<tr>
<td>(100,40,25,25,25)</td>
<td>4</td>
<td>60</td>
</tr>
</tbody>
</table>

Figure 3 examine M³O on a 1-correspondence problem under different regimes w.r.t \( |\Omega|, \eta, r \) and \( m_A/n \). Here we use \( m_A/n \) to control the difference of the magnitude of the submatrices. As we can see, the results are well aligned with our prediction in Remarks 2 and 3. We also find that the performance of M³O tends to have high variance. This is mainly because M³O is sensitive to random initialization, and more details on this phenomenon are in Appendix ??.

In practice, we recommend to run M³O a few times with different random initializations.

Finally, we examine M³O on a few d-correspondence problems. See Table I for various results, where we set \( r = 5 \) and \( \varepsilon = 0.1 \). Notice that for the 4-correspondence problem in the table, there are \( (100!)^4 \) possible correspondence. Even for such a difficult problem, M³O is able to recover 61.5% of the ground-truth correspondence with a good initialization.
4.2 MULTI-DOMAIN RECOMMENDER SYSTEM WITHOUT CORRESPONDENCE

In this section, we study the performance of M$^3$O on a real world dataset MovieLens 100K\textsuperscript{2}, which is a widely used movie recommendation dataset [Harper and Konstan, 2015]. In this application, we mainly focus on the metric Root Mean Squared Error (RMSE), i.e.,

\[
\text{RMSE} \equiv \sqrt{\frac{1}{N} \sum_{i,j} (\hat{M}_{ij} - M_{ij})^2}.
\]

Data. MovieLens 100K contains 100,000 ratings within the scale 1-5. The ratings are given by 943 users on 1,682 movies. Genre information about movies is also provided. We adopt a similar setting with [Zhang et al., 2012]. We extract five most popular genres, which are Comedy (C), Romance (R), Drama (D), Action (A), Thriller (T) respectively, to define the data from 5 different domains (or platforms). In addition to [Zhang et al., 2012], we randomly permute the indexes of the users from these five domains respectively, so that the correspondence among these data become unknown. In this way, the problem belongs to the 4-correspondence problem as discussed before. The ratings are split randomly, with 80% of them as the training data and the other 20% of them as the test data.

Algorithms. We consider the following additional algorithms for comparison.

1. SIC: Running the Soft-Impute algorithm independently for the 5 different platforms.
2. SIR: Running the Soft-Impute algorithm with Randomly generated correspondence.

Results. As discussed in experiments on the simulated data, the exact recovery of correspondence becomes impossible due to the small amount of observable entries. Therefore, in the following experiment, since exact correspondence is not needed, we fix $\epsilon = 0.05$ for M$^3$O. Table 2 shows the results by averaging the RMSE on the test data over 10 different random seeds. We can first see that the matrix completion with a wrong correspondence, i.e., SIR, can be harmful to the overall performance since it is even worse than the results of SIC. Notably, although the ground-truth correspondence can not be recovered, each platform can still benefit from M$^3$O since it improves the performance over SIC. This is mainly because M$^3$O is still able to correspond similar users for inferring missing ratings. On the contrary, since both Baseline and MUS can only establish an exact one-to-one correspondence for each user, they fail to improve SIC significantly. Remarkably, M$^3$O is only inferior to the Oracle method a little, and even achieves lower test RMSE than the Oracle method on the Comedy genre.

\textsuperscript{2}https://grouplens.org/datasets/movielens/100k/
Table 2: Test RMSE of various algorithms on MovieLens 100K

<table>
<thead>
<tr>
<th>Method</th>
<th>C</th>
<th>R</th>
<th>D</th>
<th>A</th>
<th>T</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIR</td>
<td>1.020</td>
<td>1.016</td>
<td>0.981</td>
<td>0.980</td>
<td>0.981</td>
<td>0.994</td>
</tr>
<tr>
<td>SIC</td>
<td>0.969</td>
<td>0.970</td>
<td>0.932</td>
<td>0.918</td>
<td>0.925</td>
<td>0.942</td>
</tr>
<tr>
<td>MUS</td>
<td>0.966</td>
<td>0.984</td>
<td>0.942</td>
<td>0.931</td>
<td>0.931</td>
<td>0.949</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.973</td>
<td>0.956</td>
<td>0.938</td>
<td>0.911</td>
<td>0.915</td>
<td>0.940</td>
</tr>
<tr>
<td>M3O</td>
<td><strong>0.939</strong></td>
<td><strong>0.879</strong></td>
<td><strong>0.914</strong></td>
<td><strong>0.856</strong></td>
<td><strong>0.857</strong></td>
<td><strong>0.895</strong></td>
</tr>
<tr>
<td>Oracle</td>
<td>0.944</td>
<td>0.783</td>
<td>0.906</td>
<td>0.818</td>
<td>0.810</td>
<td>0.867</td>
</tr>
</tbody>
</table>

### 4.3 VISUAL PERMUTATION RECOVERY

We also show that M3O is flexible and can also be applied to a visual jigsaw puzzle. This kind of problem is recently considered in [Santa Cruz et al., 2017], which proposes to recover the corrupted image in a data-driven way using convolutional neural networks. However, we show that it is possible to recover the image without extra data by merely exploiting the underlying low-rank structure of the image itself. A typical result is shown in Figure 4. The experiment details and more results are provided in Appendix ??.

### 5 CONCLUSION

In this paper, we study the important MRUC problem where part of the observed submatrix is shuffled. Such problem underlies the record linkage problem in VFL [Nock et al., 2021]. This problem has not been well explored in the existing literature. Theoretically, we are the first to rigorously analyze the role of low-rank model in the MRUC problem, and also provide an almost sharp sufficient condition under which minimizing nuclear norm is provably efficient for recovering permutation. For practical implementations, we propose an efficient algorithm, the M3O algorithm, which consistently achieves the best performance over several baselines in all the tested scenarios. For future works, it is important to extend the theoretical results to the scenario with missing values, and hopefully derive a theorem that can rigorously quantify the remarkable phenomenon exhibited in Figure 1.

### References


Babak Barazandeh and Meisam Razaviyayn. Solving Non-Convex Non-Differentiable Min-Max Games Using Prox-


