Journal of Machine Learning Research 21 (2020) 1-42 Submitted 7/19; Revised 6/20; Published 9/20

Two-Stage Approach to Multivariate Linear Regression with Sparsely Mismatched Data

Martin Slawski MSLAWSK3@GMU.EDU
George Mason University and Baidu Research
Emanuel Ben-David EMANUEL.BEN.DAVID@CENSUS.GOV
U.S. Census Bureau
Ping Li* LIPING11@BAIDU.COM
Baidu Research

Editor: Daniela Witten

Abstract

A tacit assumption in linear regression is that (response, predictor)-pairs correspond to identical observational units. A series of recent works have studied scenarios in which this assumption is violated under terms such as "Unlabeled Sensing and "Regression with Unknown Permutation". In this paper, we study the setup of multiple response variables and a notion of mismatches that generalizes permutations in order to allow for missing matches as well as for one-to-many matches. A two-stage method is proposed under the assumption that most pairs are correctly matched. In the first stage, the regression parameter is estimated by handling mismatches as contaminations, and subsequently the generalized permutation is estimated by a basic variant of matching. The approach is both computationally convenient and equipped with favorable statistical guarantees. Specifically, it is shown that the conditions for permutation recovery become considerably less stringent as the number of responses m per observation increase. Particularly, for $m = \Omega(\log n)$, the required signal-to-noise ratio no longer depends on the sample size n. Numerical results on synthetic and real data are presented to support the main findings of our analysis.

1. Introduction

Linear regression and its numerous extensions is an object of timeless interest in statistics and related disciplines. Continuous research efforts are being made to increase the range of situations in which it can be applied with success. A specific challenge that has attracted considerable interest recently is regression in the absence of correspondence between predictors and responses, i.e., both are given as separate samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^n$, but it is not (fully) known a priori which elements from $\mathcal{X}$ and $\mathcal{Y}$ are matching pairs in the sense of belonging to the same observational unit. Motivated by a number of applications in engineering, regression in this setting has been discussed in a series of recent papers (Emiya et al., 2014; Unnikrishnan et al., 2018; Pananjady et al., 2018; Abid et al.,

∗. Ping Li, corresponding author, Baidu Research - Bellevue, WA 98004.

©2020 Martin Slawski, Emanuel Ben-David, and Ping Li.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v21/19-645.html.

2017; Hsu et al., 2017; Haghighatshoar and Caire, 2017; Pananjady et al., 2017; Dokmanić, 2019; Shi et al., 2020; Tsakiris et al., 2020; Wang et al., 2018; Tsakiris and Peng, 2019). On the other hand, the above setup has a long history in statistics under the term "Broken Sample Problem" dating back to the early 1970s (DeGroot et al., 1971; Goel, 1975; DeGroot and Goel, 1976, 1980; Bai and Hsing, 2005; Wu, 1998; Chan and Loh, 2001) and a related line of research involving record linkage and statistical analysis based on merged data files (e.g., Neter et al. (1965); Lahiri and Larsen (2005); Goel and Ramalingam (2012); Scheuren and Winkler (1993, 1997)) partially motivated by government agencies like the U.S. Census Bureau that routinely combines data from multiple surveys and/or external data to address questions of interest. In this context, the primary interest is in the estimation of parameters (e.g., covariance matrix, regression coefficients, ...) rather than restoration of the correspondence between elements of $\mathcal{X}$ and $\mathcal{Y}$. Instead, the focus is on the adjustment of subsequent analyses for potential mismatches resulting from errors or ambiguities in record linkage based on quasi-identifiers. In fact, unique identifiers such as the social security number often need to be removed because of privacy concerns. Accordingly, in an alternative perspective on the broken sample problem, identification of matching pairs in $\mathcal{X}$ and $\mathcal{Y}$ is undesired because $\mathcal{Y}$ contains sensitive data, but an adversary makes the attempt to use external data along with identifying information stored in $\mathcal{X}$ to retrieve matching pieces in $\mathcal{Y}$. Well-known instances of such "linkage attacks" are the identification of the medical history of the former governor of Massachusetts (Sweeney, 2001) and the partial de-anonymization of Netflix movie rankings with the help of publicly available data in the Internet Movie Database (IMDb) (Narayanan and Shmatikov, 2008). Broken sample problems thus bear a relationship to data confidentiality; we refer to Domingo-Ferrer and Muralidhar (2016) for a detailed discussion.

Related Work. A starting point of recent research on the subject is the work by Unnikrishnan et al. (2018) which studies linear regression in the absence of noise with a scalar response that is observed up to an unknown permutation of the entries, i.e., $y_i = \mathbf{x}_{\pi\ (i)}^\top \beta^*$, $i = 1, \ldots, n$, for a permutation π^* on $\{1, \ldots, n\}$. The authors show that $\beta^* \in \mathsf{R}^d$ can be recovered with probability one by exhaustive enumeration over all permutations if $n \geq 2d$ and the entries of X are drawn i.i.d. from a distribution absolutely continuous w.r.t. the Lebesgue measure on R. Alternative proofs of this result have been obtained in Tsakiris (2018); Dokmanić (2019). Pananjady et al. (2018) study computational and statistical limits of recovering π^* for Gaussian $\{\mathbf{x}_i\}_{i=1}^n$ and Gaussian additive noise with variance σ^2. They show that least squares estimation recovers π^* exactly if the signal-to-noise ratio $\mathsf{SNR} = \|\beta^*\|_2^2/\sigma^2 = n^{\Omega(1)}$ which is also shown to be sharp up to a constant factor in the exponent. At the same time, least squares estimation of π^* is proved to be NP-hard. Abid et al. (2017); Hsu et al. (2017) shed light on the estimation of β^* under similar setups as in Pananjady et al. (2018). Specifically, Hsu et al. (2017) establish that the requirement $\mathsf{SNR} = \Omega(d/\log\log n)$ is necessary to ensure low relative squared ℓ_2-estimation error which is a dramatic gap compared to the requirement $\mathsf{SNR} = \Omega(d/n)$ if π^* is known. The paper (Abid and Zou, 2018) proposes Expectation-Maximization (EM) schemes to tackle the least squares problem for estimation of π^*. A clever initialization strategy for those schemes based on algebraic considerations is developed in Tsakiris et al. (2020). The paper (Slawski and Ben-David, 2019) assumes that π^* is k-sparse, i.e., $\pi^*(i) = i$ except for $k \ll n$ indices, and analyzes

a convex formulation for estimating in this setting. A similar sparsity assumption is employed in Shi et al. (2020) for spherical regression. Order-constrained regression problems with unknown permutation are discussed in Flammarion et al. (2019); Rigollet and Weed (2019); Carpentier and Schluter (2016); Ma et al. (2020).

Contributions. While several papers have elucidated important aspects of linear regression with unknown permutation for a scalar response, only few papers (Pananjady et al., 2017; Zhang et al., 2019a,b; Slawski et al., 2019; Zhang and Li, 2020) consider multivariate response, i.e., the y_i are m-dimensional, $m > 1$. This case is of independent interest for at least two reasons. First, in the context of record linkage it is natural to assume that both data sets X and Y to be merged are multi-dimensional. Second, the availability of multiple responses affected by the same permutation is expected to facilitate estimation as is confirmed by the results herein. Indeed, the requirements on the SNR to achieve permutation recovery can be considerably weaker, with potential drops from $\mathrm{SNR} = n^{(1)}$ for $m = O(1)$ to $\mathrm{SNR} = (1)$ for $m = (\log n)$. Similar benefits are shown in Pananjady et al. (2017); Zhang et al. (2019a); Slawski et al. (2019). The results in Pananjady et al. (2017) concern the prediction or denoising error rather than estimation of . Zhang et al. (2019a) provide information-theoretic lower bounds for permutation recovery; however, the computational scheme therein is only investigated empirically without theoretical support. The method in Slawski et al. (2019) requires $m \gg d$ to perform well; another downside of the approach is its cubic runtime in n. None of the aforementioned papers on the case $m > 1$ contain rigorous results regarding the estimation of the regression parameter. In order to enable the latter, the tolerable number of mismatches k herein is limited to a sufficiently small fraction of the number of samples, i.e. $k/n < c$ for c small enough. In this regime, estimation of the regression coefficients and restoration of the correct correspondence is shown to be possible based on convex optimization.

Moreover, we consider a more general notion of faulty correspondence between X and Y which goes beyond permutations, specifically allowing for missing matches and one-to-many matches. The effectiveness of the approach is demonstrated by experiments on synthetic and real data sets as well as a case study pertaining to data integration.

Outline. In §2, we state the problem and setting under consideration as well as the approach taken. Our main theoretical results are presented in §3. Empirical corroboration based on synthetic and real data is provided in §4. We conclude with a summary and an overview on potential directions of future research in §5.

Notation. The symbol $\mathbb{I}$ is used for the indicator function with value one if its argument is true and zero else. For a positive integer ℓ, I_ℓ denotes the $\ell \times \ell$ identity matrix, and $S^{\ell-1}$ denotes the unit sphere in $\mathbb{R}^\ell$. We write $|S|$ for the cardinality of a set S. The complement of S with respect to context-dependent base sets is denoted by S^c, and $\mathrm{conv}\, S$ denotes the convex hull of S. For a matrix A, $\|A\|_2 = \sigma_{\max}(A)$ denotes its spectral norm respectively maximum singular value, $\|A\|_F$ denotes its Frobenius norm, and $\mathrm{range}(A)$ denotes the column space of A. The i-th row of A is denoted by $A_{i,:}$, and is treated as column vector. For an index set I and a vector v of real numbers, v_I denotes the subvector corresponding to I. We write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Positive constants are denoted by

C, c, c_1 etc. We make use of the usual Big-O notation in terms of O, o, Ω and ω. We often use $a \lesssim b$, $b \gtrsim a$, and $a \asymp b$ as shortcuts for $a = O(b)$, $b = \Omega(a)$ and $a = \Theta(b)$, respectively.

2. Problem statement and proposed approach

We start by fixing the setup under consideration herein before outlining our approach. We then provide a toy data example in order to illustrate some of the main challenges and characteristics of the given problem and the proposed approach.

2.1 Setup

As stated in the introduction, we assume that we are given two samples $\mathcal{X} = \{x_i\}_{i=1}^n$ and $\mathcal{Y} = \{y_i\}_{i=1}^n$ taking values in $\mathbb{R}^d$ and $\mathbb{R}^m$, respectively, that are related by the model

$$s_i y_i = B^\top x_{\pi(i)} + \varepsilon_i, \quad 1 \le i \le n, \tag{1}$$

where $\pi : \{1, \ldots, n\} \to \{0, 1, \ldots, n\}$ is a map representing the (unknown) underlying correspondence between observations in $\mathcal{X}$ and $\mathcal{Y}$, with the convention that $x_0 := 0$, and $s_i = \mathbb{I}(\pi(i) \neq 0)$ indicates whether y_i has a match among $\mathcal{X}$, $1 \le i \le n$. For the set of non-matches $\mathcal{N} = \{i : s_i = 0\}$, we suppose that $\{y_i\}_{i \in \mathcal{N}}$ is independent of $\mathcal{X}$.

If $\pi(i) = i$ for $1 \le i \le n$, the above model reduces to an ordinary multivariate regression model with m responses and d predictor variables, regression coefficients $B \in \mathbb{R}^{d \times m}$, and random error variables $\{\varepsilon_i\}_{i=1}^n$. Model (1) can be expressed equivalently via

$$SY = \Pi XB + SE, \tag{2}$$

where Y and E are n-by-m matrices whose rows are given by $\{y_i^\top\}$ and $\{\varepsilon_i^\top\}$, respectively, $S = \mathrm{diag}(s_1, \ldots, s_n)$, X is an n-by-d matrix with rows $\{x_i^\top\}_{i=1}^n$, and $\Pi = (\pi_{ij})_{1 \le i,j \le n}$ has entries $\pi_{ij} = 1$ if $\pi(i) = j$ for $j \neq 0$, and zero otherwise. Observe that by construction, Π is contained in the following set of matrices

$$\mathcal{M} = \Big\{ \Pi \in \mathbb{R}^{n \times n} : \pi_{ij} \in \{0, 1\}, \ 1 \le i, j \le n, \ \sum_j \pi_{ij} \le 1, \ 1 \le i \le n \Big\} \tag{3}$$

$$\mathcal{P} = \{ \Pi \in \mathbb{R}^{n \times n} : \Pi^\top \Pi = I_n, \ \pi_{ij} \in \{0, 1\}, \ 1 \le i, j \le n \}, \tag{4}$$

which contains the set of n-by-n permutation matrices $\mathcal{P}$ in (4). Model (1) is hence more general compared to existing work in which Π is restricted to be a permutation. In particular, the generalization herein allows for missing matches via $\Pi_{i,:} = 0$ for $i \in \mathcal{N}$, as well as for one-to-many matches, i.e., more than one element in $\mathcal{Y}$ may correspond to the same element in $\mathcal{X}$; cf. Figure 1 for an illustration. We note that the case of one-to-many matches is also considered in Pananjady et al. (2017), cf. Section 2.4 therein.

Depending on the application, the goals in the setup (1) concern estimation of B and/or Π. If Π is recovered exactly by an estimator $\widehat{\Pi}$, i.e., the event $\{\widehat{\Pi} = \Pi\}$ occurs, estimation of B becomes an ordinary regression problem. In post-linkage data analysis, Π can be used to model error in the file linkage process, caused, e.g., by ambiguities resulting from the use of quasi-identifiers (say, the combination of age, gender, and race), but is typically treated as a nuisance parameter while primary interest concerns B. By contrast, in the setting of linkage attacks, the adversary aims at leveraging the linear relationship between

$$\{y_i\}_{i=1}^n \quad \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}; \quad S = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\{x_i\}_{i=1}^n$$

Figure 1: Illustration of the generalized permutation model herein for $n = 5$ including a missing match (y_3) and a one-to-many match between (y_1, y_5) and x_4.

elements of X and Y, and hence B is only regarded as a means to retrieve π. In the sequel, we adopt neither viewpoint and consider estimation of both B and π.

Assumptions. Below, we summarize and discuss the main assumptions of our analysis.

The map π is said to be k-sparse if $\pi(i) = i$ except for indices $S \subseteq \{1, \ldots, n\}$ with $|S| \le k$ for $k \le n$. Equivalently, $S = \{i : \pi_{ii} \neq 1\}$. Model (2) implies that

$$Y = XB + \Delta + SE, \tag{5}$$

where $\Delta_{i,:} = y_i - B^\top x_i$ if $\pi(i) = 0$ and $\Delta_{i,:} = B^\top x_{\pi(i)} - B^\top x_i$ otherwise, $1 \le i \le n$. Observe that k-sparsity of π implies that Δ has at most k non-zero rows. Throughout this paper, we shall impose constraints on the size of k. As of now, if $k > 0$ and k is not restricted, no practical estimation scheme with provable guarantees is known even if π is a permutation. Apart from that, the sparse regime is relevant to applications in record linkage as elaborated in detail in the case study in §4.

The matrix X has i.i.d. Gaussian rows $x_i \sim N(0, \Sigma)$, $1 \le i \le n$. Without loss of generality, we assume that $\Sigma = I_d$ as can be ensured by re-defining B accordingly.

Likewise, the matrix E has i.i.d. Gaussian rows $\epsilon_i \sim N(0, I_m)$, $1 \le i \le n$, and is independent of X.

The second assumption and the first part of the third assumption do not appear critical to our approach, but they considerably simplify results and proofs and thus aid presentation. The main results in this paper continue to hold for X and E with i.i.d. sub-Gaussian rows up to slight modifications, cf. Appendix F. Moreover, it is common to assume that the m entries of the noise terms $\{\epsilon_i\}_{i=1}^n$ are correlated; such extension can be accommodated, too.

Finally, we note that representation (5) is general enough to cover various other scenarios involving mismatched data in regression. For example, it also applies if a subset of the predictors is collected jointly with the response, i.e., we observe samples $D_1 = \{(x_i^{(1)}, y_i)\}_{i=1}^n$ and $D_2 = \{x_i^{(2)}\}_{i=1}^n$ with $\{x_i^{(1)}\}_{i=1}^n$ and $\{x_i^{(2)}\}_{i=1}^n$ having dimension d_1 and d_2, respectively, $d_1 + d_2 = d$, and associated regression model

$$y_i = B_{(1)}^\top x_i^{(1)} + B_{(2)}^\top x_{\pi(i)}^{(2)} + \epsilon_i, \; i = 1, \ldots, n; \tag{6}$$

where $\#$ is a permutation of $\{1,\dots,n\}$. Here, model (6) is subsumed by (5) by setting $B = \begin{pmatrix} B_{(1)} \\ B_{(2)} \end{pmatrix}$, $x_{i,:} = \begin{pmatrix} B_{(2)} x_{(i)}^{(2)} \\ B_{(2)} x_i^{(2)} \end{pmatrix}$, $1 \le i \le n$, and $S = I_n$. The approach and its analysis below applies to this and presumably also to other modifications with slight changes.

2.2 Approach

We suggest to tackle estimation of B and Π in a two-stage approach that we motivate as follows. Suppose first that there are no missing matches so that $\sum_j \Pi_{ij} = 1$, $1 \le i \le n$, and denote by $\overline{\mathcal{M}}$ the corresponding subset of $\mathcal{M}$ that excludes matrices with all-zero rows. Joint least squares estimation, i.e., $\min_{\Pi \in \overline{\mathcal{M}}, B \in \mathbb{R}^{d \times m}} \|Y - \Pi X B\|_F^2$, is NP-hard (Pananjady et al., 2018). However, if B is known, least squares estimation of Π reduces to a tractable optimization problem that decouples along the rows of Y:

$$\min_{\Pi \in \overline{\mathcal{M}}} \|Y - \Pi X B\|_F^2 = \sum_{i=1}^n \min_{1 \le j \le n} \|y_i - B^\top x_j\|_2^2 \; . \tag{7}$$

Assuming for simplicity that the minimizing indices $\hat{\jmath}(i)$ for the optimization problems inside the curly brackets are unique, we have $\hat{\Pi}_{i\hat{\jmath}(i)} = 1$, $1 \le i \le n$; all other entries of $\hat{\Pi}$ equal zero. If in addition Π is known to be one-to-one (i.e., a permutation), minimization over $\mathcal{M}$ can be replaced by minimization over $\mathcal{P}$ (4). The latter optimization problem reduces to a linear assignment problem (Burkard et al., 2009), a specific linear program that can be solved efficiently by specialized techniques such as the Hungarian Algorithm (Kuhn, 1955) or the Auction Algorithm (Bertsekas and Castanon, 1992).

In the case of missing matches, taking the minimum in (7) over $\mathcal{M}$ instead of over $\overline{\mathcal{M}}$ cannot be expected to ensure the successful identification of missing matches. In fact, a row of zeroes in Π means that the corresponding row of Y is paired with the zero vector rather than with any of the $\{B^\top x_j\}_{j=1}^n$, but the use of the zero vector as a reference for missing matches is not meaningful. This observation prompts the following modification of (8):

$$\text{Compute } \min_{1 \le j \le n} \|y_i - B^\top x_j\|_2^2 \text{; set } \hat{b}_{ij} = \begin{cases} 1 & \text{if } j = \hat{\jmath}(i) \text{ and } \|y_i - B^\top x_{\hat{\jmath}(i)}\|_2 \le \tau \\ 0 & \text{otherwise,} \end{cases} \quad 1 \le i,j \le n; \tag{8}$$

where $\{\hat{\jmath}(i)\}_{i=1}^n$ are the minimizing indices as above, and $\tau > 0$ is a suitably chosen threshold whose choice is discussed in Theorem 2 below.

So far, B was supposed to be known. If B is unknown, it has to be replaced by an estimator $\hat{B}$. At this point, our approach makes use of the sparsity assumption for β. In view of relation (5), we consider

$$\min_{B \in \mathbb{R}^{d \times m},\, \beta \in \mathbb{R}^m} \frac{1}{2n} \|Y - \Pi X B - \sqrt{n}\, \beta \mathbf{1}_n^\top\|_F^2 + \lambda \sum_{i=1}^n \|r_{i,:}\|_2, \tag{9}$$

for a tuning parameter $\lambda > 0$, where targets $\alpha := \frac{1}{n} \mathbf{1}_n^\top (Y - \Pi X B)^\top \mathbf{1}_n$ with Π as in (5), and $\|r_{i,:}\|_2$ being used as a convex surrogate for $\mathbb{1}(\|r_{i,:}\|_2 > 0)$, $1 \le i \le n$, in order to promote row-wise

sparsity of (Yuan and Lin, 2006; Eldar and Mishali, 2009; Lounici et al., 2011). The use of the re-scaled quantity in place of is done merely for technical reasons. We note that a variant of (9) for a single response variable has been employed in the context of linear regression with outliers (She and Owen, 2012; Laska et al., 2009; Nguyen and Tran, 2013).

Algorithm 1 Block coordinate descent for minimizing (9)

Compute the QR factorization $X = QR$ of X, and initialize $XB^{(0)} = QQ^{>}Y$, $^{(0)}$ 0.

1. Update for

$$^{(t+1)} \quad (1 \quad ^{(t)})\ ^{(t)} + \ ^{(t)}\text{GroupThreshold} \quad (Y \quad XB^{(t)};\ )=\sqrt{n}; \qquad := m \quad \sqrt{n} \quad ;$$

where for a matrix A with rows $\{a_i\}_{i=1}^n$ and 0, GroupThreshold $(A;\ )$ is defined by

$$a_i \quad a_i \quad (1 \quad = k a_i k_2)_+ \ ; \ i = 1; : : : ; n; \qquad (\ )_+ := \max\{\ ; 0\}:$$

2. Update for XB:

$$XB^{(t+1)} \quad (1 \quad ^{(t)})XB^{(t)} + \ ^{(t)}QQ^{>}(Y \quad \sqrt{n}\ ^{(t+1)}):$$

The step sizes $^{(t)}; \ ^{(t)}$ (0; 1) are chosen by back-tracking line search (Bertsekas, 1999).

Optimization problem (9) can be solved efficiently by block coordinate descent as outlined in Algorithm 1 that has performed extremely well throughout our experiments, typically converging after a small number of iterations. Formal convergence results follow immediately from the general framework in Tseng (2010).

The estimator $\hat{B}$ resulting from (9) can potentially be refined by a least squares refitting step after removing data corresponding to $\hat{S}(t) = \{1 \quad i \quad n : k\hat{b}_{i;:}k_2 \quad t\}$, where $\hat{b}$ denotes the minimizing in (9) and t is a suitably chosen threshold. The rationale is to remove mismatches as they hamper parameter estimation. This yields

$$\min_{B \in \mathbb{R}^{d \quad m}} \sum_{i \in \hat{S}(t)} k y_i \quad B^{>} x_i k_2^2: \tag{10}$$

In summary, this yields the following two-stage (or optionally three-stage) approach for estimating B and subsequently .

1. Estimate B from (9), and optionally refine via (10).
2. Estimate from (8) with B replaced by the estimator obtained in step 1.

It is worth pointing out that sparsity of is incorporated at step 1. only. The procedure (8) can be modified accordingly by applying it only for the indices corresponding to the k largest values among $\{k y_i \quad B^{>} x_i k_2^2\}_{1 \quad i \quad n}$, and setting $\hat{b}_{ii} = 1$ for all remaining i. We do not study this modification in the sequel since it does not fundamentally change the statistical limits in recovering as stated in Theorem 2 below.

Illustration. An illustration of the above approach is provided in Figure 2. The data set consists of monthly average temperatures of $n = 46$ U.S. cities as reported on Wikipedia

Jan	Mar	May	Jul	Sep	Nov	X	Y	Feb	Apr	Jun	Aug	Oct	Dec
16	33	59	74	62	34	Minneapolis	Memphis	46	63	80	82	64	44
-8	12	50	63	45	3	Fairbanks	San Antonio	56	70	83	85	71	53
1	54	72	83	75	53	Memphis	Fairbanks	-1	33	61	57	24	-4
34	44	64	78	68	47	Baltimore	Dallas	50	66	81	86	68	47
46	58	74	86	78	57	Dallas	Tampa	63	72	82	83	76	63
23	35	56	72	63	39	Milwaukee	Pittsburgh	31	51	69	72	53	33
61	67	78	83	82	69	Tampa	Minneapolis	21	48	69	71	49	20
29	40	60	73	64	43	Pittsburgh	Portland	44	52	64	70	55	40
52	62	77	85	80	61	San Antonio	Baltimore	36	54	73	76	57	37
41	48	58	69	65	47	Portland	Milwaukee	26	46	67	71	52	27

$\hat{S}$	Baltimore	Dallas	Fairbanks	Las Vegas^y	Memphis	Minneapolis
$\hat{b}(\hat{S})$	Milwaukee	Seattle	Fairbanks	Dallas	Baltimore	Minneapolis

continued:

$\hat{S}$	Phoenix	Portland	San Antonio	San Francisco^y	Seattle^y	Tampa
$\hat{b}(\hat{S})$	Las Vegas	Memphis	Phoenix	San Francisco	San Antonio	Tampa

Figure 2: Top: mismatched subset of the U.S. cities temperatures data set. Bottom: estimated subset of mismatched cities $\hat{S}$ and estimated correspondence $\hat{b}(\hat{S})$. Asterisked cities Milwaukee and Pittsburgh did not end up included in $\hat{S}$ since the mismatch resulting from shuffling happened not to be substantial enough. The superscript y refers to cities not affected by shuffling yet included in $\hat{S}$.

(2019). The data set is broken into two samples X and Y with the former containing the temperatures of the odd numbered months (January, March, ..., November) and the latter containing the temperatures of the even numbered months. For a random subset of $n = 10$ cities, we randomly permute matching records in X and Y. Linear regression is used to predict the $m = 6$ temperatures in Y from X. Due to high correlations among predictors, we work with the top $d = 3$ principal components as regressors. In the absence of partial data shuffling, this yields a reasonable goodness of fit overall in terms of a coefficient of determination $R^2 \approx 0.73$, apart from poor model fit for several west coast cities (Los Angeles, San Diego, Seattle and San Francisco) with mild winters and small seasonal differences, as well as for cities in desert regions (Las Vegas and Phoenix) with extreme temperatures during summer. After data shuffling, model fit drops to $R^2 \approx 0.4$. The approach outlined above shows some potential in this setting. With the choice of $\gamma = \frac{1}{3}\, b_0/\sqrt{n-m}$, where b_0 is the estimated error variance from the regression model in the absence of partial data shuffling, we ensure $R^2 \geq 0.62$. Subsequent restoration of the correct correspondence between X and Y is restricted to observations in $\hat{S} = \{i : |\hat{b}_{i,:}|_2 \geq \sqrt{2m}\, b_0\}$; for all other observations, no mismatches are assumed, i.e., $\hat{b}_{ii} = 1$, $i \notin \hat{S}$. The results highlight the challenges that are encountered in the estimation of γ. Most crucially, the more an observation is distinct from the rest, the easier it is identified as mismatch and the easier to retrieve its matching counterpart, with Fairbanks here being the most distinct instance. On the other hand, the temperature differences between Milwaukee and Pittsburgh are only marginal, and accordingly this mismatch remains undetected. Moreover, it is hard to disentangle cities affected by shuffling and poor fit of the linear model, respectively. Nevertheless, re-matching succeeds for three cities (Fairbanks, Minneapolis, Tampa) and gets close in case of Phoenix

Las Vegas and San Antonio! Phoenix.

Alternatives to (9). Formulation (9) treats mismatches in the same way as generic data contamination (outliers). A promising alternative approach if an upper bound on k is known and $m = 1$ can be found in Bhatia et al. (2017). A direct extension of this approach to the multiple response case with row-sparse contaminations is given by

$$\hat{B} \in \operatorname*{argmin}_{B \in \mathbb{R}^{d \times m}} \|Y - XB\|_F^2, \text{ where } \hat{e} \in \operatorname*{argmin}_{e \in \mathbb{R}^{n \times m}} \|P_X^{\perp}(Y - e)\|_F^2 \text{ subject to } \sum_{i=1}^{n} \mathbb{I}(\|e_{i,:}\|_2 > 0) \leq k, \tag{11}$$

where $P_X^{\perp}$ denotes the projection on the orthogonal complement of range(X). Following Bhatia et al. (2017), the rightmost optimization problem in (11) is tackled via iterative hard thresholding (Blumensath and Davies, 2009), and the result is substituted into the leftmost optimization problem to obtain an estimator for B. In our experiments, the performance of (11) is rather similar to that of the three-stage approach (10).

Given that both (9) and (11) treat mismatches as generic contaminations, it is worth exploring whether the additional structure under consideration here can be leveraged for improved performance. In the following, we present two approaches that are based on optimization over the polyhedron

$$C = \Big\{ \Omega \in \mathbb{R}^{n \times n} : \omega_{ij} \in [0, 1], \ 1 \leq i, j \leq n, \ \sum_j \omega_{ij} \leq 1, \ 1 \leq i \leq n \Big\}. \tag{12}$$

The first proposal can be seen as an immediate refinement of (9):

$$\min_{\Omega \in C} \frac{1}{2n - m} \|P_X^{\perp}(Y - \Omega Y)\|_F^2 + \sum_{i=1}^{n} \|Y^\top (I - \Omega)^\top e_i\|_2, \tag{13}$$

with $P_X^{\perp}$ as defined below (11) and $\{e_i\}_{i=1}^n$ denoting the canonical basis of $\mathbb{R}^n$. Similar to (9), the penalty in (13) is motivated by the fact that $(I - \Omega)Y$ has only few non-zero rows.

Given an upper bound on k, an alternative to (13) is given by the optimization problem

$$\min_{\Omega \in C} \frac{1}{2n - m} \|P_X^{\perp}(Y - \Omega Y)\|_F^2 \text{ subject to } \sum_{i=1}^{n} (1 - \omega_{ii}) \leq k. \tag{14}$$

Given a minimizer $\hat{\Omega}$ of (13) or (14), an estimate of B is obtained via least squares regression of $\hat{\Omega} Y$ on X. Both (13) and (14) are convex problems; (14) is a quadratic program. In spite of this, (13) and (14) have significant computational drawbacks compared to the approaches (9) and (11) since the former involve n^2 variables and thus scale poorly with problem size. According to own experiments, state-of-the art solvers for quadratic programs such as `cplexqp` in CPLEX[1] take prohibitively long to solve instances of (14) even for $n = 200$. In Appendix G, we present reasonably practical algorithms for obtaining approximate solutions of (13) and (14) based on the conditional gradient (aka Frank-Wolfe) method (Jaggi, 2013), which are also used in an empirical comparison with our primary proposal (9) in §4. In that comparison, neither (13) nor (14) achieves substantial improvements over (9).

1. http://www.ibm.com/us-en/marketplace/ibm-ilog-cplex

3. Main results

This section provides theoretical results on the approach introduced in the previous section. Theorem 1 quantifies the error in estimating B, while recovery of the correct correspondence in terms of Π is discussed in a separate subsection.

Theorem 1 Consider model (5) and the minimizer $(\hat{B}, \hat{\Pi})$ of (9) with $\lambda \geq \lambda_0$, where

$$\lambda_0 = \sigma \frac{\eta_{n,d}}{\sqrt{n}\sqrt{m}} \left(1 + \sqrt{\frac{4\log n}{m}}\right); \quad \eta_{n,d} := \frac{n-d}{n} + \sqrt{\frac{24\log n}{n}} \wedge 1; \tag{15}$$

and suppose $d/n < 1/4$. Then for any $\varepsilon \in (0, 1/3)$, there exist constants $c_\varepsilon, c_\varepsilon' > 0$ so that if $k \leq c_\varepsilon n/\log(n/k)$, it holds that

$$\frac{\|\hat{B} - \Pi^\top B\|_F}{\sqrt{m}} \leq 2\varepsilon^{-2} \sigma \sqrt{m} \left(\frac{\eta + \eta_0}{\eta_0}\right) \sqrt{k} \tag{16}$$

with probability at least $1 - 2/n - 3.5\exp(-c_\varepsilon' n)$. Furthermore,

$$\frac{\|\hat{B} - \Pi^\top B\|_F}{\sqrt{m}} \leq \frac{\|B\|_F}{\sqrt{m}} \left(1 - \sqrt{\frac{1}{4d\_\log\frac{n}{n}}}\right) \sqrt{\frac{5(d\_\log(n))}{n}} + \frac{\|\hat{B} - \Pi^\top B\|_F}{\sqrt{m}}$$

with probability at least $1 - 2\exp(-\frac{1}{2}(d\_\log n)) - \exp(-(d-m)\_\log(n-m))$.

In order to better understand the consequences of Theorem 1, we spell out essential scalings in (n, k, d, m) below. According to (15), the parameter λ should be chosen proportional to

$$\lambda_0 \asymp \sigma \frac{1}{\sqrt{n}\sqrt{m}} (1 + \sqrt{\log(n)/m}) \tag{17}$$

in which case $\frac{\|\hat{B} - \Pi^\top B\|_F}{\sqrt{m}} \lesssim \sigma \sqrt{\frac{k}{n}} (1 + \sqrt{\log(n)/m})$ which are familiar rates for multivariate regression with block sparsity regularization (Lounici et al., 2011). At the same time, the estimation error for the regression coefficients scales as $\frac{\|\hat{B} - \Pi^\top B\|_F}{\sqrt{m}} \lesssim \sqrt{d/n} + \frac{\|\hat{B} - \Pi^\top B\|_F}{\sqrt{m}}$, where the first term on the right hand side equals the estimation rate of least squares regression in the absence of mismatches while the second term reflects the slack arising from the presence of the latter. The bottom line is that the estimation error is in check as long as the fraction of mismatches k/n is small. In fact, the condition preceding (16) imposes a bound on that fraction as well. In experiments, performance degrades more noticeably once $k/n > 0.3$. Theorem 1 also indicates a positive influence of the number of response variables m in that one can choose $\lambda \asymp \frac{1}{\sqrt{n}\sqrt{m}}$ once $m \gtrsim \log n$ which in turn eliminates the factor $\sqrt{\log n}$ in (17) and thus also in (16). This is a known benefit of block sparsity regularization in comparison to element-wise sparsity regularization (Lounici et al., 2011).

Restoring Correspondence

In this subsection, we study recovery of Π. To begin with, we suppose that the regression parameter B is known, and establish one sufficient and one necessary condition for exact

recovery of based on the oracle estimator (8). A crucial quantity in the analysis is

$$\Delta^2 = \min_{i<j} \frac{\|B^\top (x_i - x_j)\|_2^2}{\|B\|_F^2}, \tag{18}$$

the minimum squared distance among all pairs of linear predictors scaled by $\|B\|_F^2$. A lower bound on Δ^2 is clearly needed in order to reliably match noisy responses $\{y_i\}_{i=1}^n$ to the corresponding elements in $\{B^\top x_i\}_{i=1}^n$: if there exists a pair (i, j) such that $\|B^\top (x_i - x_j)\|_2$ is smaller than the noise level, then there is a good chance that the corresponding responses get swapped. The following two lemmas provide upper and lower bounds in (18).

Lemma 1 Let $\mathrm{srank}(B) := \frac{\|B\|_F^2}{\|B\|_2^2}$ denote the stable rank of B, and consider Δ^2 as defined in (18). There exist universal constants $\delta_0 \in (0, 1)$ and such that for any $\varepsilon > 0$, with probability at least $1 - n^{-2\varepsilon}$, it holds that

$$\Delta^2 > \min\left\{ 2n^{-\frac{2(1+\varepsilon)}{\mathrm{srank}(B)}}, \delta_0 \right\}^2. \tag{19}$$

The stable rank of B as defined in the lemma crucially governs the scaling of . It is instructive to consider the extreme case $\mathrm{srank}(B) = 1$: we then obtain $\Delta^2 \gtrsim n^{-C}$ for $C > 0$. Results in Slawski and Ben-David (2019) on the case $n = 1$ show that $\Delta^2 \asymp n^{-2}$ with constant probability, which indicates sharpness of the above result in this case up to a constant in the exponent of n. On the other hand, if $\mathrm{srank}(B) = m \gtrsim \log n$, we have

$$2n^{-\frac{2(1+\varepsilon)}{\mathrm{srank}(B)}} = \exp\left(-\frac{2(1+\varepsilon)}{\mathrm{srank}(B)} \log(2n) \right) = \Theta(1),$$

i.e., the lower bound on Δ^2 does no longer decay with n. Additional insights can be obtained by considering the special case in which all non-zero singular values of B are equal to $b > 0$ and thus also $\mathrm{srank}(B) = \mathrm{rank}(B) = r$. For $r = 2(q+1)$, $q \geq 0$, the quantity (18) then becomes analytically tractable based on a closed form expression for χ^2-random variables with an even degrees of freedom.

Lemma 2 Consider Δ^2 as defined in (18) and suppose that B has exactly $r = 2(q+1)$, $q \in \{0, 1, \ldots\}$ non-zero singular values equal to $b > 0$. Then for all $\tau > 0$

$$\text{(Lower Bound):} \quad P\left(\Delta^2 \leq \frac{2}{e}(n\tau^2)^{\frac{2}{r}} \right) \leq 1 - \tau/2.$$

Moreover, if $n > 8(r/2)^{r/2}$,

$$\text{(Upper Bound):} \quad P\left(\Delta^2 \geq 2 \cdot 8^{2/r} n^{-2/r} \right) \geq 0.75.$$

Lemma 2 sheds some light on the range of the exponent in the previous Lemma 1, and provides essentially matching upper and lower bounds on Δ^2, where "essentially" refers to $n^{-4/r} \cdot \log n \cdot n^{-2/r}$, i.e., the match is up to constant factors and a factor 2 in the exponent.

In order to address the case of missing matches, we shall also consider

$$\sigma_0^2 = \min_{i \in N} \min_{1 \le j \le n} \|y_i - B^\top x_j\|_2^2 / \|B\|_F^2, \tag{20}$$

where we recall that $N = \{i : \pi(i) = 0\}$ denotes the set of missing matches. The quantity (20) exhibits scalings very similar to σ^2 (18) as discussed in the remark following Lemma B.1 in Appendix B.

Equipped with Lemma 1 & 2, we are in better position to interpret the following theorem.

Theorem 2 Let $\hat{B} = \hat{B}(X, Y)$ be an estimator of B, and let $\hat{\pi}(\hat{B}) = \{b_{ij}(\hat{B})\}$ denote the estimator (8) with $\lambda > \lambda_0 := (\sqrt{m} + 2\sqrt{\log n}) + \max_{1 \le j \le n} \|x_j\|_2 \|B - \hat{B}\|_2$ and B replaced by $\hat{B}$, i.e.,

$$b_{ij}(\hat{B}) = \begin{cases} 1, & \text{if } j = \hat{\pi}(i) \text{ and } \|y_i - \hat{B}^\top x_{\hat{\pi}(i)}\|_2 \le \lambda, \\ 0 & \text{otherwise,} \end{cases} \qquad 1 \le i, j \le n,$$

where the index $\hat{\pi}(i)$ is defined by $\|y_i - \hat{B}^\top x_{\hat{\pi}(i)}\|_2 = \min_{1 \le j \le n} \|y_i - \hat{B}^\top x_j\|_2$, $1 \le i \le n$. Let σ^2 and σ_0^2 be as in (18) and (20), respectively, and define the signal-to-noise ratio by $\mathrm{SNR} = \frac{\|B\|_F^2}{\sigma^2 m}$. Consider the event

$$\mathcal{B} = \left\{ \min\{\sigma_0^2, \sigma^2\}\,\mathrm{SNR} > 36 \max\left\{ \frac{\|\hat{B} - B\|_2^2}{\sigma^2 m} \max_{1 \le i \le n} \|x_i\|_2^2, 2\left(1 + \sqrt{\frac{4\log n}{m}}\right)^2 \right\}, \frac{\sigma^2}{\sigma^2 m} \right\}.$$

Conditional on B, with probability at least $1 - P(\mathcal{B}^c) - 1/n$, $\hat{\pi}(\hat{B}) = \pi$. Conversely, in the case that $\pi(i) \neq 0$ for $1 \le i \le n$, the following holds:

There exists $c > 0$ so that if $\mathrm{SNR} < c\frac{\log n}{m}$, $P(\hat{\pi}(\hat{B}) \neq \pi) \ge 1/3$.

If additionally $m = O(1)$, there exists $c' > 0$ so that if $\min\{\sigma_0^2, \sigma^2\}\mathrm{SNR} < c'$, $P(\hat{\pi}(\hat{B}) \neq \pi) \ge 1/3$.

The above theorem contains both an achievability result in the form of a sufficient condition for successful recovery of π given any estimator of B, as well as inachievability results concerning failure of recovery in the situation where B is known. As explained in more detail below, the above sufficient and necessary conditions agree up to multiplicative constants in certain regimes. To shed more light on the implications of the theorem, it is instructive to consider certain special cases of interest and to discuss them in connection with the error bounds stated in Theorem 1.

i) The conditions of Theorem 2 involve SNR as the ratio of the signal energy $\|B\|_F^2/m$ per response variable and noise variance σ^2. If $\hat{B} = B$ and every element of Y has match in X, the condition of the event $\mathcal{B}$ becomes

$$\min\{\sigma_0^2, \sigma^2\}\mathrm{SNR} \ge 2(1 + \sqrt{\log(n)/m})^2. \tag{21}$$

If $m = O(1)$, the scaling of σ^2 according Lemmas 1 and 2 imply that the condition $\mathrm{SNR} = \Omega(n^c)$ for a constant c depending on $\mathrm{srank}(\Pi)$ suffices for recovery of π.

ii) The second bullet in Theorem 2 implies that for $m = O(1)$, the condition $\mathrm{SNR} = \Omega(n^c)$ is also necessary (up to a constant factor in the exponent on n). In particular, Theorem 2 qualitatively recovers earlier results in Pananjady et al. (2018) and Slawski and Ben-David (2019) on $m = 1$.

iii) Regarding the scaling of m, the threshold case appears to be $m \asymp \log n \asymp \mathrm{srank}(B)$. In this regime, (21) requires only $\mathrm{SNR} = \Omega(1)$ which is a far less stringent condition compared to the regime of uniformly bounded m. Again, the sufficient condition is matched up to a constant multiplicative factor by the necessary condition stated in the first bullet of Theorem 2.

iv) Once m respectively $\mathrm{srank}(B)$ grow at a faster rate than $\log n$, the necessary condition of the first bullet is no longer aligned with (21). It remains an open question whether Theorem 2 can be sharpened in this regard.

We now discuss the situation in which B is replaced by an estimator $\widehat{B}$. In the absence of mismatches, random matrix theory (Vershynin and Rudelson, 2011) shows that ordinary least squares estimation obeys $\mathbb{E}[\|\widehat{B} - B\|_2^2/(\sigma^2 m)] \lesssim (d+m)/(n-m)$ while $\max_{1 \le i \le n} \|x_i\|_2^2 \lesssim d$ with high probability assuming that $d \gtrsim \log n$, which implies that the first term in the outer "max" of the event $\mathcal{B}$ is at best of the order $d^2/(n-m)$. A slightly less favorable condition is obtained when substituting the error bound of the proposed estimator in Theorem 1. In this case,

$$\|\widehat{B} - B\|_2^2/(\sigma^2 m) \le \|\widehat{B} - B\|_F^2/(\sigma^2 m) \lesssim (k+d)/n$$

with the stated probability, and thus Theorem 2 yields the condition $n \gtrsim d \vee (k \wedge d)$. In summary, the effect of replacing B by the proposed estimator can either be compensated by imposing a more stringent condition on SNR or the ratio d/n.

Lastly, let us comment on the case of missing matches, i.e., $N < n$, and the choice of λ. As long as λ is chosen proportional to the threshold τ_0, the requirements on the SNR remain qualitatively unchanged. The dependence of τ_0 on the noise level is intrinsic, hence approximate knowledge of σ is inevitable to guide the choice of λ. While τ_0 also depends on $\|\widehat{B} - B\|_2$, the latter can be estimated given bounds on the estimation error as discussed in the preceding paragraph. Clearly, λ can be set to zero whenever it is known that $N = n$.

Identification of Mismatched Data

In the following, we discuss a simpler task than recovery of π, namely recovery of $S = \{i : \pi(i) \neq i\}$, or equivalently, $S = \{i : \|\Delta_{i,:}\|_2 \neq 0\}$ with $\Delta = \Pi - P$ and $P = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ as defined in (5). The following statement provides a condition that ensures that we can separate mismatched data S and correctly matched data S^c in terms of $\{\|\widehat{b}_{i,:} - b_{i,:}\|_2\}_{i=1}^n$, where $\widehat{b}$ is obtained from optimization problem (9) and analyzed in Theorem 1.

Proposition 1 Let $\widehat{b}$ be as in Theorem 1, and let τ_0^2, σ^2, and SNR be as in Theorem 2. We then have $\min_{i \in S} \|\widehat{b}_{i,:} - b_{i,:}\|_2 > \max_{i \in S^c} \|\widehat{b}_{i,:} - b_{i,:}\|_2$ if

$$\min\{\tau_0^2, \sigma^2\}\mathrm{SNR} \gtrsim \frac{4 \max_{1 \le i \le n} \|\sqrt{n}(\widehat{b}_{i,:} - b_{i,:})\|_2^2}{\sigma^2 m}. \tag{22}$$

The practical consequences are as follows: if it holds that $\min_{i \in S} \|\hat{b}_{i,:}\|_2 > \max_{i \in S^c} \|\hat{b}_{i,:}\|_2$, we can sort the $\{\|\hat{b}_{i,:}\|_2\}_{i=1}^n$ and retain the observations corresponding to the $\lfloor n c \rfloor$ smallest elements for $c \in (0, (1 - k/n)]$. Any choice of $c = (1)$ in that range identifies $Q = S^c$ with $|Q| = (n)$. The least squares estimator $\tilde{B}$ of B using observations in Q only, i.e.,

$$\tilde{B} = \operatorname*{argmin}_{B \in \mathbb{R}^{d \times m}} \sum_{i \in Q} \|y_i - B^\top x_i\|_2^2$$

can substantially improve over the estimator $\hat{B}$ in Theorem 1. The condition of Proposition 1 tends to be easier to satisfy than that for recovery of π in Theorem 2. The right hand side of (22) is of the order $O(1 + \log(n)/m)$ and $O(\|f\|_1 + \log(n/m)\|g\|_\infty)$ in the best and worst case, respectively, in view of Theorem 1; the best case is obtained if $\max_i \|\hat{b}_{i,:} - b_{i,:}\|_2^2 \cdot k \approx \|\hat{b} - b\|_F^2$, i.e., the error in Frobenius norm is spread out roughly evenly over (k) rows.

4. Experiments

In the sequel, we present empirical evidence supporting central aspects of our analysis, and provide numerical comparisons to the alternative methods outlined at the end of §2 as well as to an extension of the EM scheme in Wu (1998); Abid and Zou (2018) for multiple response variables. For simplicity, we confine ourselves to the case in which π is a permutation matrix, i.e., an element of (4). Accordingly, the minimization in (7) is performed over the set of permutation matrices by means of the Auction Algorithm (Bertsekas and Castanon, 1992). We note that this modification does not affect our theoretical results. Specifically, the achievability result in Theorem 2 continues to hold because it asserts recovery over a superset of (4). Similarly, the inachievability results continue to hold if π is required to be a permutation.

Synthetic data

Setup Data is generated according to the model

$$y_i = B^\top x_{\pi(i)} + \varepsilon_i, \; i = 1, \ldots, n,$$

where the $\{x_i\}_{i=1}^n$ and $\{\varepsilon_i\}_{i=1}^n$, are i.i.d. from $N(0, I_d)$ and $N(0, I_m)$, respectively, π is a random permutation that shuffles $\{1, \ldots, k\}$ uniformly at random, and is the identity map when restricted to the remaining indices, i.e, $\pi(i) = i$ for $i > k$. The matrix B is obtained by first generating a d-by-d matrix (i.e., $d = m$) with i.i.d. $N(0, 1)$-entries, then computing its singular value decomposition $B = USV^\top$, and replacing the diagonal entries $\{s_1, \ldots, s_d\}$ of S according to $s_j \mapsto s_j^q$, $1 \le j \le d$ for $q \in \{0, 0.05, 0.1, 0.2, 0.5, 1, 2, 5\}$; finally, B is re-scaled such that $\|B\|_F^2 = m$. This construction ensures that the stable rank $\mathrm{srank}(B)$, which has a critical influence on the recovery of π, varies between $m = d$ (achieved for $q = 0$) and 1 (achieved for $q \to \infty$). In addition, the signal-to-noise ratio then results as $\mathrm{SNR} = \gamma^2$ with $\gamma \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2\}$. Lastly, the fraction of mismatches k/n varies between 0.05 and 0.4 in steps of 0.05 with $n \in \{200, 500, 1000\}$ and $d/n \in \{0.03, 0.06, 0.12\}$. For each configuration of (n, d, k, q, γ), 100 independent replications are performed.

In the experiments, the following approaches are compared:

naive, oracle. Plain least squares and estimation of B with knowledge of S, respectively.

proposed. B is estimated according to (9) with the choice $\lambda = \lambda^{\star} = 4\sqrt{\frac{1}{n}\frac{n+m}{m}}\sqrt{4\log(n)/m}$ which is the lower bound on λ suggested by Theorem 1 when treating ρ as 1.

proposed+. The re-fitting approach (10) building on proposed, cf. also Proposition 1. Assuming that k is known, the set of mismatches S is estimated by $\widehat{S} = \{1 \le i \le n : \|b_{i,:} - k_2\|_2 > t_{(n-k)}\}$, where $t_{(i)}, 1 \le i \le n$, denotes the i-th order statistic of the $\{\|b_{i,:} - k_2\|_2\}_{i=1}^n$.

CRR. "Consistent Robust Regression", following the title for the approach (11) used in Bhatia et al. (2017). The number of mismatches k is assumed to be known.

EM. The EM-scheme in Wu (1998); Abid and Zou (2018) in which Π is treated as missing data in conjunction with the use of the EM algorithm. Since the E-step involves intractable integration over the set of permutation matrices, MCMC is employed to approximate this step. In our implementation, the permutation is initialized as the identity, and the number of MCMC iterations per EM iteration is set to 10,000 given a "burn-in period" of 1,000.

DS-reg. The approach (13) that arises as a refinement of proposed, and here involves optimization over the set of doubly stochastic matrices of size n. We consider $\lambda = 2^p \lambda^{\star}$, $p \in \{-1, 0, \ldots, 3\}$, with $\lambda^{\star}$ as in the description of proposed above, and choose p by replication to minimize the estimation error w.r.t. $\|\cdot\|_F$ of the resulting estimator of B.

DS-cons. The approach (14) with k assumed to be known.

DS-reg+, DS-cons+. Re-fitting approaches associated with DS-reg and DS-cons. The set S is estimated by $\widehat{S} = \{1 \le i \le n : e_{ii} < t_{(n-k)}\}$, where e is the estimator of Π from (13) and (14), respectively, and $t_{(i)}, 1 \le i \le n$, denotes the i-th order statistic of $\{e_{ii}\}_{i=1}^n$.

Since solving the optimization problems associated with DS-reg and DS-cons entails substantial additional efforts even with customized solvers (Appendix G) given $O(n^2)$ variables, we only consider a reduced set of configurations (n, d, k, q, ρ) with $n \in \{200, 500\}$, $d/n = 0.03$, and $q = 0$, while the ranges for k/n and ρ remain unchanged. In addition, the number of replications per configuration is lowered to 20.

Results (I): Estimation of B. For better comparison across experimental configurations, we visualize the following "standardized" estimation error

$$\frac{1}{m}\|m^{1/2}\|\,\|B^{\text{est}} - B\|_F \sqrt{d/n}, \tag{23}$$

where B^{est} is a placeholder for the various estimators mentioned in the previous paragraph. Note that (23) approximately equals zero in expectation for the oracle estimator equipped with S, thus (23) can be interpreted as the excess error relative to that oracle. For the estimator $\widehat{B}$ analyzed in Theorem 1, the quantity (23) is expected to be proportional to $\sqrt{k/n}$. Selected results are shown in Figure 3, which displays averages of (23) for $n \in \{500, 1000\}$ and $\rho \in \{0.05, 0.1, 0.2\}$; the number of different values for k considered in a single plot had to be limited to ensure readability since for naive and EM, (23) still depends substantially

on . To account for that, shaded areas are used to represent the ranges of (23) for those two approaches; the upper and lower margins of the shaded areas represent the normalized estimation error for = 0:05 and = 0:2, respectively, while the dashed lines inside the shaded areas correspond to = 0:1. Accordingly, the performance of naive and EM (initialized by naive) relative to (23) improves, which is unsurprising given that as $\sqrt{\overline{m}} \% \|B\|_F$ (recall that $\|B\|_F = \sqrt{\overline{m}}$), the error induced by mismatches is of the same order as the noise in which case the gap between naive and oracle narrows. With the same reasoning, remedies for mismatches compared here are most effective if $\sqrt{\overline{m}}/\|B\|_F$ is small: for example, proposed achieves a roughly tenfold reduction in standardized estimation error over naive for = 0:05; that margin reduces gradually with increasing .

Figure 3: Average standardized estimation errors (23) on a $\log_{10}$-scale, with one curve for each $\in \{0.05, 0.1, 0.2\}$. For naive (in red) and EM (in green), the resulting curves do not cluster together, and are hence captured by the upper (= 0:05) and lower (= 0:2) boundaries of the shaded areas plus a dashed line (= 0:1).

Figure 3 also shows that refitting after applying proposed and estimating S considerably boosts performance. The performance of the resulting approach proposed+ is indistinguishable from CRR. While EM performs on par with the oracle for n = 500 (and n = 200, not shown), the approach degrades with n. One likely explanation is that the challenges associated with the E-step become more severe with m: specifically, the MCMC approximation tends to be less reliable for larger values of n. For the same reason, EM is at least an order of magnitude slower than proposed+ and CRR.

Figure 4: Average standardized estimation errors $\frac{1}{(d/(n-k))^{1/2}} m^{-1/2} \|B^{est} - B\|_F$ ($\log_2$-scale) of the refitting approach proposed+ (lines) and EM (shaded areas) for different rates of decay of the singular values of B corresponding to decreasing srank(B) from left to right. Curves for different combinations of n and appear in the same plots; due to poor clustering of those curves for EM in conjunction with the chosen error normalization, their range is indicated by shaded areas for better readability.

In Figure 4, the performance of proposed+ relative to EM is investigated in more detail. In addition to poor scalability with n, the competitiveness of EM also hinges on the stable rank of B not to be too small. The sequence of three plots in Figure 4 indicates a transition from superior to comparable and eventually not competitive performance of EM as the singular values in B decay more rapidly.

Finally, Figure 5 provides a comparison to the approaches DS-reg and DS-cons. Despite the additional sophistication involved, the results only indicate minor improvements, which largely disappear when considering refitting. In particular, the observed gains in performance do not appear to justify the massive computational effort associated with the solution of the optimization problems underlying DS-reg and DS-cons.

Figure 5: Average standardized estimation errors (23) of DS-reg and DS-cons in comparison to proposed along with their counterparts for refitting.