A Sparse Representation-Based Approach to Linear Regression with Partially Shuffled Labels

Martin Slawski, Mostafa Rahmani, Ping Li
Cognitive Computing Lab, Baidu Research USA
10900 NE 8th ST. Bellevue, WA 98004, USA
martin.dot.slawski@gmail.com, mostafarahmani@baidu.com, liping11@baidu.com

Abstract

Several recent papers have discussed a modification of linear regression in which the correspondence between input variables and labels is missing or erroneous, referred to as "Linear Regression with Unknown Permutation", or "Linear Regression with Shuffled Data". Prior studies of this setup have shed light on the associated statistical limits. However, practical and well-founded approaches that overcome the computational challenges of the setup are still scarce. In this paper, we translate the problem of linear regression with unknown permutation to a robust subspace recovery problem, or alternatively, outlier recovery problem. Outliers correspond to mismatched data, and a specific sparse representation of the data is used to detect the outliers. The proposed approach requires at least a reasonably large fraction (but potentially significantly less than 50%) of the response-predictor pairs to be in correct correspondence. This assumption is often justified, e.g., if record linkage based on additional contextual information is sufficiently accurate to enable correct matching of such fraction of the data. It turns out that in this situation, estimation of the regression parameter on the one hand and recovery of the underlying permutation on the other hand can be decoupled so that the computational hardness associated with the latter can be sidestepped.

1 INTRODUCTION

Suppose we are given two samples $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \subseteq \mathbb{R}^m$ that contain two different pieces of information about a common set of observational units (entities), but there is only incomplete knowledge about which datum in $\mathcal{Y}$ belongs to the same entity as another datum in $\mathcal{X}$. As an illustrative example, $\mathcal{X}$ and $\mathcal{Y}$ might correspond to measurements from two different sensors, but each measurement is observed without or only inaccurate time stamp. This is a common scenario in engineering applications [4, 18].

In statistics, this situation has been studied under the term "Broken Sample Problem" [3, 10, 14, 15, 16, 20, 48]. Specifically, $\{(\mathbf{x}_{\pi^*(i)}, \mathbf{y}_i)\}_{i=1}^n$ for some unknown permutation π^* are assumed to be i.i.d. pairs from a joint distribution P^*, typically belonging to some parametric family $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, and it is of interest to recover π^* as well as the parameter θ^* of $P^* = P_{\theta^*}$. Perhaps the most straightforward instance of this setup is the bivariate normal case ($d = m = 1$) with θ^* as the correlation coefficient. "Broken Samples" naturally arise in record linkage [11, 25]. Broadly speaking, the latter term refers to the task of combining data sets obtained from multiple sources into a single, more comprehensive data set by matching records corresponding to the same entity. This process can be challenging and error-prone in the absence of accurate (quasi-) identifying variables. On the other hand, an adversary may try to use record linkage in order to identify sensitive information. A well-known example is due to Sweeney (1997) who demonstrated that voter registration data can be used to disclose individual health records even if the latter do not contain explicit identifiers. In this regard, broken samples also pertain to the area of data confidentiality.

Model and Problem Statement.
The task of re-pairing a broken sample, i.e., recovering the permutation π^* from $\mathcal{X}$ and $\mathcal{Y}$ appears hopeless unless there is a strong association between the underlying random variables. In a motivating example in [16], $\mathcal{X}$ and $\mathcal{Y}$ consist of photos of n movie stars taken during adulthood and childhood, respectively. It is expected that by extracting suitable features from both sets of photos, matching of corresponding photos is facilitated. Several works have studied the case of a monotone functional re-

lationship between both samples [9, 19, 35]. Monotonicity is a natural assumption as it prompts a straightforward way of estimating π^* via sorting. Various recent papers have considered a linear functional relationship under the terms "Unlabeled Sensing", "Linear Regression with Unknown Permutation", or "Linear Regression with Shuffled Data" [1, 18, 24, 26, 33, 34, 39, 45]. Specifically, the model considered therein is of the form

$$\mathbf{Y} = \mathbf{\Pi}^* \mathbf{X} \mathbf{B}^* + \sigma \mathbf{E}, \tag{1}$$

where $\mathbf{Y}$ has rows $\{\mathbf{y}_i^\top\}_{i=1}^n$, $\mathbf{X}$ has rows $\{\mathbf{x}_i^\top\}_{i=1}^n$, $\mathbf{B}^* \in \mathbb{R}^{d\times m}$ is a regression parameter, $\mathbf{\Pi}^*$ is a permutation matrix, and $\mathbf{E}$ represents random additive error scaled by $\sigma > 0$. The goal is to recover $\mathbf{B}^*$ and/or $\mathbf{\Pi}^*$ given $(\mathbf{X}, \mathbf{Y})$. Applications of (1) include the reconstruction of spatial fields using mobile sensors [44], time-domain sampling in the presence of clock jitter [4], multi-target tracking in radar [6], header-free communication in sensor networks [33], regression analysis after record linkage [13, 23, 27, 32, 36, 37], linkage of electronic health records [38], correspondence problems in computer vision, and gated flow cytometry [1].

Main Challenges.
The presence of the permutation poses unprecedented statistical and computational challenges. The paper [45] shows that in case that $\mathbf{E} = \mathbf{0}$, $\mathbf{B}^*$ can be uniquely recovered by exhaustive search over permutations almost surely if the entries of $\mathbf{X}$ are i.i.d. from a continuous distribution and $n > 2d$. Multiple subsequent papers consider the situation in which $\mathbf{X}$ and $\mathbf{E}$ have i.i.d. standard Gaussian entries. In this setting, a series of properties have been established for the least squares problem

$$\min_{\mathbf{\Pi}\in\mathcal{P}_n,\, \mathbf{B}\in\mathbb{R}^{d\times m}} \|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}\|_F^2, \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\mathcal{P}_n$ denotes the set of n-by-n permutation matrices. Problem (2) is a specific quadratic assignment problem [7]. A result in [34] shows that (2) is NP-hard. For $m = 1$ (i.e., the response variable is scalar), the paper [34] also derives some necessary and some sufficient conditions for exact and approximate recovery of $\mathbf{\Pi}^*$ based on (2), and elaborates on the significance of the quantity $\mathtt{SNR} = \|\mathbf{B}^*\|_F^2/\sigma^2$ (signal-to-noise ratio) in this context. The requirement $\mathtt{SNR} = \Omega(n^2)$, i.e., an excessively large signal-to-noise ratio, is proved to be a necessary condition for approximate permutation recovery with respect to the Hamming distance. In a similar spirit, the work [26] shows that for $m = 1$, $\mathtt{SNR} = \Omega(d/\log\log n)$ is necessary for approximate recovery of $\mathbf{B}^*$. Tractable algorithms with provable guarantees are scarce at this point: the scheme in [26] has polynomial time complexity, but is "not meant for practical deployment" as the authors state themselves. The convex relaxation of (2) in which $\mathcal{P}_n$ is replaced by its convex hull, the set of doubly stochastic matrices, was shown to perform poorly [18]. The papers [33, 51] study (2) for multivariate response ($m > 1$). While [33] contains results on the denoising error rather than on recovery of $\mathbf{B}^*$ or $\mathbf{\Pi}^*$ as in [51], both papers provide substantial support for the hypothesis that multiple responses resulting from the same permutation reduce the required $\mathtt{SNR}$ in order to reliably estimate those two quantities.

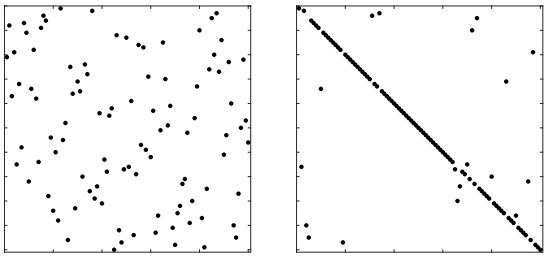


Figure 1: Left: Permutation matrix without additional structure. Right: Sparse Permutation.

Proposed Approach and Contributions.
Despite the significant body of work on setup (1), approaches that are both computationally feasible and equipped with statistical guarantees are scarce. One possible remedy is to impose additional assumptions on the permutation $\mathbf{\Pi}^*$. We herein consider the case in which $\mathbf{\Pi}^*$ moves only a fraction $\alpha = k/n$ of indices, or equivalently, the correspondence between $\mathcal{X}$ and $\mathcal{Y}$ is known except for subsets of $\mathcal{X}$ and $\mathcal{Y}$ of size k each. Formally,

$$|S_*| \le k, \quad \text{where } S_* = \{1 \le i \le n : \ \mathbf{\Pi}^*_{ii} \neq 1\} \tag{3}$$

An illustration is provided in Figure 1. The above assumption is often sensible in applications. For example, measurements observed over time may be received in their correct chronological order except for selected periods with higher latencies; in record linkage, auxiliary information such as demographic variables can be used to facilitate the identification of matching records.

In the above setup, recovery of $\mathbf{\Pi}^*$ can be performed in a two-stage manner. In the first stage, one aims at the identification of a reasonably large subset $Q \subseteq S_*^c$, where S_*^c denotes the complement of S_* in (3) associated with shuffled data. The pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i\in Q}$ are correctly matched and can thus be used to estimate the regression parameter $\mathbf{B}^*$ by ordinary least squares. Denote the resulting estimator by $\widehat{\mathbf{B}}$. In the second stage, $\widehat{\mathbf{B}}$ is plugged into the least squares problem (2) in order to obtain the following minimization problem in $\mathbf{\Pi}$ only:

$$\min_{\mathbf{\Pi}\in\mathcal{P}_n} \|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\widehat{\mathbf{B}}\|_F^2 = \min_{\mathbf{\Pi}\in\mathcal{P}_n} -2\,\mathrm{tr}(\mathbf{\Pi}\mathbf{X}\widehat{\mathbf{B}}\mathbf{Y}^\top) + c, \tag{4}$$

where c does not depend on $\mathbf{\Pi}$. The optimization prob-

lem on the right hand side, a so-called linear assignment problem, is computationally tractable. Indeed, Birkhoff's theorem asserts that the constraint set in (4) can be relaxed to its convex hull, the set of doubly stochastic matrices. This yields a specific linear programming problem for which various specialized algorithms have been developed [7].

The main contributions of this paper are as follows. First, we propose and analyze a particularly effective approach based on sparse data representations for identifying a subset of observations $Q \subseteq S_*^c$ not affected by $\mathbf{\Pi}^*$. Clearly, this task becomes more challenging as the fraction of mismatched data $\alpha = k/n$ increases. It is demonstrated that the suggested approach deals more successfully with substantial values of α in the range $(0.4, 0.7)$ than competing methods. Second, we provide a statistical analysis of permutation recovery based on (4). That analysis reveals that the requirement on the signal-to-noise ratio in [34] for a single outcome ($m = 1$) can be considerably relaxed as m increases.

Related Work.
The paper [39] is the first to study (1) under assumption (3) that is referred to as "sparse permutation" therein. While [39] studies the case of a single outcome only, their approach is straightforward to extend to multiple outcomes. Specifically, after introducing an auxiliary variable $\mathbf{\Xi}^* = n^{-1/2}(\mathbf{\Pi}^* - I)\mathbf{X}\mathbf{B}^*$ model (1) becomes

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^* + \sqrt{n}\mathbf{\Xi}^* + \sigma\mathbf{E}.$$

Observing that $\mathbf{\Xi}^*$ has at most k non-zero rows if $\mathbf{\Pi}^*$ is k-sparse in the sense of (3) motivates the formulation

$$\min_{\mathbf{B},\mathbf{\Xi}} \frac{1}{n \cdot m}\|\mathbf{Y} - \mathbf{X}\mathbf{B} - \sqrt{n}\mathbf{\Xi}\|_F^2 + \lambda \sum_{i=1}^{n} \|\mathbf{\Xi}_{i,:}\|_2, \quad (5)$$

where $\{\mathbf{\Xi}_{i,:}\}_{i=1}^n$ denote the rows of $\mathbf{\Xi} \in \mathbb{R}^{n\times m}$. Row sparsity is promoted via the group lasso penalty [50]. Approach (5) is both computationally convenient and amenable to statistical analysis. However, a serious drawback of (5) is that the fraction of permuted data $\alpha = k/n$ that can be tolerated is comparatively small, and particularly must not exceed 0.5. By contrast, if the set S_* were known in advance, $\mathbf{B}^*$ could be estimated at the usual rate as long as a constant (but potentially small) fraction of the observations is correctly matched.

The paper [38] considers setting (1) under a spherical regression model, i.e., the columns of both $\mathbf{Y}$ and $\mathbf{X}$ are assumed to be elements of the respective unit sphere. While the sparsity assumption (3) is adopted in [38] as well, the authors additionally assume $\mathbf{\Pi}^*$ to be block diagonal with known block structure. They suggest to estimate $\mathbf{B}^*$ and $\mathbf{\Pi}^*$ in an alternating fashion, starting from the ordinary least squares estimator of $\mathbf{B}^*$ without adjustment for mismatches. The approach is thus not well-suited to the case in which α is substantial.

2 APPROACH: SPARSE DATA REPRESENTATION

We start with the noiseless case $\sigma = 0$. We let $\mathbf{Z} = [\mathbf{X}\ \ \mathbf{Y}]^\top \in \mathbb{R}^{D\times n}$, $D = d + m$, with columns $\mathbf{z}_i = [\mathbf{x}_i^\top\ \mathbf{y}_i^\top]^\top$, $1 \le i \le n$. We further write $\mathbf{Z}_{S_*}$ and $\mathbf{Z}_{S_*^c}$ for the column sub-matrices corresponding to $S_* = \{1 \le i \le n : \mathbf{\Pi}^*_{ii} \neq 1\}$ and its complement S_*^c, respectively. Without loss of generality, $S_* = \{1, \ldots, k\}$. We let $\mathbf{\Pi}^*_{S_*}$ be the principal submatrix of $\mathbf{\Pi}^*$ corresponding to S_*. For simplicity, we also assume that $\mathrm{rank}(\mathbf{B}^*) = m \le d$.

The first observation is that the points $\{\mathbf{z}_i\}_{i\in S_*^c}$ are contained in a d-dimensional subspace, say $\mathcal{Z}_{S_*^c}$, of $\mathbb{R}^D$. This follows immediately from the fact that

$$\mathbf{Z}_{S_*^c} = \begin{bmatrix} I_{d\times d} \\ \mathbf{B}^{*\top} \end{bmatrix} \mathbf{X}_{S_*^c}^\top, \quad (6)$$

where $\mathbf{X}_{S_*^c}$ denotes the row submatrix of $\mathbf{X}$ corresponding to S_*^c. By contrast, the points $\{\mathbf{z}_i\}_{i\in S_*}$ are generally not contained in a lower-dimensional subspace. Moreover, if the entries of $\mathbf{X}$ are drawn i.i.d. from a continuous distribution, then $\mathcal{Z}_{S_*^c} \subset \mathcal{Z}_{S_*} = \mathrm{span}\{\mathbf{z}_i\}_{i\in S_*}$ with probability one given $|S_*| \ge 2d$. To see this, note that

$$\mathbf{Z}_{S_*} = \begin{bmatrix} I_{d\times d} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^{*\top} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{S_*}^\top \\ \mathbf{X}_{S_*}^\top \mathbf{\Pi}_{S_*}^\top \end{bmatrix}, \quad (7)$$

$$\begin{bmatrix} I_{d\times d} \\ \mathbf{B}^{*\top} \end{bmatrix} = \begin{bmatrix} I_{d\times d} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^{*\top} \end{bmatrix} \begin{bmatrix} I_{d\times d} \\ I_{d\times d} \end{bmatrix}. \quad (8)$$

The rightmost matrix in (7) has $2d$ linearly independent rows with probability one once $|S_*| \ge 2d$, hence the range of that matrix equals $\mathbb{R}^{2d}$. Combining this with (6) and (8) yields $\mathcal{Z}_{S_*^c} \subset \mathcal{Z}_{S_*}$. In summary, the points $\{\mathbf{z}_i\}_{i=1}^n$ can be split into low-dimensional data $\{\mathbf{z}_i\}_{i\in S_*^c}$ (inliers) and outliers $\{\mathbf{z}_i\}_{i\in S_*}$. Fig. 2 provides an illustration for $D = 2$ $(d, m = 1)$ and $D = 3$ $(d = 2, m = 1)$.

According to the above considerations, if $\sigma = 0$, $\mathbf{B}^*$ can be recovered exactly by identifying at least d linearly independent inliers forming a basis of $\mathcal{Z}_{S_*^c}$. Given the latter, it is then also straightforward to classify each of the $\{\mathbf{z}_i\}_{i=1}^n$ as either inlier or outlier.

In order to distinguish between inliers and outliers, we use the fact that each element in $\{\mathbf{z}_i\}_{i\in S_*^c}$ can be expressed as a linear combination of d other points, whereas the elements of $\{\mathbf{z}_i\}_{i\in S_*}$ generally require $d+m$ other points. In other words, elements in S_* and S_*^c can

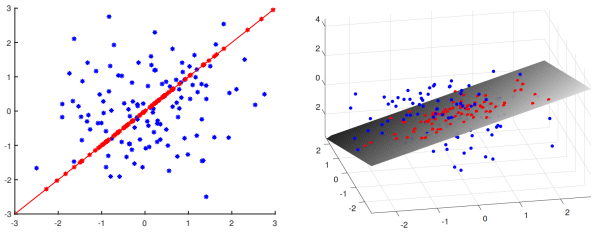


Figure 2: Illustration of the inlier/outlier representation as described in the text. Left: $d = 1$, $m = 1$. Right: $d = 2, m = 1$. Red dots correspond to inliers (correct matches), blue dots to outliers (mismatches).

be separated in terms of the optimum value of the following sequence of optimization problems:

$$\min_{\boldsymbol{\alpha}_i \in \mathbb{R}^{n-1}} \|\boldsymbol{\alpha}_i\|_0 \quad \text{subject to } \mathbf{z}_i = \sum_{j \neq i} \alpha_i^{(j)} \mathbf{z}_j, \qquad \boldsymbol{\alpha}_i = (\alpha_i^{(j)}), \quad i \in \{1, \ldots, n\}, \tag{9}$$

where $\|\cdot\|_0$ returns the number of non-zero entries of a vector. Denoting by $\mathsf{optval}_i^{\ell_0}$ the optimal function value of (9) for observation i, we have $\max_{i \in S_*^c} \mathsf{optval}_i^{\ell_0} < \min_{i \in S_*} \mathsf{optval}_i^{\ell_0}$. Optimization problem (9) appears in the literature on sparse subspace clustering (SSC), e.g. [17, 31, 41, 46, 47]. A common approach is to replace $\|\cdot\|_0$ in (9) by the ℓ_1-norm $\|\cdot\|_1$ [17, 31, 41]. This yields Algorithm 1 below. Its refined (and computationally more demanding) version, in which (10) is solved repeatedly for increasingly smaller portions of the data to achieve successive outlier removal, yields a further boost in performance and often succeeds even in regimes close to the definite limit of recovery with $k = n - d$.

Algorithm 1 Identifying the Subspace $\mathcal{Z}_{S_*^c}$

Data Preparation: Normalize the columns of the matrix $\mathbf{Z} = [\mathbf{X}\ \mathbf{Y}]^\top$ to unit ℓ_2-norm.

1. Sparse Representation: Obtain $\widehat{\mathbf{A}} \in \mathbb{R}^{n \times n}$ as the minimizer of the optimization problem

$$\min_{\mathbf{A}} \|\mathbf{A}\|_1 \ \text{ subject to } \ \mathbf{Z} = \mathbf{Z}\mathbf{A}, \ \ \mathrm{diag}(\mathbf{A}) = \mathbf{0} \tag{10}$$

2. Basis Selection: Let $b_i = \|\widehat{\mathbf{A}}_{:,i}\|_1$, $1 \leq i \leq n$. Select the columns of $\mathbf{Z}$ that correspond to the d smallest values of $\{b_i\}_{i=1}^n$ as guess for the basis of the subspace $\mathcal{Z}_{S_*^c}$.

Refinement: Modify step 2. as follows: remove the columns of $\mathbf{Z}$ corresponding to the $\lfloor \eta n \rfloor$, $\eta \in (0, 1)$, largest values of $\{b_i\}_{i=1}^n$ and iterate both steps until $\mathcal{Z}_{S_*^c}$ is obtained from the d smallest values of $\{b_i\}_{i=1}^n$.

Rationale.
It is important to the study the consequences of replacing the ℓ_0-norm by the ℓ_1-norm when moving from (9) to (10). Intuitively, we expect the corresponding objective value $\mathsf{optval}_i^{\ell_1}$ to be of the order of $\sqrt{d}$ for $i \in S_*^c$ and $\sqrt{d+m}$ for $i \in S_*$: after normalizing each point $\mathbf{z}_i \leftarrow \mathbf{z}_i / \|\mathbf{z}_i\|_2$, $1 \leq i \leq n$, each problem is supposed to have $O(d)$ and $O(d+m)$ non-zero coefficients of magnitude $O(1/\sqrt{d})$ and $O(1/\sqrt{d+m})$, respectively. A rigorous analysis of the ℓ_1-relaxation for generic SSC for specific data-generating models is given in [41]. In the following, we leverage results in [41] to deduce separation of $\{\mathsf{optval}_i^{\ell_1}\}_{i \in S_*}$ and $\{\mathsf{optval}_i^{\ell_1}\}_{i \in S_*^c}$.

Proposition 1. *Suppose that the entries of* $\mathbf{X}$ *are i.i.d. standard Gaussian, and that the matrix* $\mathbf{B}^*$ *has orthonormal columns. Then for any* $t > 0$

$$\min_{i \in S_*} \mathsf{optval}_i^{\ell_1} \ \geq (1-t)\lambda(\tau)\sqrt{d+m}$$

$$\max_{i \in S_*^c} \mathsf{optval}_i^{\ell_1} \ \leq \frac{\sqrt{d}}{c(\rho)\sqrt{\log \rho}}$$

with probability $\geq 1 - n\exp(-ct^2(d+m)) - n\exp(-d)$, *where* $\lambda(\tau)$ *is a function of* $\tau = n/(d+m)$, $\rho = (n-k)/d$ *is the oversampling ratio, and* $c(\rho) \geq \frac{1}{\sqrt{8}}$ *is a constant depending only on* ρ [41].

Note that separation of $\{\mathbf{z}_i\}_{i \in S_*}$ and $\{\mathbf{z}_i\}_{i \in S_*^c}$, i.e., $\max_{i \in S_*^c} \mathsf{optval}_i^{\ell_1} < \min_{i \in S_*} \mathsf{optval}_i^{\ell_1}$ is not required for Algorithm 1 to succeed since we only need that

$$\min_{T \subseteq S_*^c, |T| \geq d} \max_{i \in T} \mathsf{optval}_i^{\ell_1} < \min_{i \in S_*} \mathsf{optval}_i^{\ell_1}.$$

Proposition 1 indicates that separation is more likely to hold as m, d, or $(n-k)/d$ increase, where we recall that $m = \mathrm{rank}(\mathbf{B}^*) \leq d$. Our numerical results in §4 confirm the insights conveyed by the above proposition.

Extension to the noisy case.
In the presence of noise ($\sigma > 0$), the subspace geometry as outlined above only holds approximately. However, Algorithm 1 extends to a noisy regime by replacing the ℓ_1-minimization problem (10) by a lasso-type formulation [31], and sampling $\nu \cdot d$, $\nu > 1$, observations rather than just d in order to stabilize subsequent least squares fitting. If the noise level σ is known, the oversampling factor ν can be chosen in a data-driven manner. Details are provided in Algorithm 2 above. Following [31], we suggest to choose the regularization parameter λ in (11) as $\lambda = cd^{-1/2}$ for $c > 0$. In our experiments we observed good performance for $c \in [0.05, 0.2]$. The rationale behind the selection of the oversampling factor via (12) is that if not enough points are sampled, we observe overfitting, i.e., $\widehat{\sigma}_\ell \ll \sigma$. By contrast, $\widehat{\sigma}_\ell \gg \sigma$ indicates that we have sampled too many points including outliers. One possible improvement is to replace the least squares fits in (12) and (13) by a robust regression procedure like the one based on the group lasso (5) since the subsets $\mathcal{I}_\ell$ potentially may contain few outliers as well.

Algorithm 2 Estimation of $\mathbf{B}^*$ by outlier removal

Data Preparation: Normalize the columns of the matrix $\mathbf{Z} = [\mathbf{X}\ \mathbf{Y}]^\top$ to unit ℓ_2-norm.

1. Sparse Representation: Obtain $\widehat{\mathbf{A}} \in \mathbb{R}^{n\times n}$ as the minimizer of the optimization problem

$$\|\mathbf{Z} - \mathbf{Z}\mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_1 \ \text{ subject to } \ \mathrm{diag}(\mathbf{A}) = \mathbf{0}. \quad (11)$$

2. Choosing the Oversampling Factor: Let $b_i = \|\widehat{\mathbf{A}}_{:,i}\|_1$, $i = 1, \dots, n$, and $\{\nu_1, \dots, \nu_M\} \subset [1, n/d]$.

For $\ell = 1, \dots, M$:

Obtain $\mathcal{I}_\ell$ as the subset of $\{1, \dots, n\}$ corresponding to the $\lfloor \nu_\ell \cdot d \rfloor$ smallest values among the $\{b_i\}_{i=1}^n$, and let

$$\widehat{\sigma}_\ell = \min_{\mathbf{B}\in\mathbb{R}^{d\times m}} \frac{1}{\sqrt{n\cdot m}} \|\mathbf{Y}_{\mathcal{I}_\ell} - \mathbf{X}_{\mathcal{I}_\ell}\mathbf{B}\|_F \quad (12)$$

Let $\ell^* = \mathrm{argmin}_{1\le\ell\le M} |\widehat{\sigma}_\ell/\sigma - 1|$, and return

$$\widehat{\mathbf{B}} = \operatorname*{argmin}_{\mathbf{B}\in\mathbb{R}^{d\times m}} \|\mathbf{Y}_{\mathcal{I}_{\ell^*}} - \mathbf{X}_{\mathcal{I}_\ell}\mathbf{B}\|_F. \quad (13)$$

Robust PCA via Matrix Decomposition.
The subspace geometry discussed above also suggests an alternative route in which one aims to detect outliers by means of a low rank plus column sparse decomposition of the matrix $\mathbf{Z}$, an approach often referred to as robust PCA [8, 29, 49]. Specifically, for $\sigma = 0$ we have

$$\mathbf{Z} = \underbrace{\begin{bmatrix} \mathbf{X}^\top \\ \mathbf{X}^\top\mathbf{B}^{*\top} \end{bmatrix}}_{=\mathbf{L}^*} + \underbrace{\begin{bmatrix} \mathbf{0} \\ (\mathbf{\Pi}^* - I)\mathbf{X}^\top\mathbf{B}^{*\top} \end{bmatrix}}_{=\mathbf{C}^*}$$

This suggests the following optimization problem:

$$\min_{\mathbf{L},\mathbf{C}} \|\mathbf{C}\|_{2,1} \text{ sb. to } \mathbf{Z} = \mathbf{L} + \mathbf{C},\ \|\mathbf{L}\|_\star \le \sqrt{d}\|\mathbf{Z}\|_F \quad \mathbf{C}_{i,:} = \mathbf{0},\ 1 \le i \le d, \quad (14)$$

where $\|\cdot\|_{2,1}$ returns the sum of column ℓ_2-norms and $\|\cdot\|_\star$ denotes the sum of singular values. These norms serve as convex surrogate for the number of non-zero columns and the matrix rank, respectively; the bound on $\|\mathbf{L}\|_\star$ follows from $\|\mathbf{\Pi}^*\mathbf{X}\mathbf{B}^*\|_F = \|\mathbf{X}\mathbf{B}^*\|_F$ and

$$\|\mathbf{L}^*\|_\star \le \sqrt{\mathrm{rank}(\mathbf{L}^*)}\|\mathbf{L}^*\|_F = \sqrt{d}\|\mathbf{Z}\|_F.$$

In the presence of noise, (14) can be modified as follows:

$$\min_{\mathbf{L},\mathbf{C}} \frac{1}{2}\|\mathbf{Z} - \mathbf{L} - \mathbf{C}\|_F^2 + \lambda\|\mathbf{C}\|_{2,1} \text{ sb. to } \|\mathbf{L}\|_\star \le R \quad (15)$$

for suitable choices of $\lambda, R > 0$. While (14) and (15) are worth a consideration, they are outperformed by the proposed approach in the challenging regime with a substantial fraction of mismatches. The benefit over the regression formulation (5) is unclear as well since the latter seems to better incorporate the specific underlying low-rank structure. We refer to §4 for empirical comparisons.

3 PERMUTATION RECOVERY

In this section we discuss estimation of $\mathbf{\Pi}^*$ given an estimator $\widehat{\mathbf{B}}$ of $\mathbf{B}^*$. To begin with, we suppose that $\mathbf{B}^*$ is perfectly known. In this case, the problem

$$\min_{\mathbf{\Pi}\in\mathcal{P}_n} \|\mathbf{Y} - \mathbf{\Pi}\mathbf{X}\mathbf{B}^*\|_F^2 = \min_{\mathbf{\Pi}\in\mathcal{P}_n} -2\,\mathrm{tr}(\mathbf{\Pi}\mathbf{X}\mathbf{B}^*\mathbf{Y}^\top) + c \quad (16)$$

reduces to a specific linear program as discussed in the introduction subsequent to (4). While it is tempting to impose a sparsity constraint on $\mathbf{\Pi}$ also at this point, this is not adequate as such additional constraint generally affects the integrality of the convex relaxation of (16).

A sufficient condition for permutation recovery, i.e., the event $\mathcal{R} = \{\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\}$, where $\widehat{\mathbf{\Pi}}$ denotes a minimizer of (4), is given by the following statement.

Lemma 1. *Suppose that the noise matrix* $\mathbf{E}$ *has i.i.d.* $N(0,1)$*-entries. Define*

$$\gamma := \min_{i<j} \|\mathbf{B}^{*\top}(\mathbf{x}_j - \mathbf{x}_i)\|_2. \quad (17)$$

Consider the event $\mathcal{E} = \{\gamma \ge 4\sigma\sqrt{\log(n/(2\delta))}\}$*. We then have* $\mathbf{P}(\mathcal{R}|\mathcal{E}) \ge 1 - \delta$ *for all* $\delta \in (0,1)$*.*

The appearance of the quantity γ as a measure of minimum separation between two pairs of data points is unsurprising given that in the presence of noise it becomes hard to correctly identify matching pairs without such separation. In the sequel, we provide a lower bound on γ for Gaussian $\mathbf{X}$ as we shall assume throughout the rest of the section. The following result is an immediate consequence of Proposition 6 in [28].

Lemma 2. *There exist universal constants* $\alpha_0 \in (0,1)$ *and* $\kappa > 0$ *such that for any* $\alpha \in (0, \alpha_0)$

$$\mathbf{P}\left(\min_{i<j} \|\mathbf{B}^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2 \le \alpha\|\mathbf{B}^*\|_F\right) \le \exp\left(\kappa\log(\alpha)\,\mathrm{srank}(\mathbf{B}^*) + \log(n^2/2)\right),$$

where $\mathrm{srank}(\mathbf{B}^*) = (\|\mathbf{B}^*\|_F/\|\mathbf{B}^*\|_{\mathrm{op}})^2 \le \mathrm{rank}(\mathbf{B}^*) \le \min\{d, m\}$ *denotes the stable rank of* $\mathbf{B}^*$*.*

Using Lemmas 1 and 2 yields the following theorem.

Theorem 1. *Define the signal-to-noise ratio by* $\mathtt{SNR} = \|\mathbf{B}^*\|_F^2/\sigma^2$*, let* $\kappa > 0$*,* $\alpha_0 \in (0,1)$ *be as in Lemma 2, and* $\epsilon > 0$*,* $\delta \in (0,1)$*. We have* $\mathbf{P}(\mathcal{R}) \ge 1 - n^{-\epsilon} - \delta$ *if*

$$\mathtt{SNR} > 16\log(n/(2\delta)) \cdot \max\{\alpha_0^{-2},\ n^{\frac{4\cdot(1+\epsilon)}{\kappa\cdot srank(\mathbf{B}^*)}}\}. \quad (18)$$

In the special case srank$(\mathbf{B}^*) = O(1)$, Theorem 1 recovers the requirement $\texttt{SNR} \gtrsim n^C$ for $C > 0$ shown in [34] for permutation recovery based on the (computationally intractable) problem (2) and a single outcome ($m = 1$).

As an important implication of Theorem 1 we obtain that if srank$(\mathbf{B}^*) \gtrsim \log n$, the requirement on the SNR becomes far less stringent. For instance, if $\mathbf{B}^* = b\mathbf{Q}$, $b > 0$, with $\mathbf{Q}$ having $m \gtrsim \log n$ orthonormal columns so that $\|\mathbf{B}^*\|_F^2 \gtrsim b^2 \log n$, constant signal-to-noise ratio $b^2/\sigma^2 = \Omega(1)$ per dimension suffices for permutation recovery. This constitutes a dramatic improvement compared to the case $m = 1$ studied earlier [34, 39].

While condition (18) is sufficient for permutation recovery based on (4) with $\mathbf{B}^*$ given, a similar condition can be shown to be necessary as a consequence of the next statement that can be seen as a converse to Lemma 2.

Proposition 2. *Let $\mathbf{B}^* = b\mathbf{Q}$ as above, assuming additionally that $m = r = \text{rank}(\mathbf{B}^*) = 2(q+1)$ for a non-negative integer q is an even number. Then if $n > 8(r/2)^{r/2}$, with probability at least .75*

$$\min_{i<j} \|\mathbf{B}^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2 \leq \|\mathbf{B}^*\|_F \cdot 8^{1/r} n^{-1/r}.$$

Proposition 2 confirms that if rank$(\mathbf{B}^*) = O(1)$ an excessively large SNR is needed for permutation recovery. From the analysis in [12], it is known that a separation condition in terms of a lower bound on γ (17) is necessary for any estimator and not only (16).

Plug-in of $\widehat{\mathbf{B}}$.
Optimization problem (16) is not practical since $\mathbf{B}^*$ is not known. A natural approach is to replace $\mathbf{B}^*$ by the estimator $\widehat{\mathbf{B}}$ obtained from the approach in §2. Regarding the pivotal quantity γ (17), the triangle inequality yields

$$\begin{aligned}
&\min_{i<j} \|\widehat{\mathbf{B}}^\top(\mathbf{x}_i - \mathbf{x}_j)\|_2 \\
&= \min_{i<j} \|((\widehat{\mathbf{B}} - \mathbf{B}^*)^\top + \mathbf{B}^{*\top})(\mathbf{x}_i - \mathbf{x}_j)\|_2 \\
&\geq \min_{i<j} \|\mathbf{B}^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2 - 2 \max_{1\leq i\leq n} \|\mathbf{x}_i\|_2 \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{\mathrm{F}} \quad (19)
\end{aligned}$$

For Gaussian $\{\mathbf{x}_i\}_{i=1}^n$, standard concentration arguments show that with high probability

$$\max_{1\leq i\leq n} \|\mathbf{x}_i\|_2 \leq \sqrt{d} + 2\sqrt{\log n}.$$

Given that Algorithm 2 allows rate-optimal estimation of $\mathbf{B}^*$, i.e., $\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F \lesssim \sigma\sqrt{(d \cdot m)/n}$, (19) becomes

$$\min_{i<j} \|\widehat{\mathbf{B}}^\top(\mathbf{x}_i - \mathbf{x}_j)\|_2 \geq \gamma - O\left(\frac{(d \vee \log n) \cdot \sqrt{m}}{\sqrt{n}}\right).$$

In view of Lemma 2, if srank$(\mathbf{B}^*) = \Omega(m)$ and $m = \Omega(\log n)$, γ will be proportional to $\|\mathbf{B}\|_F^* \gtrsim \sqrt{m}$. The $O(\cdot)$ term will be of lower order if $d = o(\sqrt{n})$, and the analysis for known $\mathbf{B}^*$ continues to apply.

4 EXPERIMENTS

In this section, we present the results of experiments with synthetic and real data in order illustrate some central features of the approach proposed herein.

Synthetic Data

For all synthetic data experiments, the entries of $\mathbf{X}$ are drawn i.i.d. from the $N(0,1)$-distribution. The coefficient matrix $\mathbf{B}^*$ is generated in the same way, and subsequently its columns are orthornormalized.

Noiseless case.
We let $n = 200$, $d \in \{20, 50\}$, and $m = m_d \in \{1, 2, 5, 10, 20\}$ for $d = 20$ and $m = \{1, 5, 10, 20, 50\}$ for $d = 50$. Moreover, for the fraction of mis-matched data $\alpha = k/n$, we consider $\{0.1, 0.15, \ldots, 0.85\}$. We compare the results of Algorithm 1 to the noiseless counterpart of the group lasso formulation (5)

$$\min_{\mathbf{B},\boldsymbol{\Xi}} \sum_{i=1}^n \|\boldsymbol{\Xi}_{i,:}\|_2 \quad \text{subject to } \mathbf{Y} = \mathbf{X}\mathbf{B} + \boldsymbol{\Xi}, \qquad (20)$$

as well as to the robust PCA formulation (14).
For each triplet of (d, m, α), we perform twenty independent replications. Optimization problems (9) in Algorithm 1 as well as (14) and (20) are solved with CVX [21]. The refined version of Algorithm 1 is run with a custom ADMM solver for efficiency.

Selected results are shown in Figure 3. We observe that larger m leads to improvements since it increases the separation of inliers and outliers in terms of optval^{ℓ_1} as explained in §2. Moreover, smaller n/d as well as larger fraction of mismatches α make the problem more challenging. The group lasso approach (20) is outperformed by Algorithm 1 whenever m and α are large. In fact, in all scenarios shown, (20) cannot handle values of α larger than .55. On the other hand, for small m and small α (20) performs better. Robust PCA (14) is outperformed by at least one of these two approaches in all regimes.

Noisy case.
We generate data according to model (1) for $\sigma \in \{0.05, 0.1, 0.2, 0.5, 1\}$, $d = m \in \{20, 50\}$ with $\mathbf{X}$ and $\mathbf{B}^*$ and the grid for α as in the noiseless case. Twenty replications are considered per setup. Algorithm 2 is run with $\lambda = 0.1/\sqrt{d}$ using the code from [40] and oversampling factors $\nu \in \{1.2, 1.5, 2, 2.5, 3\}$. As competitors, we consider the group lasso formulation (5) with $\lambda = 2\sqrt{2} \cdot \sigma \cdot \sqrt{1/(n \cdot m)}$, cf. [30] and robust PCA (15) with $R = \|[\mathbf{X}^\top; \mathbf{B}^{*\top}\mathbf{X}]\|_\star$ and $\lambda = \frac{\sigma}{2} \cdot (\sqrt{m} + \sqrt{\log n})$[2]. Subsequently, we plug in the resulting estimators for $\mathbf{B}^*$ in (4) to recover $\mathbf{\Pi}^*$. Problem (4) is solved by an implementation of the auction algorithm [5]. While not

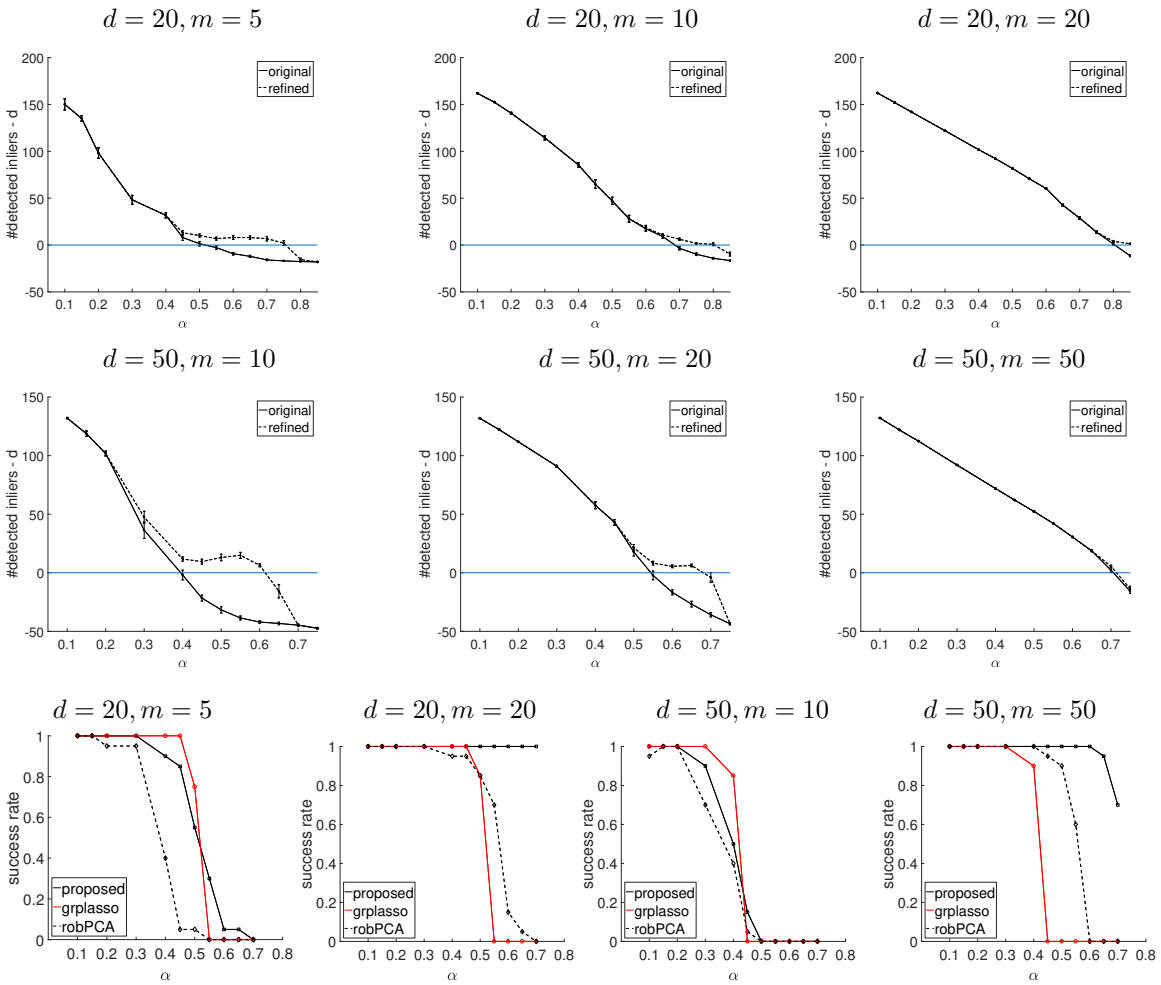


Figure 3: Selected results for the noiseless case. Top/Middle panel: #detected inliers $-d$ (average $\pm$ standard error) based on Algorithm 1 depending on $\alpha = k/n$ (horizontal axis). #detected inliers is defined as the number of observations in S_*^c for which $b_i = \mathsf{optval}_i^{\ell_1} < \min_{j \in S_*} \mathsf{optval}_j^{\ell_1}$. Success of the approach is equivalent to #detected inliers $\geq d$. The horizontal line at zero indicates the threshold for success. Bottom panel: Percentage of successful recovery depending on α for the proposed approach (black) and the baseline competitors (20) (red) and (14) (dashed).

included in this paper, the situation for inlier recovery looks similar as depicted in Figure 3 for moderate levels of σ, i.e., the presented approach is stable with respect to noise. Even more, as shown in Figure 4, the errors $\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F/\sqrt{m}$ are not far from the oracle rate $\sigma\sqrt{d/n}$. Similar as in the noiseless case, the two baselines group lasso and robust PCA are competitive, but fall short of the suggested method as α increases. A similar observation can be made for permutation recovery (Figure 5). Interestingly, in our experiments the permutation is recovered perfectly in all runs for small level of σ by all competitors. Notable differences arose only for $\sigma \geq 0.5$.

Real Data

We consider data from the trading agent competition in supply chain management [22, 42] available via [43]. The goal is to forecast the prices of $m = 16$ different computer systems 20 days in the future using 61 features capturing market conditions (bank interest rate, current and lagged product demands in three different price segments), current product prices and component prices. The number of observations (days) equals $n = 8966$. Due to high feature correlations, the matrix of inputs $\mathbf{X}$ is reduced to its top $d = 35$ principal components. We here work with $\mathbf{B}^*$ as the resulting least squares estimator, i.e., $\mathbf{B}^* = \operatorname{argmin}_{\mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2$, where the n-by-m matrix $\mathbf{Y}$ contains the 16 outcome variables. We subsequently create shuffled versions of $\mathbf{Y}$ in the form of $\mathbf{\Pi}^*\mathbf{Y}$, where the permutation $\mathbf{\Pi}^*$ is generated by first selecting $\lfloor \alpha n \rfloor$ indices uniformly at random from

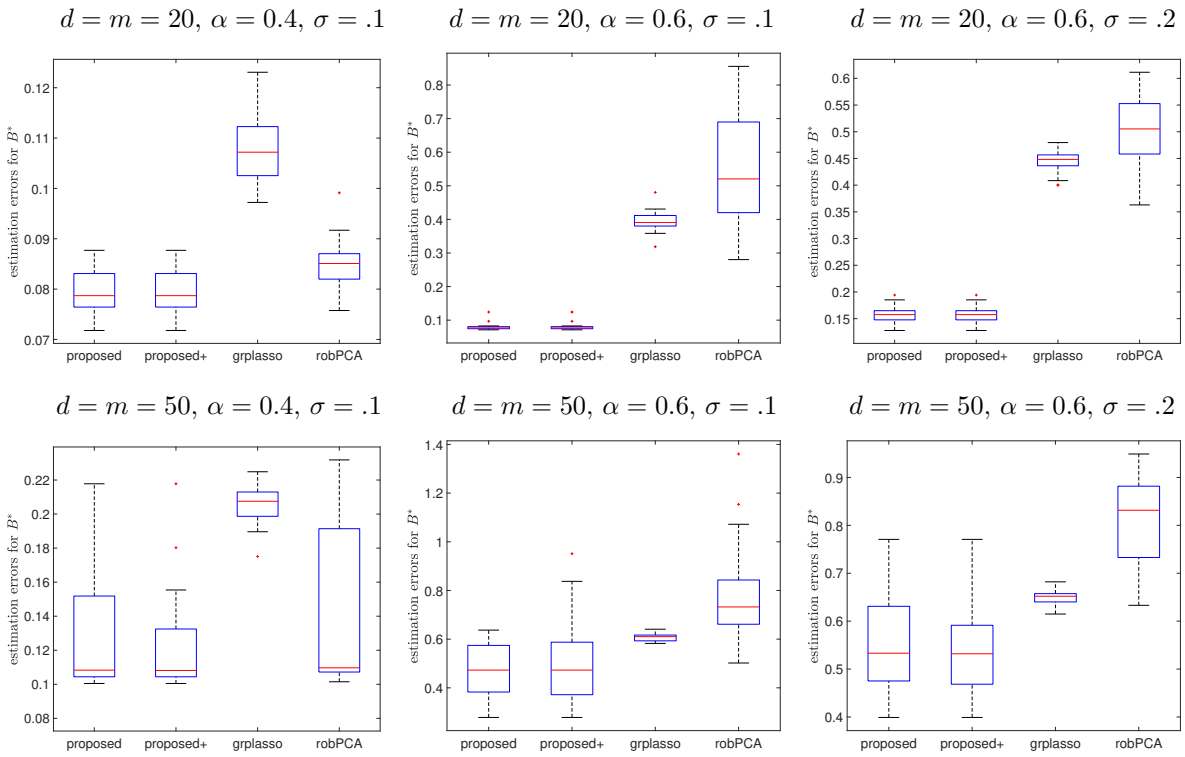


Figure 4: Boxplots of the estimation errors $\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F/\sqrt{m}$ for selected settings. "proposed" refers to Algorithm 2, "proposed+" refers to a hybrid of Algorithm 2 and (5) in which the least squares fits (12) and (13) are replaced by group lasso fits (5). "grplasso" and "robPCA" refer to the plain group lasso and robust PCA, respectively.

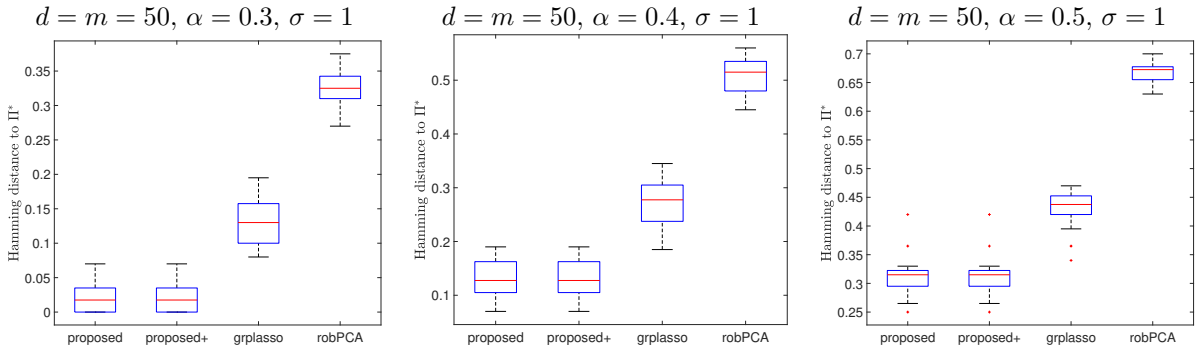


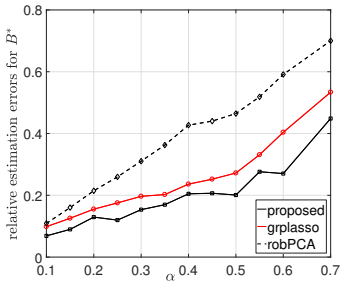
Figure 5: Boxplots of the (averaged) Hamming distance between estimated and underlying permutation for selected settings. The annotation of this plot is as for Figure 4.

$\{1, \ldots, n\}$, and then randomly permuting them. All other indices are not affected by $\mathbf{\Pi}^*$.

We evaluate the competitors above with regard to $\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F/\|\mathbf{B}^*\|_F$ and $\|\mathbf{\Pi}^*\mathbf{Y} - \widehat{\mathbf{\Pi}}\mathbf{Y}\|_F/\|\mathbf{Y}\|_F$, where given $\widehat{\mathbf{B}}$, the estimate $\widehat{\mathbf{\Pi}}$ is obtained from the plug-in approach in §3. We use the latter metric to assess performance in estimating $\mathbf{\Pi}^*$ since the separability condition in Lemma 1 for permutation recovery w.r.t. Hamming distance is not satisfied here. The results are shown in Fig. 6.

5 CONCLUSION

Broken Sample Problems such as regression with arbitrarily permuted data are notoriously challenging, and no practical algorithms with theoretical guarantees are available at this point. In this paper, we have studied a more benign sub-case in which a sufficiently large fraction (but potentially less than 0.5) of the observations are in the right correspondence. We have demonstrated herein that this situation can be solved by means of a formulation based on sparse representations for the cor-



	α=.2	α=.3	α=.4	α=.5	α=.6	α=.7
proposed	**22.49**	**26.87**	31.30	**34.90**	38.77	**42.05**
grplasso	22.59	26.91	**31.28**	34.95	38.67	42.18
robPCA	22.62	27.07	31.27	34.97	**38.64**	42.45

Figure 6: Results on the supply chain management data set. The plot shows $\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{\mathrm{F}}$ vs. fraction of mismatches α. The table shows the relative error (in percent) in recovering $\mathbf{\Pi}^*\mathbf{Y}$.

rectly matched data. A large number of (not too strongly correlated) outcomes appears to be crucial to the success of the approach: it not only allows for a clear separation between correctly matched data and mismatched data, but also significantly facilitates permutation recovery as elaborated in §3. While we have shown (Lemma 2) that at least of the order of $\log n$ outcomes are necessary for permutation recovery at a constant SNR, it is an interesting direction of future research whether at least the regression parameter can be recovered in case of a large fraction of mismatches and a small number of outcomes.

Appendix

Proof of Proposition 1. We first note that the assumption on $\mathbf{B}^*$ implies that $\mathbf{z}_i/\|\mathbf{z}_i\|_2$ follow uniform distributions on the unit spheres of $\mathbb{R}^D$ (for $i \in S_*$) and range$([I_{d\times d}; \mathbf{B}^{*\top}])$ (for $i \in S_*^c$), respectively. In fact, for any $i \in S_*$, $\pi^*(i) = j \neq i$ and hence

$$\mathbf{z}_i = [\mathbf{x}_i^\top \;\; \mathbf{y}_i^\top] = [\mathbf{x}_i^\top \;\; (\mathbf{x}_j^\top \mathbf{B}^*)^\top]^\top \sim N(0, I_D).$$

On the other hand, uniformity on range$([I_{d\times d}; \mathbf{B}^{*\top}])$ for $i \in S_*^c$ is immediate from the observation that $[I_{d\times d}; \mathbf{B}^{*\top}]$ the matrix has orthonormal columns. The above observations allow application of Theorem 2.9 in [41] which then yields the claim of the proposition.

Proof of Lemma 1. Event $\mathcal{R} = \{\widehat{\mathbf{\Pi}} = \mathbf{\Pi}^*\}$ is implied by

$$\big\{\|\mathbf{B}^{*\top}\mathbf{x}_i - \mathbf{y}_i\|_2^2 < \min_{j\neq i}\|\mathbf{B}^{*\top}\mathbf{x}_j - \mathbf{y}_i\|_2^2,\ 1 \le i \le n\big\}$$

Expanding the squares and elementary re-arrangements show that the above event is in turn implied by

$$\gamma > \max_{i<j} 2\sigma \underbrace{\left\langle \boldsymbol{\epsilon}_i, \frac{\mathbf{B}^{*\top}(\mathbf{x}_j - \mathbf{x}_i)}{\|\mathbf{B}^{*\top}(\mathbf{x}_j - \mathbf{x}_i)\|_2} \right\rangle}_{\xi_{ij}}$$

with γ defined in (17), and $\{\boldsymbol{\epsilon}_i\}_{i=1}^n$ are the rows of the noise matrix $\mathbf{E}$. Note that conditional on $\mathbf{X}$, $\xi_{ij} \sim N(0,1)$ for all $i < j$. Standard concentration arguments then show that the right hand side in the previous display is upper bounded by $2\sigma\sqrt{2\log\left(\binom{n}{2}/\delta\right)}$ with probability at least $1-\delta$, $\delta \in (0,1)$.

Proof of Lemma 2. This lemma follows immediately from Proposition 6 in [28] and a union bound.

Proof of Theorem 1. First invoke Lemma 2 with the choice $\alpha = \min\{\alpha_0, n^{\frac{-2\cdot(1+\epsilon)}{\kappa\cdot \mathrm{srank}(\mathbf{B}^*)}}\}$ to conlude that $\gamma \ge \alpha\|\mathbf{B}^*\|_F$ with probability at least $1 - n^{-\epsilon}$. Plugging this lower bound on γ into the condition of event $\mathcal{E}$ in Proposition 1 yields $\alpha\|\mathbf{B}^*\|_F \ge 4\sigma\sqrt{\log(n/(2\delta))} \Leftrightarrow \mathtt{SNR} = \|\mathbf{B}^*\|_F^2/\sigma^2 \ge 16\alpha^{-2}\log(n/(2\delta))$.

Proof of Proposition 2. Let $n_2 = n/2$, and let $\{\chi_i^2(k)\}_{i=1}^{n_2}$ be independent χ^2-random variables with k degrees of freedom. For any $t \ge 0$, we have

$$\begin{aligned}
&\mathbf{P}\left(\min_{i<j}\|\mathbf{B}^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 < t^2\right)\\
&\ge \mathbf{P}\left(\min_{1\le i\le n/2}\|\mathbf{B}^{*\top}(\mathbf{x}_{2i} - \mathbf{x}_{2i-1})\|_2^2 < t^2\right)\\
&= \mathbf{P}\left(2b^2 \min_{1\le i\le n/2}\chi_i^2(r) < t^2\right)\\
&= 1 - \mathbf{P}(\chi_1^2(r) > t^2/2b^2)^{n_2} \qquad (21)
\end{aligned}$$

Now if $r = 2(q+1)$ is even, the CDF of $\chi_1^2(r)$ has the following closed form expression:

$$\mathbf{P}(\chi_1^2(r) \le z) = 1 - \exp(-z/2)\textstyle\sum_{s=0}^q \frac{(z/2)^s}{s!}, \quad z \ge 0.$$

Choosing $t^2 = c\cdot 2b^2$ in (21) for $c > 0$ to be determined below, we obtain that

$$\begin{aligned}
&\mathbf{P}\left(\min_{i<j}\|\mathbf{B}^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 < t^2\right)\\
&\ge 1 - \left(\textstyle\sum_{s=0}^q \frac{c^s}{s!}\exp(-c)\right)^{n_2}\\
&= 1 - \left(1 - \textstyle\sum_{s=q+1}^\infty \frac{c^s}{s!}\exp(-c)\right)^{n_2}\\
&\ge 1 - \left(1 - \frac{c^{q+1}}{(q+1)!}\exp(-c)\right)^{n_2} \qquad (22)
\end{aligned}$$

Choosing $c = \theta^{1/(q+1)}n^{-1/(q+1)}(q+1)$ and using that

$$(q+1)! > (q+1)^{q+1}$$

we obtain the following lower bound on (22)

$$1 - \left(\left(1 - \tfrac{\theta}{n}\exp(-c)\right)^n\right)^{1/2} \ge 1 - \exp\left(-\tfrac{\theta}{2}\exp(-c)\right)$$

as long as $n \ge \theta$. Setting $\theta = 8$, the above probability is lower bounded by .75 if $n > 8(q+1)^{q+1}$. Combining the choice of $t^2 = c2b^2$ with the above choice of c and noting that $\|\mathbf{B}^*\|_F^2 = 2(q+1)b^2$ concludes the proof.

References

[1] A. Abid, A. Poon, and J. Zou. Linear Regression with Shuffled Labels. arXiv:1705.01342, 2017.

[2] A. Agarwal, S. Negahban, and M. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40:1171–1197, 2012.

[3] Z. Bai and T. Hsing. The broken sample problem. *Probability Theory and Related Fields*, 131 (4):528–552, 2005.

[4] A. Balakhrisnan. On the problem of time jitter in sampling. *IRE Transactions on Information Theory*, 8:226–236, 1962.

[5] F. Bernhard. Fast Linear Assignment Problem using Auction Algorithm. mathworks.com.

[6] S. Blackman. *Multiple target tracking with radar applications*. Artech House Radar Library, 1986.

[7] R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems: Revised Reprint*. SIAM, 2009.

[8] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11, 2011.

[9] A. Carpentier and T. Schlüter. Learning relationships between data obtained independently. In *Proceedings of the International Conference on Artifical Intelligence and Statistics (AISTATS)*, pages 658–666, 2016.

[10] H.-P. Chan and W.-L. Loh. A file linkage problem of DeGroot and Goel revisited. *Statistica Sinica*, 11:1031–1045, 2001.

[11] P. Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.

[12] O. Collier and A. Dalalyan. Minimax Rates in Permutation Estimation for Feature Matching. *Journal of Machine Learning Research*, 17:1–31, 2016.

[13] N. Dalzell and J. Reiter. Regression Modeling and File Matching Using Possibly Erroneous Matching Variables. *Journal of Computational and Graphical Statistics*, 27:728–738, 2018.

[14] M. DeGroot and P. Goel. The Matching Problem for Multivariate Normal Data. *Sankhya, Series B*, 38:14–29, 1976.

[15] M. DeGroot and P. Goel. Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, 8:264–278, 1980.

[16] M. DeGroot, P. Feder, and P. Goel. Matchmaking. *The Annals of Mathematical Statistics*, 42: 578–593, 1971.

[17] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:2765–2781, 2013.

[18] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval. Compressed sensing with unknown sensor permutation. In *Acoustics, Speech and Signal Processing (ICASSP)*, pages 1040–1044, 2014.

[19] N. Flammarion, C. Mao, and P. Rigollet. Optimal Rates of Statistical Seriation. *Bernoulli*, 25:623–653, 2019.

[20] P. Goel. On Re-Pairing Observations in a Broken Sample. *The Annals of Statistics*, 3:1364–1369, 1975.

[21] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[22] W. Groves and M. Gini. Improving prediction in tac scm by integrating multivariate and temporal aspects via pls regression. In *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets*, pages 28–43. 2011.

[23] R. Gutman, C. Afendulis, and A. Zaslavsky. A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs. *Journal of the American Statistical Association*, 108(501):34–47, 2013.

[24] S. Haghighatshoar and G. Caire. Signal Recovery from Unlabeled Samples. In *International Symposium on Information Theory (ISIT)*, 2017.

[25] T. Herzog, F. Scheuren, and W. Winkler. *Data quality and record linkage techniques*. Springer, 2007.

[26] D. Hsu, K. Shi, and X. Sun. Linear regression without correspondence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1531–1540, 2017.

[27] P. Lahiri and Michael D. Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, 2005.

[28] R. Latala, P. Mankiewicz, K. Oleskiewicz, and N. Tomczak-Jaegermann. Banach-Mazur distances and projections on random subgaussian polytopes. *Discrete and Computational Geometry*, 38:29–50, 2007.

[29] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[30] K. Lounici, M. Pontil, A. Tsybakov, and S. van de Geer. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39: 2164–2204, 2011.

[31] E. Elhamifar M. Soltanolkotabi and E. Candes. Robust subspace clustering. *The Annals of Statistics*, 40:2195–2238, 2014.

[32] J. Neter, S. Maynes, and R. Ramanathan. The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, 60(312):1005–1027, 1965.

[33] A. Pananjady, M. Wainwright, and T. Cortade. Denoising Linear Models with Permuted Data. arXiv:1704.07461, 2017.

[34] A. Pananjady, M. Wainwright, and T. Cortade. Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, 3826–3300, 2018.

[35] P. Rigollet and J. Weed. Uncoupled isotonic regression via minimum Wasserstein deconvolution. arXiv:1806.10648, 2018.

[36] F. Scheuren and W. Winkler. Regression analysis of data files that are computer matched I. *Survey Methodology*, 19:39–58, 1993.

[37] F. Scheuren and W. Winkler. Regression analysis of data files that are computer matched II. *Survey Methodology*, 23:157–165, 12 1997.

[38] X. Shi, X. Lu, and T. Cai. Spherical regresion under mismatch corruption with application to automated knowledge translation. arXiv:1810.05679, 2018.

[39] M. Slawski and E. Ben-David. Linear Regression with Sparsely Permuted Data. *Electronic Journal of Statistics*, 1:1–36, 2019.

[40] M. Soltanolkotabi. MATLAB Code for Sparse Subspace Clustering. http://www-bcf.usc.edu/ soltanol/datacode.html.

[41] M. Soltanolkotabi and E. Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40:2195–2238, 2012.

[42] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 194:55–98, 2016.

[43] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12:2411–2414, 2011.

[44] J. Unnikrishnan and M. Vetterli. Sampling and reconstruction of spatial fields using mobile sensors. *IEEE Transactions on Signal Processing*, 61:2328–2340.

[45] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli. Unlabeled sensing with random linear measurements. *IEEE Transactions on Information Theory*, 64:3237–3253, 2018.

[46] R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28:52–68, 2011.

[47] Y.-X. Wang and H. Xu. Noisy sparse subspace clustering. *The Journal of Machine Learning Research*, 17(1):320–360, 2016.

[48] Y. N. Wu. A note on broken sample problem. Technical report, Department of Statistics, University of Michigan, 1998.

[49] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2496–2504, 2010.

[50] M. Yuan and Y. Lin. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 2006.

[51] H. Zhang, M. Slawski, and P. Li. Permutation Recovery from Multiple Measurement Vectors in Unlabeled Sensing. In *IEEE International Symposium on Information Theory (ISIT)*, 2019.