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.

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 $\beta^* \in \mathbb{R}^d$ 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 $\mathbb{R}$. 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 β^* 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 Schlüter (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 $\{\mathbf{y}_i\}_{i=1}^n$ 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 $\mathcal{X}$ and $\mathcal{Y}$ 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 $\mathsf{SNR} = n^{\Omega(1)}$ for $m = O(1)$ to $\mathsf{SNR} = \Omega(1)$ for $m = \Omega(\log n)$. 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 π^*. 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 $m \gtrsim d$ 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 $\mathcal{X}$ and $\mathcal{Y}$ 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 $\mathbb{S}^{\ell-1}$ 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 $\operatorname{conv} S$ 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 $\operatorname{range}(A)$ 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 Θ. 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 $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^n$ taking values in $\mathbb{R}^d$ and $\mathbb{R}^m$, respectively, that are related by the model

$$\mathfrak{s}_i \mathbf{y}_i = B^{*\top} \mathbf{x}_{\theta^*(i)} + \sigma \mathfrak{s}_i \boldsymbol{\epsilon}_i, \quad 1 \leq i \leq n, \tag{1}$$

where $\theta^* : \{1, \ldots, n\} \to \{0, 1, \ldots, n\}$ is a map representing the (unknown) underlying correspondence between observations in $\mathcal{X}$ and $\mathcal{Y}$, with the convention that $\mathbf{x}_0 := 0$, and $\mathfrak{s}_i = \mathbb{I}(\theta^*(i) \neq 0)$ indicates whether $\mathbf{y}_i$ has a match among $\mathcal{X}$, $1 \leq i \leq n$. For the set of non-matches $\mathcal{N} = \{i : \mathfrak{s}_i = 0\}$, we suppose that $\{\mathbf{y}_i\}_{i \in \mathcal{N}}$ is independent of $\mathcal{X}$.

If $\theta^*(i) = i$ for $1 \leq i \leq n$, 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 $\{\boldsymbol{\epsilon}_i\}_{i=1}^n$. Model (1) can be expressed equivalently via

$$\mathcal{S} Y = \Theta^* X B^* + \sigma \mathcal{S} E, \tag{2}$$

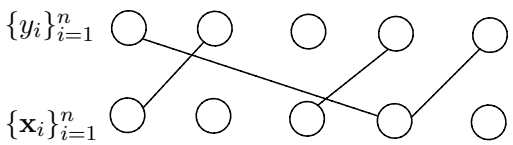
where Y and E are n-by-m matrices whose rows are given by $\{\mathbf{y}_i^\top\}$ and $\{\boldsymbol{\epsilon}_i^\top\}$, respectively, $\mathcal{S} = \mathrm{diag}(\mathfrak{s}_1, \ldots, \mathfrak{s}_n)$, X is an n-by-d matrix with rows $\{\mathbf{x}_i^\top\}_{i=1}^n$, and $\Theta^* = (\Theta^*_{ij})_{1 \leq i,j \leq n}$ has entries $\Theta^*_{ij} = 1$ if $\theta^*(i) = j$ for $j \neq 0$, and zero otherwise. Observe that by construction, Θ^* is contained in the following set of matrices

$$\mathcal{M} = \Big\{ \Theta \in \mathbb{R}^{n \times n} : \ \Theta_{ij} \in \{0, 1\}, \, 1 \leq i, j \leq n, \, \textstyle\sum_j \Theta_{ij} \leq 1, \, 1 \leq i \leq n \Big\} \tag{3}$$

$$\supset \mathcal{P} = \{ \Theta \in \mathbb{R}^{n \times n} : \ \Theta^\top \Theta = I_n, \, \Theta_{ij} \in \{0, 1\}, \, 1 \leq i, j \leq n \}, \tag{4}$$

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 θ^* is restricted to be a permutation. In particular, the generalization herein allows for missing matches via $\Theta^*_{i,:} = 0$ 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 Θ^*. If Θ^* is recovered exactly by an estimator $\widehat{\Theta}$, i.e., the event $\{\widehat{\Theta} = \Theta^*\}$ occurs, estimation of B^* becomes an ordinary regression problem. In post-linkage data analysis, Θ^* 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



$$\Theta^* = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \ \mathcal{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 1: Illustration of the generalized permutation model herein for $n = 5$ including a missing match ($\mathbf{y}_3$) and a one-to-many match between ($\mathbf{y}_1, \mathbf{y}_5$) and $\mathbf{x}_4$.

elements of $\mathcal{X}$ and $\mathcal{Y}$, and hence B^* is only regarded as a means to retrieve Θ^*. In the sequel, we adopt neither viewpoint and consider estimation of both B^* and Θ^*.

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

- The map θ^* is said to be k-sparse if $\theta^*(i) = i$ except for indices $S_* \subset \{1, \dots, n\}$ with $|S_*| \leq k$ for $k \ll n$. Equivalently, $S_* = \{i : \Theta^*_{ii} \neq 1\}$. Model (2) implies that

$$Y = XB^* + \Phi^* + \sigma \mathcal{S} E, \tag{5}$$

  where $\Phi^*_{i,:} = \mathbf{y}_i - B^{*\top}\mathbf{x}_i$ if $\theta^*(i) = 0$ and $\Phi^*_{i,:} = B^{*\top}\mathbf{x}_{\theta^*(i)} - B^{*\top}\mathbf{x}_i$ otherwise, $1 \leq i \leq n$. Observe that k-sparsity of θ^* implies that Φ^* has at most k non-zero rows. Throughout this paper, we shall impose constraints on the size of k. As of now, if $\sigma > 0$ and k is not restricted, no practical estimation scheme with provable guarantees is known even if θ^* 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 $\mathbf{x}_i \sim N(0, \Sigma)$, $1 \leq i \leq n$. 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 $\boldsymbol{\epsilon}_i \sim N(0, I_m)$, $1 \leq i \leq n$, 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 $\{\boldsymbol{\epsilon}_i\}_{i=1}^n$ 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 $\mathcal{D}_1 = \{(\mathbf{x}_i^{(1)}, \mathbf{y}_i)\}_{i=1}^n$ and $\mathcal{D}_2 = \{\mathbf{x}_i^{(2)}\}_{i=1}^n$ with $\{\mathbf{x}_i^{(1)}\}_{i=1}^n$ and $\{\mathbf{x}_i^{(2)}\}_{i=1}^n$ having dimension d_1 and d_2, respectively, $d_1 + d_2 = d$, and associated regression model

$$\mathbf{y}_i = B^{*\top}_{(1)}\mathbf{x}_i^{(1)} + B^{*\top}_{(2)}\mathbf{x}^{(2)}_{\theta^*(i)} + \sigma\boldsymbol{\epsilon}_i, \ i = 1, \dots, n, \tag{6}$$

where θ^* is a permutation of $\{1,\ldots,n\}$. Here, model (6) is subsumed by (5) by setting $B^* = \begin{bmatrix} B^*_{(1)} \\ B^*_{(2)} \end{bmatrix}$, $\Phi^*_{i,:} = B^{*\top}_{(2)}\mathbf{x}^{(2)}_{\theta^*(i)} - B^{*\top}_{(2)}\mathbf{x}^{(2)}_i$, $1 \le i \le n$, and $\mathcal{S} = I_n$. 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 Θ^* in a two-stage approach that we motivate as follows. Suppose first that there are no missing matches so that $\sum_j \Theta^*_{ij} = 1$, $1 \le i \le n$, 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., $\min_{\Theta\in\overline{\mathcal{M}},\, B\in\mathbb{R}^{d\times m}} \|Y - \Theta X B\|_F^2$, is NP-hard (Pananjady et al., 2018). However, if B^* is known, least squares estimation of Θ^* reduces to a tractable optimization problem that decouples along the rows of Y:

$$\min_{\Theta\in\overline{\mathcal{M}}} \|Y - \Theta X B^*\|_F^2 = \sum_{i=1}^n \left\{ \min_{1\le j\le n} \|\mathbf{y}_i - B^{*\top}\mathbf{x}_j\|_2^2 \right\}. \tag{7}$$

Assuming for simplicity that the minimizing indices $\widehat{j}(i)$ for the optimization problems inside the curly brackets are unique, we have $\widehat{\Theta}_{i\widehat{j}(i)} = 1$, $1 \le i \le n$; all other entries of $\widehat{\Theta}$ equal zero. If in addition θ^* 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 Θ means that the corresponding row of Y is paired with the zero vector rather than with any of the $\{B^{*\top}\mathbf{x}_j\}_{j=1}^n$, but the use of the zero vector as a reference for missing matches is not meaningful. This observation prompts the following modification of (8):

$$\text{Compute } \min_{1\le j\le n} \|\mathbf{y}_i - B^{*\top}\mathbf{x}_j\|_2^2 \text{; set } \widehat{\Theta}_{ij} = \begin{cases} 1 & \text{if } j = \widehat{j}(i) \text{ and } \|\mathbf{y}_i - B^{*\top}\mathbf{x}_{\widehat{j}(i)}\|_2 \le \tau, \\ 0 & \text{otherwise,} \end{cases} \quad 1 \le i,j \le n, \tag{8}$$

where $\{\widehat{j}(i)\}_{i=1}^n$ 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 $\widehat{B}$. At this point, our approach makes use of the sparsity assumption for θ^*. In view of relation (5), we consider

$$\min_{B\in\mathbb{R}^{d\times m},\, \Xi\in\mathbb{R}^{n\times m}} \frac{1}{2n\cdot m}\|Y - XB - \sqrt{n}\Xi\|_F^2 + \lambda\sum_{i=1}^n \|\Xi_{i,:}\|_2, \tag{9}$$

for a tuning parameter $\lambda > 0$, where Ξ targets $\Xi^* := \Phi^*/\sqrt{n}$ with Φ^* as in (5), and $\|\Xi_{i,:}\|_2$ being used as a convex surrogate for $\mathbb{I}(\|\Xi_{i,:}\|_2 > 0)$, $1 \le i \le n$, in order to promote row-wise

sparsity of Ξ (Yuan and Lin, 2006; Eldar and Mishali, 2009; Lounici et al., 2011). The use of the re-scaled quantity Ξ^* in place of Φ^* 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 $XB^{(0)} = QQ^\top Y$, $\Xi^{(0)} \equiv 0$.

1. Update for Ξ

$$\Xi^{(t+1)} \leftarrow (1-\alpha^{(t)})\Xi^{(t)} + \alpha^{(t)}\textsc{GroupThreshold}(Y - XB^{(t)}, \tau)/\sqrt{n}, \qquad \tau := m \cdot \sqrt{n} \cdot \lambda,$$

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

$$a_i \leftarrow a_i \cdot (1 - \eta/\|a_i\|_2)_+ \,,\ i = 1, \ldots, n, \qquad (\cdot)_+ := \max\{\cdot, 0\}.$$

2. Update for XB:

$$XB^{(t+1)} \leftarrow (1-\gamma^{(t)})XB^{(t)} + \gamma^{(t)}QQ^\top(Y - \sqrt{n}\Xi^{(t+1)}).$$

The step sizes $\alpha^{(t)}, \gamma^{(t)} \subset (0,1)$ 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 $\widehat{B}$ resulting from (9) can potentially be refined by a least squares refitting step after removing data corresponding to $\widehat{S}(t) = \{1 \leq i \leq n : \|\widehat{\Xi}_{i,:}\|_2 \geq t\}$, where $\widehat{\Xi}$ denotes the minimizing Ξ in (9) and t is a suitably chosen threshold. The rationale is to remove mismatches as they hamper parameter estimation. This yields

$$\min_{B \in \mathbb{R}^{d \times m}} \sum_{i \notin \widehat{S}(t)} \|\mathbf{y}_i - B^\top \mathbf{x}_i\|_2^2. \tag{10}$$

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

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

It is worth pointing out that sparsity of Θ^* 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 $\{\|\mathbf{y}_i - B^{*\top}\mathbf{x}_i\|_2^2\}_{1 \leq i \leq n}$, and setting $\widehat{\Theta}_{ii} = 1$ for all remaining i. We do not study this modification in the sequel since it does not fundamentally change the statistical limits in recovering Θ^* 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	$\mathcal{X}$	$\mathcal{Y}$	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

$\widehat{S}$	Baltimore	Dallas	Fairbanks	Las Vegas†	Memphis	Minneapolis
$\widehat{\theta}(\widehat{S})$	Milwaukee	Seattle	Fairbanks	Dallas	Baltimore	Minneapolis

continued:

$\widehat{S}$	Phoenix	Portland	San Antonio	San Francisco†	Seattle†	Tampa
$\widehat{\theta}(\widehat{S})$	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 $\widehat{S}$ and estimated correspondence $\widehat{\theta}(\widehat{S})$. Asterisked cities Milwaukee and Pittsburgh did not end up included in $\widehat{S}$ since the misfit resulting from shuffling happened not to be substantial enough. The superscript † refers to cities not affected by shuffling yet included in $\widehat{S}$.

(2019). The data set is broken into two samples $\mathcal{X}$ and $\mathcal{Y}$ 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 $k = 10$ cities, we randomly permute matching records in $\mathcal{X}$ and $\mathcal{Y}$. Linear regression is used to predict the $m = 6$ temperatures in $\mathcal{Y}$ from $\mathcal{X}$. 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 $\lambda = \frac{1}{3} \cdot \widehat{\sigma}_0 / \sqrt{n \cdot m}$, where $\widehat{\sigma}_0$ is the estimated error variance from the regression model in the absence of partial data shuffling, we ensure $R^2 \approx 0.62$. Subsequent restoration of the correct correspondence between $\mathcal{X}$ and $\mathcal{Y}$ is restricted to observations in $\widehat{S} = \{i : \ \|\widehat{\Xi}_{i,:}\|_2 \geq \sqrt{2m}\widehat{\sigma}_0\}$; for all other observations, no mismatches are assumed, i.e., $\widehat{\Theta}_{ii} = 1$, $i \notin \widehat{S}$. The results highlight the challenges that are encountered in the estimation of Θ^*. 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 $\rightarrow$

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

$$\widetilde{B} \in \operatorname*{argmin}_{B\in\mathbb{R}^{d\times m}} \|Y - \widetilde{\Phi} - XB\|_F^2, \text{ where } \widetilde{\Phi} \in \operatorname*{argmin}_{\Phi\in\mathbb{R}^{n\times m}} \|\mathsf{P}_X^{\perp}(Y-\Phi)\|_F^2 \text{ subject to } \sum_{i=1}^{n} I(\Phi_{i,:} \neq \mathbf{0}) \leq k, \tag{11}$$

where $\mathsf{P}_X^{\perp}$ 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

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

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

$$\min_{\Theta\in\mathcal{C}} \frac{1}{2n\cdot m}\|\mathsf{P}_X^{\perp}\Theta Y\|_F^2 + \lambda \textstyle\sum_{i=1}^n \|Y^\top(I - \ \Theta)^\top e_i\|_2, \tag{13}$$

with $\mathsf{P}_X^{\perp}$ 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 $(I - \Theta^*)Y$ has only few non-zero rows.

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

$$\min_{\Theta\in\mathcal{C}} \frac{1}{2n\cdot m}\|\mathsf{P}_X^{\perp}\Theta Y\|_F^2 \text{ subject to } \textstyle\sum_{i=1}^n \Theta_{ii} \geq n - k. \tag{14}$$

Given a minimizer $\widetilde{\Theta}$ of (13) or (14), an estimate of B^* is obtained via least squares regression of $\widetilde{\Theta}Y$ 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 Θ^* is discussed in a separate subsection.

Theorem 1 *Consider model* (5) *and the minimizer* $(\widehat{B}, \widehat{\Xi})$ *of* (9) *with* $\lambda \geq 2\lambda_0$, *where*

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

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*

$$\frac{\|\widehat{\Xi} - \Xi^*\|_F}{\sqrt{m}} \leq 2\varepsilon^{-2} \cdot \lambda \sqrt{m} \cdot \frac{\lambda + \lambda_0}{\lambda - \lambda_0} \sqrt{k} \tag{16}$$

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

$$\frac{\|\widehat{B} - B^*\|_F}{\sqrt{m}} \leq \frac{1}{1 - \sqrt{\frac{4d \vee \log n}{n}}} \left(\sigma \sqrt{\frac{5(d \vee \log(n))}{n}} + \frac{\|\widehat{\Xi} - \Xi^*\|_F}{\sqrt{m}}\right)$$

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

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

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

in which case $\frac{\|\widehat{\Xi} - \Xi^*\|_F}{\sqrt{m}} \lesssim \sqrt{\frac{k}{n}}(1 + \sqrt{\log(n)/m})$ 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 $\frac{\|\widehat{B} - B^*\|_F}{\sqrt{m}} \lesssim \sqrt{d/n} + \frac{\|\widehat{\Xi} - \Xi^*\|_F}{\sqrt{m}}$, 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 $\lambda \asymp \frac{1}{\sqrt{n \cdot m}}$ 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 Θ^*. To begin with, we suppose that the regression parameter B^* is known, and establish one sufficient and one necessary condition for exact

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

$$\gamma^2 = \min_{i<j} \frac{\|B^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2}{\|B^*\|_F^2}, \tag{18}$$

the minimum squared distance among all pairs of linear predictors scaled by $\|B^*\|_F^2$. A lower bound on γ^2 is clearly needed in order to reliably match noisy responses $\{\mathbf{y}_i\}_{i=1}^n$ to the corresponding elements in $\{B^{*\top}\mathbf{x}_i\}_{i=1}^n$: if there exists a pair (i,j) such that $\|B^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2$ 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* $\operatorname{srank}(B^*) := \frac{\|B^*\|_F^2}{\|B^*\|_2^2}$ *denote the stable rank of* B^**, and consider* γ^2 *as defined in* (18)*. There exist universal constants* $\alpha_0 \in (0,1)$ *and* κ *such that for any* $\varepsilon > 0$*, with probability at least* $1 - n^{-2\varepsilon}$*, it holds that*

$$\gamma^2 > \min\left\{2n^{\frac{-2(1+\varepsilon)}{\kappa \cdot \operatorname{srank}(B^*)}}, \alpha_0\right\}^2. \tag{19}$$

The stable rank of B^* as defined in the lemma crucially governs the scaling of γ. It is instructive to consider the extreme case $\operatorname{srank}(B^*) = 1$: we then obtain $\gamma^2 \gtrsim n^{-C}$ for $C > 0$. Results in Slawski and Ben-David (2019) on the case $m = 1$ show that $\gamma^2 \lesssim n^{-2}$ 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 $\operatorname{srank}(B^*) = m \gtrsim \log n$, we have

$$2n^{\frac{-2(1+\varepsilon)}{\kappa \cdot \operatorname{srank}(B^*)}} = \exp\left(-\frac{2(1+\varepsilon)}{\kappa \cdot \operatorname{srank}(B^*)} \log(2n)\right) = \Omega(1),$$

i.e., the lower bound on γ^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 $\operatorname{srank}(B^*) = \operatorname{rank}(B^*) = r$. 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* γ^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* $\delta > 0$

$$\textit{(Lower Bound):} \quad \mathbf{P}\left(\gamma^2 \geq \frac{2}{e}(n^{-2}\,\delta)^{\frac{2}{r}}\right) \geq 1 - \delta/2.$$

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

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

Lemma 2 sheds some light on the range of the exponent κ in the previous Lemma 1, and provides essentially matching upper and lower bounds on γ^2, where "essentially" refers to $n^{-4/r} \lesssim \gamma \lesssim n^{-2/r}$, 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

$$\gamma_0^2 = \min_{\substack{i \in \mathcal{N} \\ 1 \le j \le n}} \|\mathbf{y}_i - B^{*\top}\mathbf{x}_j\|_2^2 / \|B^*\|_F^2, \tag{20}$$

where we recall that $\mathcal{N} = \{i : \theta^*(i) = 0\}$ denotes the set of missing matches. The quantity (20) exhibits scalings very similar to γ^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 $\widehat{B} = \widehat{B}(X,Y)$ be an estimator of B^*, and let $\widehat{\Theta}(\widehat{B}) = \big(\widehat{\Theta}_{ij}(\widehat{B})\big)$ denote the estimator (8) with $\tau > \tau_0 := \sigma(\sqrt{m} + 2\sqrt{\log n}) + \max_{1\le j\le n}\|\mathbf{x}_j\|_2 \|B^* - \widehat{B}\|_2$ and B^* replaced by $\widehat{B}$, i.e.,*

$$\widehat{\Theta}_{ij}(\widehat{B}) = \begin{cases} 1, & \text{if } j = \widehat{j}(i) \text{ and } \|\mathbf{y}_i - \widehat{B}^\top \mathbf{x}_{\widehat{j}(i)}\|_2 \le \tau, \\ 0 & \text{otherwise,} \end{cases} \qquad 1 \le i, j \le n,$$

where the index $\widehat{j}(i)$ is defined by $\|\mathbf{y}_i - \widehat{B}^\top \mathbf{x}_{\widehat{j}(i)}\|_2 = \min_{1\le j\le n}\|\mathbf{y}_i - \widehat{B}^\top \mathbf{x}_j\|_2$, $1 \le i \le n$. Let γ^2 and γ_0^2 be as in (18) and (20), respectively, and define the signal-to-noise ratio by $\mathsf{SNR} = \frac{\|B^*\|_F^2}{\sigma^2 m}$. Consider the event

$$\mathcal{B} = \left\{ \min\{\gamma_0^2, \gamma^2\}\,\mathsf{SNR} > 36 \max\left\{ \frac{\|\widehat{B} - B^*\|_2^2}{\sigma^2 m} \max_{1\le i\le n}\|\mathbf{x}_i\|_2^2, 2\left(1 + \sqrt{\frac{4\log n}{m}}\right)^2, \frac{\tau^2}{\sigma^2 m}\right\}\right\}.$$

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

- *There exists $c > 0$ so that if $\mathsf{SNR} < c\frac{\log n}{m}$, $\mathbf{P}(\widehat{\Theta}(B^*) \neq \Theta^*) \ge 1/3$.*
- *If additionally $m = O(1)$, there exists $c' > 0$ so that if $\min\{\gamma_0^2, \gamma^2\}\mathsf{SNR} < c'$, $\mathbf{P}(\widehat{\Theta}(B^*) \neq \Theta^*) \ge 1/3$.*

The above theorem contains both an achievability result in the form of a sufficient condition for successful recovery of Θ^* given any estimator of $\widehat{B}$, 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 $\widehat{B} = B^*$ and every element of $\mathcal{Y}$ has match in $\mathcal{X}$, the condition of the event $\mathcal{B}$ becomes

$$\min\{\gamma_0^2, \gamma^2\}\mathsf{SNR} \ge 2(1 + \sqrt{\log(n)/m})^2. \tag{21}$$

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

ii) The second bullet in Theorem 2 implies that for $m = O(1)$, the condition $\mathsf{SNR} = \Omega(n^c)$ is also necessary (up to a constant factor in the exponent of 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 $\mathsf{SNR} = \Omega(1)$ 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 $\mathbf{E}[\|\widehat{B} - B^*\|_2^2/(\sigma^2 m)] \lesssim (d+m)/(n \cdot m)$ while $\max_{1\le i\le n}\|\mathbf{x}_i\|_2^2 \lesssim d$ 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 $d^2/(n\cdot m)$. 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 $n \gtrsim d \cdot (k \vee d)$. 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., $\mathcal{N} \neq \emptyset$, and the choice of τ. As long as τ 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 τ. 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, τ can be set to zero whenever it is known that $\mathcal{N} = \emptyset$.

Identification of Mismatched Data

In the following, we discuss a simpler task than recovery of Θ^*, namely recovery of $S_* = \{1 \le i \le n :\ \theta^*(i) \neq i\}$, or equivalently, $S_* = \{1 \le i \le n :\ \Xi^*_{i,:} \neq 0\}$ with $\Xi^* = \Phi^*/\sqrt{n}$ 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 $\{\|\widehat{\Xi}_{i,:}\|_2\}_{i=1}^n$, where $\widehat{\Xi}$ is obtained from optimization problem (9) and analyzed in Theorem 1.

Proposition 1 *Let $\widehat{\Xi}$ be as in Theorem 1, and let γ_0^2, γ^2, and SNR be as in Theorem 2. We then have $\min_{i\in S_*}\|\widehat{\Xi}_{i,:}\|_2 > \max_{i\in S_*^c}\|\widehat{\Xi}_{i,:}\|_2$ if*

$$\min\{\gamma_0^2, \gamma^2\}\mathsf{SNR} \ge \frac{4\max_{1\le i\le n}\|\sqrt{n}(\widehat{\Xi}_{i,:} - \Xi^*_{i,:})\|_2^2}{\sigma^2 m}. \tag{22}$$

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

$$\widetilde{B} = \operatorname*{argmin}_{B\in\mathbb{R}^{d\times m}} \sum_{i\in Q} \|\mathbf{y}_i - B^\top \mathbf{x}_i\|_2^2$$

can substantially improve over the estimator $\widehat{B}$ in Theorem 1. The condition of Proposition 1 tends to be easier to satisfy than that for recovery of Θ^* in Theorem 2. The right hand side of (22) is of the order $O(1+\log(n)/m)$ and $O(k\{1+\log(n/m)\})$ in the best and worst case, respectively, in view of Theorem 1; the best case is obtained if $\max_i \|\widehat{\Xi}_{i,:} - \Xi^*_{i,:}\|_2^2 \lesssim \|\widehat{\Xi} - \Xi^*\|_F^2/k$, i.e., the error in Frobenius norm is spread out roughly evenly over $\Omega(k)$ 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 Θ^* 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 Θ^* is required to be a permutation.

Synthetic data

Setup. Data is generated according to the model

$$\mathbf{y}_i = B^{*\top}\mathbf{x}_{\theta^*(i)} + \sigma\boldsymbol{\epsilon}_i,\ i = 1,\dots,n,$$

where the $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\boldsymbol{\epsilon}_i\}_{i=1}^n$, are i.i.d. from $N(0, I_d)$ and $N(0, I_m)$, respectively, θ^* is a random permutation that shuffles $\{1,\dots,k\}$ uniformly at random, and is the identity map when restricted to the remaining indices, i.e, $\theta^*(i) = i$ 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 $\{s_1,\dots,s_d\}$ of S according to $s_j \leftarrow j^{-q}$, $1 \le j \le d$ 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 $\operatorname{srank}(B^*)$, which has a critical influence on the recovery of Θ^*, 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 $\mathsf{SNR} = \sigma^{-2}$ with $\sigma \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2\}$. 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, σ), 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 Θ^*, respectively.

proposed. B^* is estimated according to (9) with the choice $\lambda = \lambda^\star = 4\sigma\frac{1}{\sqrt{n\cdot m}}$ which is the lower bound on λ suggested by Theorem 1 when treating $\sqrt{4\log(n)/m}$ simply 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 $\widehat{S} = \{1 \leq i \leq n : \|\widehat{\Xi}_{i,:}\|_2 > t_{(n-k)}\}$, where $t_{(i)}$, $1 \leq i \leq n$, denotes the i-th order statistic of the $\{\|\widehat{\Xi}_{i,:}\|_2\}_{i=1}^n$.

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 Θ^* 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 $\lambda \in 2^{-p}\lambda^\star$, $p \in \{-1, 0, \ldots, 3\}$, with $\lambda^\star$ as in the description of **proposed** above, and choose p replication 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 $\widetilde{S} = \{1 \leq i \leq n : \widetilde{\Theta}_{ii} < \widetilde{t}_{(n-k)}\}$, where $\widetilde{\Theta}$ is the estimator of Θ^* from (13) and (14), respectively, and $\widetilde{t}_{(i)}$, $1 \leq i \leq n$, denotes the i-th order statistic of $\{\widetilde{\Theta}_{ii}\}_{i=1}^n$.

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 for (n, d, k, q, σ) with $n \in \{200, 500\}$, $d/n = 0.03$, and $q = 0$, while the ranges for k/n and σ 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

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

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 Θ^*, 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 $\sigma \in \{0.05, 0.1, 0.2\}$; the number of different values for σ considered in a single plot had to be limited to ensure readability since for **naive** and **EM**, (23) still depends substantially

on σ. 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 $\sigma = 0.05$ and $\sigma = 0.2$, respectively, while the dashed lines inside the shaded areas correspond to $\sigma = 0.1$. Accordingly, the performance of **naive** and **EM** (initialized by **naive**) relative to (23) improves, which is unsurprising given that as $\sigma\sqrt{m} \nearrow \|B^*\|_F$ (recall that $\|B^*\|_F = \sqrt{m}$), 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 $\sigma\sqrt{m}/\|B^*\|_F$ is small: for example, **proposed** achieves a roughly tenfold reduction in standardized estimation error over **naive** for $\sigma = 0.05$; that margin reduces gradually with increasing σ.

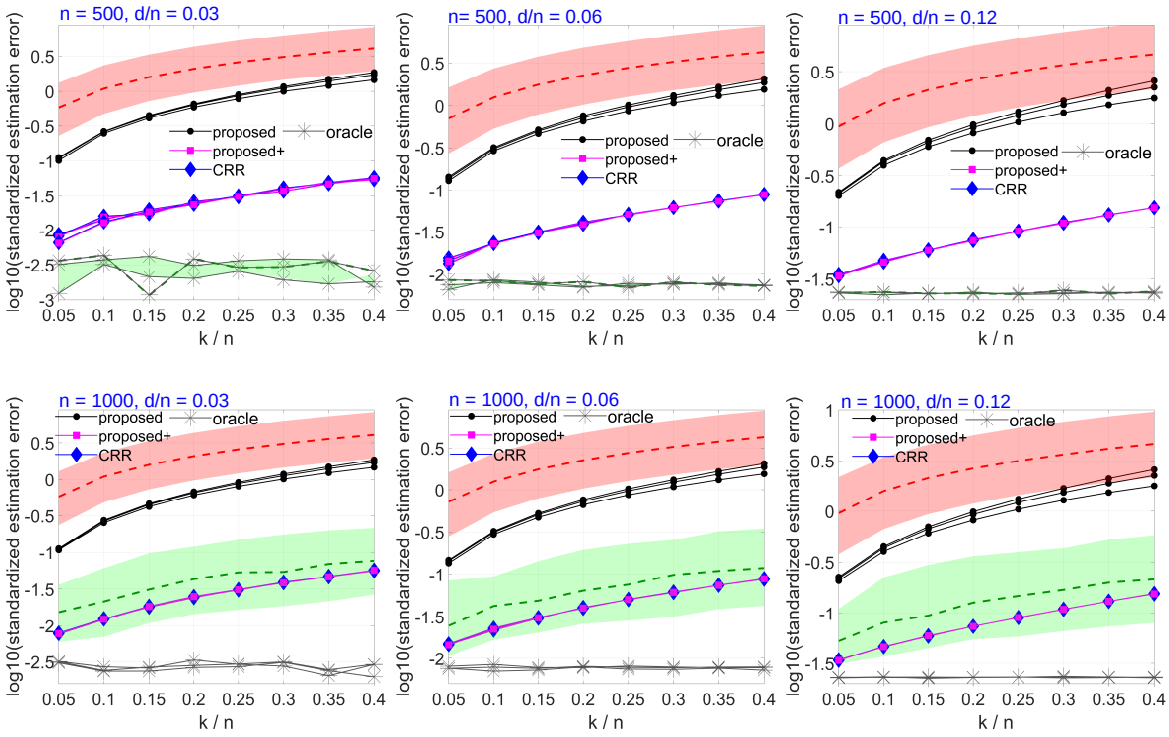


Figure 3: Average standardized estimation errors (23) on a $\log_{10}$-scale, with one curve for each $\sigma \in \{0.05, 0.1, 0.2\}$. For **naive** (in red) and **EM** (in green), the resulting curves do not cluster together, and are hence captured by the upper ($\sigma = 0.05$) and lower ($\sigma =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

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 n: specifically, the MCMC approximation tends to be less reliable for larger values n. For the same reason, **EM** is at least an order of magnitude slower than **proposed+** and **CRR**.

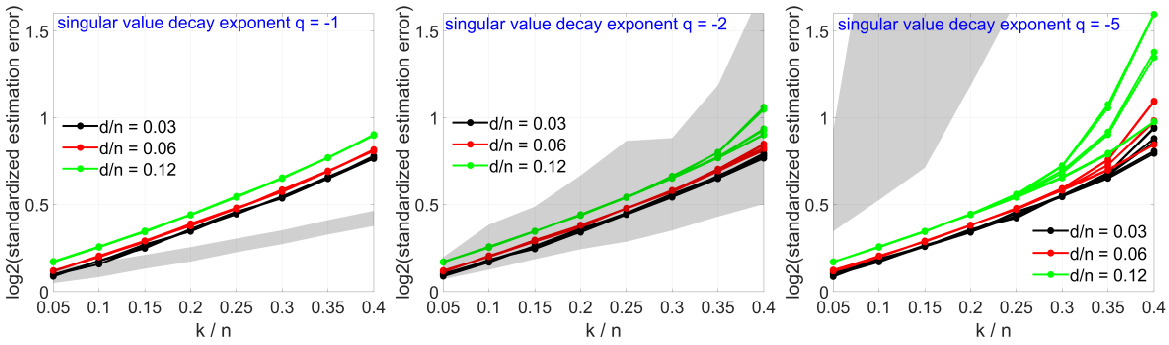


Figure 4: Average standardized estimation errors $\frac{\sigma^{-1}m^{-1/2}\|B^{\mathrm{est}}-B^*\|_F}{(d/(n-k))^{1/2}}$ ($\log_2$-scale) of the refitting approach **proposed+** (lines) and **EM** (shaded areas) for different rates of decay of the singular values of B^* corresponding to decreasing srank(B^*) from left to right. Curves for different combinations of n and σ 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**.

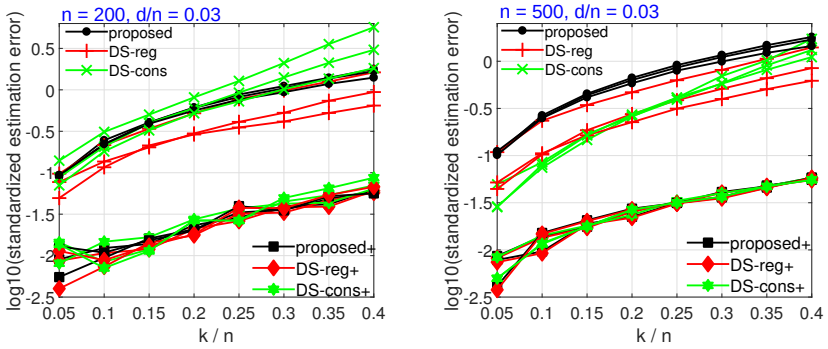


Figure 5: Average standardized estimation errors (23) of **DS-reg** and **DS-cons** in comparison to **proposed** along with their counterparts for refitting.

Results (II): Recovery of Θ^*. We evaluate the normalized Hamming distance $\frac{1}{n}\sum_{i=1}^n \mathbb{I}(\theta^*(i) \neq \widehat{\theta}(i))$, where the matrix counterpart of $\widehat{\theta}$ is given by $\widehat{\Theta}$, i.e., the plug-in estimator (7) (modified to incorporate the constraint that Θ^* is a permutation) with B^* replaced by $\widehat{B}$ from (9). In light of Theorem 2 and Lemmas 1 & 2, recovery of Θ^* is successful if $\gamma^2 \cdot \mathsf{SNR} \asymp n^{-c/\mathrm{srank}(B^*)} \cdot \mathsf{SNR}$ is large enough. We therefore plot the normalized Hamming distance in dependency of the (log)"normalized" SNR $-c/\mathrm{srank}(B^*)\log(n) - 2\log(\sigma)$, where the choice $c = 0.7$ was found to ensure a reasonable alignment of the results across different experimental configurations. Figure 6 indicates that recovery of Θ^* follows a phase transition: if the normalized SNR drops below a certain threshold, the normalized Hamming distance rises sharply. This observation is in alignment with the inachievability results in Theorem 2. Interestingly, plug-in estimation (lower panel) does not lead to a significant degradation in performance compared to the situation in which B^* is known (upper panel) even if the fraction of mismatches is noticeable ($k/n = 0.4$).

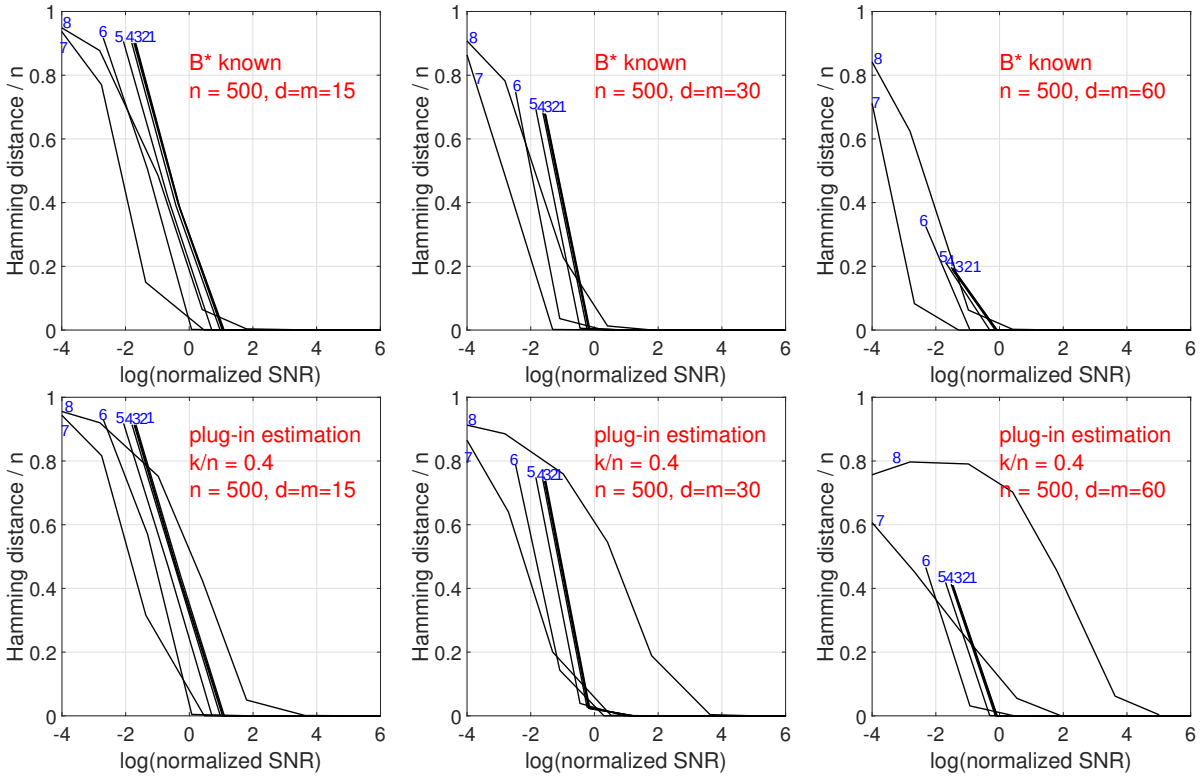


Figure 6: Average Hamming distance (scaled by $1/n$) between $\widehat{\Theta}(B^*)$ and Θ^* (top row) and between $\widehat{\Theta}(\widehat{B})$ and Θ^* (bottom row) vs. the (log) normalized SNR $= -c/\mathrm{srank}(B^*)\log(n) - 2\log(\sigma)$. The numbering indicates different values of the parameter q controlling $\mathrm{srank}(B^*)$, with higher numbers for larger q (smaller $\mathrm{srank}(B^*)$). The better the curves align, the more accurate the predicted dependence on the normalized SNR.

Table 1: Overview on the data sets considered in this paragraph. R^2 here refers to the coefficient of determination in the absence of shuffling.

Full Name	**Short Name**	n	d	m	R^2
SARCOS robot arm (Rasmussen and Williams, 2019)	sarcos	44,484	10	6	0.76
Flight Ticket Prices (Tsoumakas et al., 2011)	ftp	335	30	6	0.89
Supply Chain Management (Tsoumakas et al., 2011)	scm	8,966	35	16	0.58

Real Data

We consider three benchmark data sets for multivariate regression as tabulated in Table 1. The data sets are preprocessed versions of their original counterparts. The columns of the matrices X and Y were centered, and X was subsequently reduced to an adequate number of principal components since due to (almost) linearly independent predictors the oracle least squares estimator (here assigned the role of B^*) would (essentially) not be defined. For sarcos, one of the response variables was removed to improve goodness of fit, and hence to observe a better contrast in performance with an increasing fraction of mismatches. Likewise, two outliers with Cook's distance > 0.7 were removed from ftp. We randomly permute varying fractions (between 0.05 and 0.4) of the rows of Y, and investigate to what extent the proposed approach is able to restore the goodness-of-fit (in terms of the coefficient of determination $R^{2\dagger}$) and the regression coefficients of the least squares estimator in the complete absence of mismatches that here takes the role of B^*. The performance of the proposed approach is compared to naive least squares based on the permuted data. For each data set, we consider 20 independent random permutations for each value of k/n. Performance with regard to permutation recovery is assessed via $\|(\widehat{\Theta}(B^{\text{est}}) - \Theta^*)Y\|_F / \|(I_n - \Theta^*)Y\|_F$, i.e., via the relative reduction in error induced by random shuffling. This is a somewhat less stringent metric than the Hamming distance reported for synthetic data. The change in metric is motivated by the fact that exact permutation recovery cannot be expected for the data sets under consideration given that separability in terms of (18) relative to the noise level is poor. Approach (9) is run with the choice $\lambda = M \cdot \frac{\widehat{\sigma}_0}{\sqrt{n \cdot m}}$ for $M \in \{0.25, 0.5, 1, 2\}$ and $\widehat{\sigma}_0$ denoting the root mean square error of the least squares estimator in the absence of shuffling. We consider the same list of competitors and associated settings as for the synthetic data experiments, apart from the omission of **DS-reg** and **DS-cons** given the aforementioned scalability issues.

As can be seen from Figure 7, the results are not sensitive to the choice of the multiplier M. The proposed approach consistently improves over naive least squares once the fraction of mismatches exceeds 0.2, and yields more pronounced improvements as that fraction increases. Two-stage estimation of Θ^* yields noticeable reductions of the error $\|(I_n - \Theta^*)Y\|_F$ induced by shuffling. Approaches **proposed+** and **CRR** (equipped with knowledge of k), yield only occasional and rather minor improvements over **proposed**. Interestingly, **EM** exhibits poor performance even for moderate n (data set ftp), often falling short of **naive**

†. Here and in the sequel, the reported R^2 refers to the R^2 on the original data (i.e., before shuffling) given an estimator B^{est} obtained from the shuffled data (cf. caption of Figure 7).

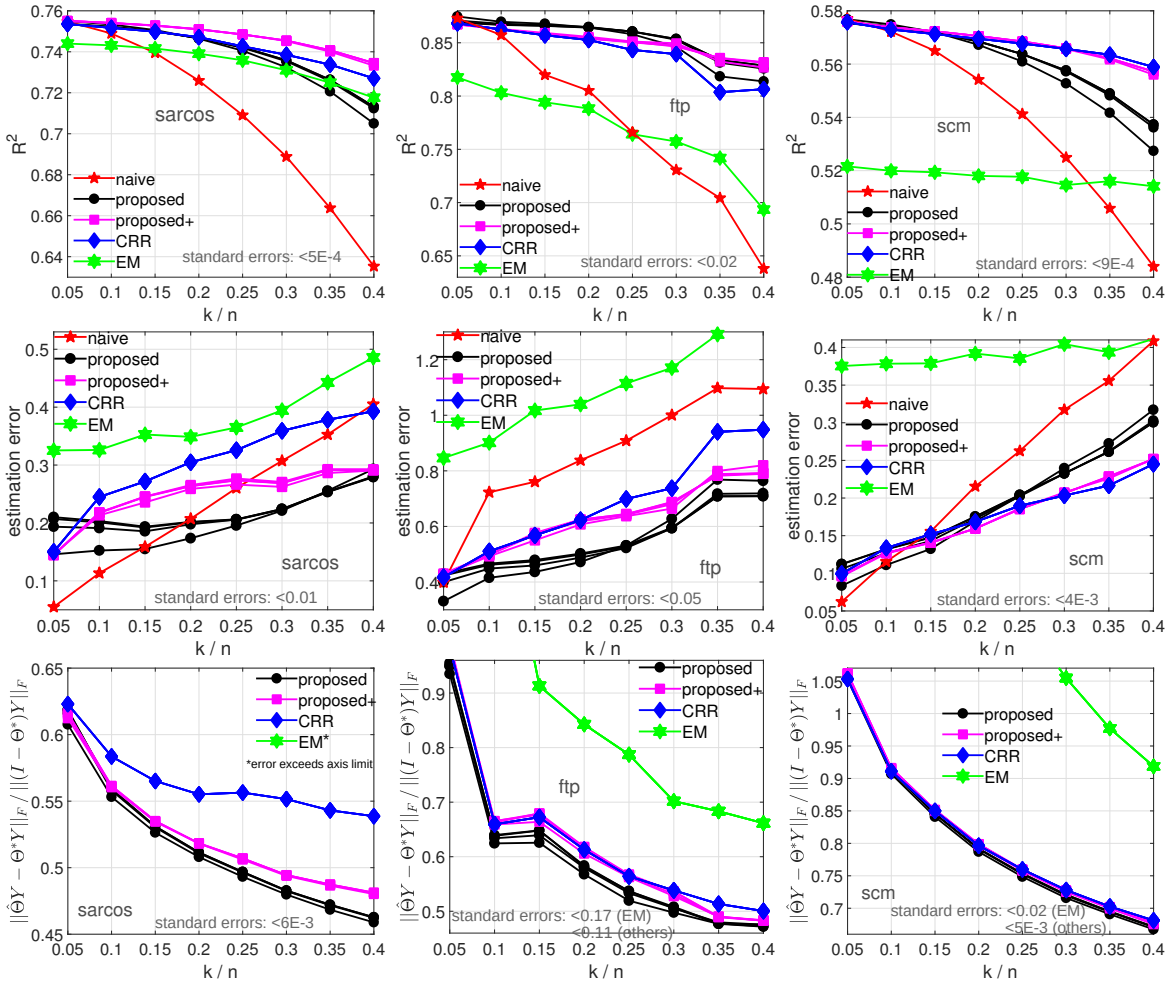


Figure 7: Top panels: Goodness of fit in terms of the coefficient of determination $R^2 = \|Y - XB^{\text{est}}\|_F^2/\|Y\|_F^2$. Middle panels: Relative estimation errors $\|B^{\text{est}} - B^*\|_F/\|B^*\|_F$, where B^* here refers to the oracle least squares estimator equipped with knowledge of Θ^*. Bottom panels: Performance in approximate recovery of Θ^* evaluated in terms of $\|(\widehat{\Theta}(B^{\text{est}}) - \Theta^*)Y\|_F/\|(I_n - \Theta^*)Y\|_F$. Each of the black lines corresponds to one specific value of the multiplier M in $\lambda = M\widehat{\sigma}_0/\sqrt{n \cdot m}$.

in sharp contrast to the results observed for the synthetic data. This raises the question whether competitive performance of **EM** is tied to specific properties of Gaussian design.

Case Study

We here illustrate the use of the proposed approach and its competitors in data integration scenarios based on a setting designed to mimic the analysis of data obtained from multiple sensors in an asynchronous fashion. The specific example presented in the sequel is based on the Multi-Site Beijing Air Quality data set (Chen, 2017) which contains measurements of various air pollutants and climate parameters recorded at an hourly rate from March 1st,

2013 to February 28th, 2017. For demonstration purposes, we confine ourselves to complete records from the site Nongzhanguan for the years 2016 and 2017 ($n = 9,726$). A linear regression model is fitted in which the response variables are given by the square roots of the air concentrations of the pollutants PM2.5, PM10, SO2, NO2, O3 ($m = 5$) and the predictor variables are given by temperature, dew point temperature, air pressure, precipitation, wind speed, CO concentration, and all associated quadratic terms plus intercept ($d = 28$). This model achieves an $R^2 \approx 0.725$.

At the next stage, we suppose that the response and predictor variables are collected by two different sensors, with temperature and air pressure collected by both sensors. In order to recreate the situation of mismatch error in record linkage that commonly results from the use of inexact or erroneous identifiers (Christen, 2012), the two sets of measurements are merged based on incomplete time stamps (day and hour are missing) and inaccurate temperature and air pressure measurements (rounded to integers). Requiring that linked records must agree on this combination of four matching variables implies that the merged file is of the form $[\Theta^* X \;\; Y]$, where Θ^* is a permutation matrix that can be arranged in block diagonal structure with the blocks corresponding to groups of measurements having the same combination of matching variables. It is assumed that the data analyst has no knowledge about the linkage process, in particular about the use of matching variables and the resulting block structure of Θ^*; this setting is typically referred to as "secondary analysis" in the record linkage literature (Chambers and da Silva, 2019).

Only $1,379$ out of $n = 9,726$ observations yield singleton blocks, i.e., they are uniquely identifiable based on the matching variables, while all other observations belong to blocks of size two up to 20. To fix $\Theta^* = \mathrm{bdiag}(\Theta^*_{(1)}, \ldots, \Theta^*_{(K)})$, we set $\Theta^*_{(l)} = \mathrm{argmax}_\Theta \|Y_{(l)} - \Theta Y_{(l)}\|_F^2$‡ where $Y_{(l)}$ denotes the rows of Y corresponding to the l-th block, $1 \leq l \leq K = 3,625$, and the argmax is over all permutations associated with the respective block. While the resulting nominal fraction of mismatches $|\{i : \Theta^*_{ii} \neq 1\}|/n \approx 0.63$ does not appear to fit the sparse regime, the majority of mismatches do not introduce substantial contamination in the sense that $\|Y_{i,:} - Y_{\theta^*(i),:}\|_F$ is within the noise level; to a good extent, this can be attributed to the fact that the responses tend to be more similar within blocks than across blocks.

The same regression model as above is fitted based on the merged records $[\Theta^* X \;\; Y]$. Naive least squares regression leads to a noticeable drop of the $R^2 \approx 0.66$ and a root mean squared error (RMSE) of 431.4 relative to the original (i.e., based on $[X \; Y]$) regression parameter estimate B^*. Application of the approach (9) with the choice $\lambda = \frac{\widehat{\sigma}}{\sqrt{n \cdot m}}$, where $\widehat{\sigma}$ can be taken as the root mean squared prediction error of either the original or the naive least squares fit, lifts the R^2 to 0.70 and reduces the RMSE for the regression parameter to 318.1. Following the proposed two-stage method, we use the resulting estimator $\widehat{B}$ to correct mismatches by solving the following optimization problem:

$$\min_{\Pi \in \mathcal{P}} \|Y - \Pi(\Theta^* X)\widehat{B}\|_F^2 \quad \text{subject to} \;\; \Pi_{ii} = 1 \text{ if } \|Y_{i,:} - \Theta^*_{i,:} X \widehat{B}\|_F \leq \sqrt{2m}\widehat{\sigma} \qquad (24)$$
$$\Pi_{ij} = 0 \text{ if } \|Y_{i,:} - \Theta^*_{i,:} X \widehat{B}\|_F \leq \|Y_{i,:} - \Theta^*_{j,:} X \widehat{B}\|_F,$$

for $1 \leq i, j \leq n$, where $\mathcal{P}$ denotes the set of all permutation matrices (4). Note that perfect recovery corresponds to $\Pi = (\Theta^*)^{-1}$. The additional constraints are imposed as a means

‡. This optimization problem reduces to a linear assignment problem.

to achieve sparsity of Π in the sense of small Hamming distance to the identity: the first constraint sets diagonal elements to one for which the discrepancy between observed and fitted values is within a factor of $\sqrt{2}$ of the noise level, and the second constraint excludes pairings that do not lead to improvements in terms of fit.

Given the minimizer $\widehat{\Pi}$ of (24), it is worth attempting a re-fit of the regression model based on data $[\widehat{\Pi}(\Theta^* X) \;\; Y]$. As shown in the top panel of Figure 8, the solution $\widehat{\Pi}$ is able to reduce mismatch error to an extent that is comparable to the error of the original regression model. Moreover, the bottom panel of Figure 8 shows that the fitted values of the re-fit agree considerably better with the fitted values based on $[X \; Y]$ relative to the fitted values of naive least squares (plot of the first principal component is meaningful here since here $\mathrm{srank}(Y) \approx 1$). Accordingly, the R^2 of the refit increases to 0.715 close to the original 0.725.

In addition, we consider the competitors **CRR** and **EM** as alternatives. **CRR** achieves slightly better performance than (9) with an oracular choice of its tuning parameter (sparsity level k); choosing the latter so as to minimize the R^2 at 0.717 yields the choice $k/n = 0.19$ while an R^2 of 0.71 or higher is achieved within the entire range $k/n \in [0.09, 0.32]$. The "effective" fraction of mismatches is expected to be contained in that interval. By contrast, the performance of **EM** is rather poor, with an additional drop of the R^2 compared to naive least squares. At the same time, the R^2 achieved by **EM** on the mismatched data is close to 0.8 (i.e., much larger than 0.725), which indicates substantial overfitting. A numerical summary of the performance of the approaches compared here can be found in Table 2.

Table 2: **oracle**: least squares fit based on the original data $[X \; Y]$; **prop**: short for **proposed**; **prop**-$\widehat{\Pi}$, **CRR**-$\widehat{\Pi}$: least squares refit after solving (24) with $\widehat{B}$ obtained according to (9) and (11), respectively. The second and third row contain the RMSE in estimating B^* including intercepts (a) and not including intercepts (b). Note that the combination of both tends to provide a more accurate picture: **EM** achieves a decent value for (a) despite poor performance based on R^2 and confirmed by (b).

	oracle	**naive**	**prop**	**prop+**	**CRR**	**EM**	**prop**-$\widehat{\Pi}$	**CRR**-$\widehat{\Pi}$
R^2	0.725	0.66	0.70	0.712	.717	0.625	0.715	0.715
B^*-RMSE^a	0	431.4	318.1	295.81	259.1	280.6	298.9	304.8
B^*-RMSE^b	0	4.11	3.94	3.98	3.42	5.97	3.67	3.58

5. Conclusion

In this paper, we have presented a computationally appealing two-stage approach to multivariate linear regression in the presence of a small to moderate number of mismatches. The proposed approach can be used to safeguard against a potentially dramatic increase in the estimation error that can be incurred when ignoring the possibility of mismatches, as demonstrated in terms of statistical analysis and supported by a series of empirical results. Moreover, under certain conditions involving "separability" of pairs of data points and the signal-to-noise ratio, it is shown that the true correspondence between those pairs can be

RMSE	(Y, XB^*)	$(Y, \Theta^{*-1}Y)$	$(\widehat{\Pi}Y, \Theta^{*-1}Y)$ **proposed**	$(\widehat{\Pi}Y, \Theta^{*-1}Y)$ **CRR**	$(\widehat{\Pi}Y, \Theta^{*-1}Y)$ **EM**
	1.8	2.53	1.89	1.86	2.13

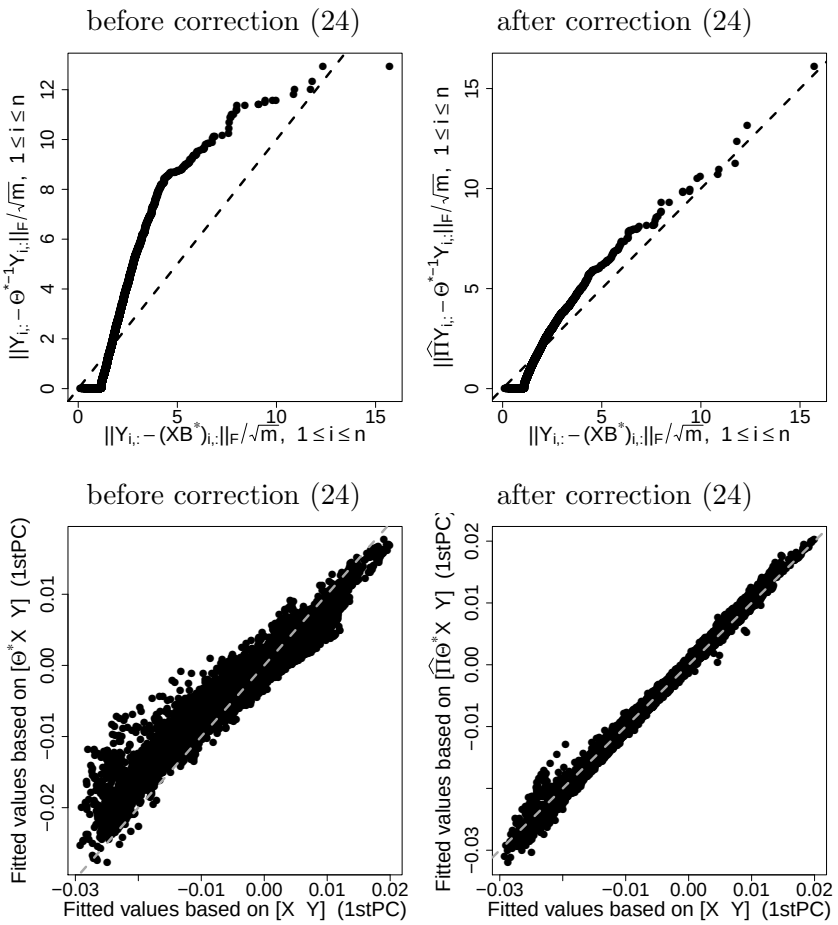


Figure 8: Table at the top: RMSEs of various quantities (A, B), i.e., $\|A - B\|_F/\sqrt{n \cdot m}$. The first entry equals the RMSE of the original least squares fit, the second entry equals the mismatch error introduced by Θ^*, and the remaining entries show the reduction based on (24) in combination with three methods for obtaining $\widehat{B}$. Top plots: Mismach error vs. residual error, before (left) and after correction based on (24) with $\widehat{B}$ from (9) (right). Bottom plots: the fitted values based on $[\Theta^* X \;\; Y]$ vs. fitted values based on $[X \; Y]$ (left), and the fitted values based on $[\Theta^* X \;\; Y]$ vs. fitted values based on $[\widehat{\Pi}(\Theta^* X) \;\; Y]$ (right). "Fitted values" here refer to the projection on the leading eigenvector (first principal component) of XB^*.

perfectly recovered. A key result in this paper asserts that the availability of multiple, linearly independent response variables (as measured by the stable rank of the regression coefficients) considerably simplifies the problem as it increases separability.

A limitation of the proposed approach is that it imposes a stringent limit on the allowed fraction of mismatches. In fact, as long as a sufficiently large superset of correctly matched data (of size $\Omega(n)$) can be identified, the regression parameter can still be estimated at the usual rate. Accordingly, the given problem does not appear hopeless even for significantly larger fraction of mismatches, say, up to $1 - \delta$ for δ bounded away from zero. Closing this gap is a worthwhile endeavor for future research. A second direction of future work concerns extension of the setup beyond classical linear models, specifically more flexibility regarding the range of the response variables (binary, mixed discrete/continuous etc.).

Acknowledgments

The first author was partially supported by the NSF Grant CCF-1849876. The authors would like to thank the Reviewers and Action Editor for their thoughtful and encouraging comments that have led to numerous improvements over an earlier draft. The authors also thank Zhenbang Wang for providing an implementation of the EM-based method in §4.

References

A. Abid and J. Zou. Stochastic EM for Shuffled Linear Regression. In *Allerton Conference on Communication, Control, and Computing*, pages 470–477, 2018.

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

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

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition edition, 1999.

D. Bertsekas and D. Castanon. A forward/reverse auction algorihtm for asymmetric assignment problems. *Computational Optimization and Applications*, 1:277–297, 1992.

K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2110–2119, 2017.

T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27:265–274, 2009.

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

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.

R. Chambers and A. da Silva. Improved secondary analysis of linked data: a framework and an illustration. *Journal of the Royal Statistical Society Series A*, 2019.

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

S. X. Chen. Beijing Multi-Site Air-Quality Data Data Set. `https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data`, 2017.

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

S. DasGupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22:60–65, 2003.

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

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

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

I. Dokmanić. Permutations unlabeled beyond sampling unknown. *IEEE Signal Processing Letters*, 26:823–827, 2019.

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

Y. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009.

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.

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

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

P. Goel and T. Ramalingam. *The Matching Methodology: Some Statistical Properties*. Springer Lecture Notes in Statistics, 2012.

Y. Gordon. *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$, pages 84–106. Springer Berlin Heidelberg, Berlin, Heidelberg, 1988.

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

D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of sub-Gaussian random vectors. *Electronic Communications in Probability*, 52:1–6, 2012.

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

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, pages 427–435, 2013.

H. Kuhn. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

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

J. Laska, M. Davenport, and R. Baraniuk. Exact Signal Recovery from Sparsely Corrupted Measurements through the Pursuit of Justice. In *Asilomar Conference on Signals, Systems and Computers*, pages 1556–1560, 2009.

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.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

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.

R. Ma, T. Cai, and H. Li. Optimal permutation recovery in permuted monotone matrix model. *to appear in Journal of the American Statistical Association*, 2020.

A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

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.

N. Nguyen and T. Tran. Robust Lasso with Missing and Grossly Corrupted Observations. *IEEE Transactions on Information Theory*, 59:2036–2058, 2013.

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

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.

Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. *IEEE Transactions on Information Theory*, 59:482–494, 2013a.

Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66:1275–1297, 2013b.

C. Rasmussen and C. Williams. Gaussian processes for machine learning: Data. http://www.gaussianprocess.org/gpml/data/, January 2019.

P. Rigollet and J. Weed. Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information and Inference*, 8:691–717, 2019.

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

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

Y. She and A. Owen. Outlier Detection Using Nonconvex Penalized Regression. *Journal of the American Statistical Association*, 106:626–639, 2012.

X. Shi, X. Lu, and T. Cai. Spherical regresion under mismatch corruption with application to automated knowledge translation. to appear in Journal of the American Statistical Association, 2020.

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

M. Slawski, M. Rahmani, and P. Li. A Robust Subspace Recovery Approach to Linear Regression with Partially Shuffled Labels. In *Uncertainty in Artificial Intelligence (UAI)*, 2019.

L. Sweeney. *Computational disclosure control: A primer on data privacy protection.* PhD thesis, Massachusetts Institute of Technology, 2001.

M. Tsakiris. Eigenspace conditions for homomorphic sensing. arXiv:1812.07966, December 2018.

M. Tsakiris and L. Peng. Homomorphic sensing. In *International Conference on Machine Learning (ICML)*, pages 6335–6344, 2019.

M. Tsakiris, L. Peng, A. Conca, L. Kneip, Y. Shi, and H. Choi. An algebraic-geometric apporach to shuffled linear regression. to appear in IEEE Transactions on Information Theory, 2020.

P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming Series B*, 12:263–295, 2010.

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.

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

R. Vershynin. *In: Compressed Sensing: Theory and Applications*, chapter 'Introduction to the non-asymptotic analysis of random matrices'. Cambridge University Press, 2012.

R. Vershynin. *High-Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

R. Vershynin and M. Rudelson. Spectral norm of products of random and deterministic matricess. *Probability Theory and Related Fields*, 150:471–509, 2011.

G. Wang, J. Zhu, R. Blum, P. Willett, S. Marano, V. Matta, and P. Braca. Signal Amplitude Estimation and Detection From Unlabeled Binary Quantized Samples. *IEEE Transactions on Signal Processing*, 66:4291–4303, 2018.

Wikipedia. List of cities by average temperature. `https://en.wikipedia.org/wiki/List_of_cities_by_average_temperature`, January 2019.

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

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.

H. Zhang and P. Li. Optimal estimator for unlabeled linear regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

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

H. Zhang, M. Slawski, and P. Li. The benefits of diversity: Permutation recovery in unlabeled sensing from multiple measurement vectors. *arXiv preprint arXiv:1909.02496*, 2019b.

Appendix A. Proof of Theorem 1

(I) *Bound on* $\|\Xi^* - \widehat{\Xi}\|_F$.

A crucial observation is that the joint optimization problem (9) in B and Ξ can be decomposed into two optimization problems involving only B and Ξ, respectively, as stated in the following Lemma.

Lemma A.1 *Consider optimization problem* (9) *with solution* $(\widehat{B}, \widehat{\Xi})$ *and denote by* $\mathtt{P}_X^\perp$ *the projection on the orthogonal complement of* $\mathrm{range}(X)$. *Then, if* $n \geq d$, *with probability one*

$$\widehat{\Xi} \in \mathfrak{X}, \quad \mathfrak{X} := \operatorname*{argmin}_{\Xi} \frac{1}{2n \cdot m} \|\mathtt{P}_X^\perp (Y - \sqrt{n}\Xi)\|_2^2 + \lambda \sum_{i=1}^n \|\Xi_{i,:}\|_2, \tag{25}$$

$$\widehat{B} \in \left\{ \left(\frac{X^\top X}{n} \right)^{-1} \frac{X^\top (Y - \sqrt{n}\widehat{\Xi})}{n}, \ \widehat{\Xi} \in \mathfrak{X} \right\}. \tag{26}$$

The proof is along the lines of the proof of Lemma 1 in Slawski and Ben-David (2019), and is hence omitted. Note that $\mathtt{P}_X^\perp Y = \mathtt{P}_X^\perp(\sqrt{n}\Xi^* + \sigma \widetilde{E})$ with $\widetilde{E} = \mathcal{S}E$. The optimization problem in (25) thus becomes

$$\min_{\Xi} \frac{1}{2n \cdot m} \|\mathtt{P}_X^\perp (\sqrt{n}\Xi^* + \sigma\widetilde{E} - \sqrt{n}\Xi)\|_2^2 + \lambda \sum_{i=1}^n \|\Xi_{i,:}\|_2 \tag{27}$$

In the sequel, we study an equivalent vectorized problem. Accordingly, we define

$$\begin{aligned} &\xi^* = [(\Xi^*_{:,1})^\top; \ldots; (\Xi^*_{:,m})^\top] \in \mathbb{R}^{n \cdot m}, \quad \widetilde{e} = [\widetilde{E}_{:,1}^\top; \ldots; \widetilde{E}_{:,m}^\top] \\ &\mathtt{P}_X^{\perp \otimes} = I_m \otimes \mathtt{P}_X^\perp = \begin{pmatrix} \mathtt{P}_X^\perp & 0 & \ldots & 0 \\ 0 & \mathtt{P}_X^\perp & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \mathtt{P}_X^\perp \end{pmatrix}, \end{aligned} \tag{28}$$

with $\otimes$ denoting the Kronecker product, the subscripts $_{:,j}$ refer to the j-th column, $j = 1, \ldots, m$, and ";" here means row-wise concatenation. Moreover, for any $v \in \mathbb{R}^{n \cdot m}$, we let

$$v^{[i]} = (v_j)_{j \in G_i}, \ i = 1, \ldots, n, \quad G_i = \{i, i+n, \ldots, i + (m-1) \cdot n\}.$$

With this in place, the $(2,q)$-norm with respect to $G_1, \ldots, G_n$ is defined by

$$\|v\|_{2,q} := \left(\sum_{i=1}^n \|v^{[i]}\|_2^q \right)^{1/q}, \ 1 \leq q < \infty, \quad \text{and } \|v\|_{2,\infty} := \max_{1 \leq i \leq n} \|v^{[i]}\|_2, \tag{29}$$

$$\|v\|_{2,0} := \sum_{i=1}^n \mathbb{I}(\|v^{[i]}\|_2 > 0), \tag{30}$$

where the latter is not a norm; it counts the number of non-zero groups of components, with each of the $\{G_i\}_{i=1}^n$ forming a group. Note that $\|\xi^*\|_{2,0} \leq k$ with support

$$S_* = \{1 \leq i \leq n : \Theta^*_{ii} \neq 1\} = \{1 \leq i \leq n : \|\xi^{*[i]}\|_2 > 0\}.$$

We also observe that for all $v, w \in \mathbb{R}^{n \cdot m}$

$$\|v\|_{2,2} = \|v\|_2, \qquad |\langle v, w \rangle| = \left| \sum_{i=1}^n v^{[i]\top} w^{[i]} \right| \leq \sum_{i=1}^n \|v^{[i]}\| \|w^{[i]}\|_2 \leq \|v\|_{2,1} \|w\|_{2,\infty} \tag{31}$$

by the inequalities of Cauchy-Schwarz and Hölder.

After these preparations, we are in position to state another Lemma. First note that optimization problem (27) can be expressed in vectorized form as

$$\min_{\xi} \frac{1}{2n \cdot m} \|\mathtt{P}_X^{\perp\otimes}(\sqrt{n}\xi^* + \sigma\widetilde{e}) - \mathtt{P}_X^{\perp\otimes}\xi\sqrt{n}\|_2^2 + \lambda \sum_{i=1}^n \|\xi^{[i]}\|_2, \tag{32}$$

Letting $\widehat{\delta} = \xi^* - \widehat{\xi}$, where $\widehat{\xi}$ is a minimizer of (32), we have the following basic inequality

$$\frac{1}{2n \cdot m} \|\mathtt{P}_X^{\perp\otimes}\sqrt{n}\widehat{\delta}\|_2^2 + \lambda \sum_{i=1}^n \|\widehat{\xi}^{[i]}\|_2 \leq \frac{1}{\sqrt{n} \cdot m} |\langle \mathtt{P}_X^{\perp\otimes}\widehat{\delta}, \sigma\widetilde{e}\rangle| + \lambda \sum_{i \in S_*} \|\xi^{*[i]}\|_2, \tag{33}$$

which is obtained by evaluating (32) at $\xi = 0$, expanding squares and re-arranging.

Lemma A.2 *Consider $\widehat{\delta}$ in* (33) *and Let λ_0 be a number such that*

$$\frac{1}{\sqrt{n} \cdot m} \|\mathtt{P}_X^{\perp\otimes}\sigma\widetilde{e}\|_{2,\infty} \leq \lambda_0. \tag{34}$$

Then for any $\lambda \geq 2\lambda_0$, it holds that either $\widehat{\delta} = 0$ or $\widehat{\delta}/\|\widehat{\delta}\|_2 \in 2\,\mathrm{conv}(B_0(k')) \cap \mathbb{S}^{n\cdot m-1}$, where for $r \geq 0$, $B_0(r) = \{v \in \mathbb{R}^{n\cdot m} : \ \|v\|_{2,0} \leq r, \ \|v\|_2 \leq 1\}$ according to (29) *and $k' = \left(1 + \frac{\lambda+\lambda_0}{\lambda-\lambda_0}\right)^2 k \leq 16k$.*

Proof As an immediate consequence of (33) and the triangle inequality, we obtain that

$$\lambda \sum_{i \in S_*^c} \|\widehat{\delta}^{[i]}\|_2 \leq \frac{1}{\sqrt{n} \cdot m} |\langle \mathtt{P}_X^{\perp\otimes}\widehat{\delta}, \sigma\widetilde{e}\rangle| + \lambda \sum_{i \in S_*} \|\widehat{\delta}^{[i]}\|_2 \leq \lambda_0 \|\widehat{\delta}\|_{2,1} + \lambda \sum_{i \in S_*} \|\widehat{\delta}^{[i]}\|_2,$$

where the second inequality is a result of (31) and (34). If $k = 0$, $S_* = \emptyset$, we must have $\widehat{\delta} = \widehat{\xi} = \xi^* = 0$ as the above inequality would be violated otherwise, and the claim of the lemma follows. On the other hand, if $k \geq 1$, combination of the left and right hand side of the above chain of inequalities yields

$$\begin{aligned}
&\lambda \sum_{i \in S_*^c} \|\widehat{\delta}^{[i]}\|_2 \leq \lambda_0 \|\widehat{\delta}\|_{2,1} + \lambda \sum_{i \in S_*} \|\widehat{\delta}^{[i]}\|_2 = \lambda_0 \left(\sum_{i \in S_*} \|\widehat{\delta}^{[i]}\|_2 + \sum_{i \in S_*^c} \|\widehat{\delta}^{[i]}\|_2 \right) + \lambda \sum_{i \in S_*} \|\widehat{\delta}^{[i]}\|_2 \\
\Rightarrow &\sum_{i \in S_*^c} \|\widehat{\delta}^{[i]}\|_2 \leq \frac{\lambda + \lambda_0}{\lambda - \lambda_0} \sum_{i \in S_*} \|\widehat{\delta}^{[i]}\|_2 \\
\Rightarrow &\|\widehat{\delta}\|_{2,1} \leq \left(1 + \frac{\lambda + \lambda_0}{\lambda - \lambda_0}\right) \sum_{i \in S_*} \|\widehat{\delta}^{[i]}\|_2 \leq \left(1 + \frac{\lambda + \lambda_0}{\lambda - \lambda_0}\right) \sqrt{k} \|\widehat{\delta}\|_2.
\end{aligned} \tag{35}$$

The assertion then follows from Lemma E.1 provided in a separate section below. ■

As in the above Lemma, under event (34), inequality (33) implies

$$\frac{1}{2n \cdot m} \|\mathtt{P}_X^{\perp\otimes}\sqrt{n}\widehat{\delta}\|_2^2 + \leq \left(\lambda_0 \left(1 + \frac{\lambda + \lambda_0}{\lambda - \lambda_0}\right) + \lambda \right) \sqrt{k} \|\widehat{\delta}\|_2 = \lambda \left(\frac{\lambda + \lambda_0}{\lambda - \lambda_0} \right) \sqrt{k} \|\widehat{\delta}\|_2 \tag{36}$$

by following the steps leading to (35). We now lower bound the l.h.s. of (36). Let $\Lambda = \{(\lambda_s)_{s=1}^N \subset \mathbb{R}_+ :\ N \in \{1,2,\dots,\},\ \sum_{s=0}^N \lambda_s \le 2\}$. In light of Lemma A.2, we have

$$\frac{1}{n}\|\mathsf{P}_X^{\perp\otimes}\sqrt{n}\widehat{\delta}\|_2^2 \ge \|\widehat{\delta}\|_2^2 \min_{\substack{\{\lambda_s\}\in\Lambda,\,\{v_s\}\subset B_0(k'),\\ \sum_s \lambda_s v_s \in \mathbb{S}^{n\cdot m-1}}} \|\mathsf{P}_X^{\perp\otimes}\textstyle\sum_s \lambda_s v_s\|_2^2.$$

Structuring each v_s into sub-vectors $v_s^{(l)} \in \mathbb{R}^n$, $l = 1,\dots,m$, we obtain

$$\min_{\substack{\{\lambda_s\}\in\Lambda,\,\{v_s\}\subset B_0(k'),\\ \sum_s \lambda_s v_s \in \mathbb{S}^{n\cdot m-1}}} \|\mathsf{P}_X^{\perp\otimes}\textstyle\sum_s \lambda_s v_s\|_2^2 = \min_{\substack{\{\lambda_s\}\in\Lambda,\,\{v_s\}\subset B_0(k'),\\ \sum_s \lambda_s v_s \in \mathbb{S}^{n\cdot m-1}}} \sum_{l=1}^m \|\mathsf{P}_X^{\perp}\textstyle\sum_s \lambda_s v_s^{(l)}\|_2^2.$$

Since each v_s is k'-group sparse according to the partitioning defined by $\{G_i\}_{i=1}^n$, each $v_s^{(l)}$ is at most k'-sparse in the ordinary sense, i.e., having at most k' non-zero entries. Letting $\mathcal{B}_0(k') = \{v \in \mathbb{R}^n : \|v\|_0 \le k'\}$ denote the usual k'-sparsity ball in $\mathbb{R}^n$, we have

$$\begin{aligned}
&\min_{\substack{\{\lambda_s\}\in\Lambda,\,\{v_s\}\subset B_0(k'),\\ \sum_s \lambda_s v_s \in \mathbb{S}^{n\cdot m-1}}} \sum_{l=1}^m \|\mathsf{P}_X^{\perp}\textstyle\sum_s \lambda_s v_s^{(l)}\|_2^2 \\
=&\min_{\substack{\{\lambda_s\}\in\Lambda,\,\{v_s^{(l)}\}\subset\mathcal{B}_0(k'),\\ \{\sum_s \lambda_s v_s^{(l)}\}\subset\mathbb{S}^{n-1}\\ \{\gamma^{(l)}\}\subset\mathbb{R}_+,\,\sum_{l=1}^m\{\gamma^{(l)}\}^2=1}} \sum_{l=1}^m \|\mathsf{P}_X^{\perp}\,\gamma^{(l)}\textstyle\sum_s \lambda_s v_s^{(l)}\|_2^2 \\
=&\min_{\{\gamma^{(l)}\}\subset\mathbb{R}_+,\,\sum_{l=1}^m\{\gamma^{(l)}\}^2=1} \sum_{l=1}^m \{\gamma^{(l)}\}^2 \times \min_{u\in 2\mathrm{conv}(\mathcal{B}_0(k'))\cap\mathbb{S}^{n-1}} \|\mathsf{P}_X^{\perp}u\|_2^2 \\
=&\min_{u\in 2\mathrm{conv}(\mathcal{B}_0(k'))\cap\mathbb{S}^{n-1}} \|\mathsf{P}_X^{\perp}u\|_2^2 \\
=&\mathrm{dist}^2(2\mathrm{conv}(\mathcal{B}_0(k'))\cap\mathbb{S}^{n-1}, \mathrm{range}(X)).
\end{aligned} \tag{37}$$

In order to lower bound this squared distance, we apply Gordon's Theorem (cf. Lemma E.3 below) with $K = 2\mathrm{conv}(\mathcal{B}_0(k'))\cap\mathbb{S}^{n-1}$ and $V = \mathrm{range}(X)$ noting that the latter random subspace follows a uniform distribution on the Grassmannian $\mathsf{G}(n,d)$, thus we identify $p = n$, $p - q = d \Leftrightarrow q = n - d$. It is well-known that $\nu_r = \sqrt{r^2/(r+1)} = (1 - O(1/\sqrt{r}))\sqrt{r} \sim \sqrt{r}$ as $r \to \infty$; to simplify our argument, we henceforth replace ν_r by $\sqrt{r}$. Translated to the setting under consideration, the condition $w(K) < (1-\varepsilon)\nu_q - \varepsilon\nu_p$ in Lemma E.3 reads

$$\frac{1}{1-\varepsilon}\, w(2\mathrm{conv}(\mathcal{B}_0(k'))\cap\mathbb{S}^{n-1}) < \sqrt{n-d} - \frac{\varepsilon}{1-\varepsilon}\sqrt{n}. \tag{38}$$

Invoking the assumption $d/n \le 1/4$, the r.h.s. of (38) evaluates as $(\sqrt{3}/2 - \frac{\varepsilon}{1-\varepsilon})\sqrt{n}$. Regarding the l.h.s. of (38), it follows from standard results (cf. Plan and Vershynin (2013a), Lemma 2.3) that the Gaussian width $w(2\,\mathrm{conv}(\mathcal{B}_0(k'))\cap\mathbb{S}^{n-1}) \le 7\sqrt{k'\log(en/k')}$. It thus follows that for any $\varepsilon \in (0, 1/3)$, there exists $c_\varepsilon, c'_\varepsilon > 0$ so that if

$$k \le c_\varepsilon \cdot n/\log(n/k)$$

inequality (38) is satisfied, so that with probability at least $1-3.5\cdot\exp(-c'_\varepsilon n)$, (37) is lower bounded by ε^2. Combining (36) and this lower bound on (37), we conclude that

$$m^{-1/2}\|\widehat{\Xi}-\Xi^*\|_F = m^{-1/2}\|\widehat{\delta}\|_2 \leq \varepsilon^{-2}\cdot 2\lambda\sqrt{m}\cdot\frac{\lambda+\lambda_0}{\lambda-\lambda_0}\sqrt{k}.$$

The lemma below elaborates on the choice of λ_0, which completes the proof of the bound on $m^{-1/2}\|\widehat{\Xi}-\Xi^*\|_F$.

Lemma A.3 *With probability at least $1-2/n$, it holds that*

$$\frac{1}{\sqrt{n}\cdot m}\|\mathtt{P}_X^{\perp\otimes}\sigma\widetilde{e}\|_{2,\infty}\leq\lambda_0 \quad \textit{with } \lambda_0=\frac{\mu_{n,d}\,\sigma}{\sqrt{n\cdot m}}\left(1+\sqrt{\frac{4\log n}{m}}\right),\ \mu_{n,d}:=\left(\tfrac{n-d}{n}+\sqrt{24\tfrac{\log n}{n}}\right)\wedge 1.$$

Proof

$$\frac{1}{\sqrt{n}\cdot m}\|\mathtt{P}_X^{\perp\otimes}\sigma\widetilde{e}\|_{2,\infty}=\frac{\sigma}{\sqrt{n}\cdot m}\max_{1\leq i\leq n}\|E^\top\mathcal{S}^\top\mathtt{P}_X^\perp\mathfrak{e}_i\|_2,$$

where $\{\mathfrak{e}_i\}_{i=1}^n$ is the canonical basis of $\mathbb{R}^n$. Observe that conditional on $\mathtt{P}_X^\perp$, $E^\top\mathcal{S}^\top\mathtt{P}_X^\perp\mathfrak{e}_i$ is a zero mean-Gaussian random vector with covariance matrix $\|\mathcal{S}\,\mathtt{P}_X^\perp\mathfrak{e}_i\|_2^2\cdot I_m$, $1\leq i\leq n$. Since $\|\mathcal{S}\|_2\leq 1$ and since $\mathtt{P}_X^\perp$ is a random projection in the sense of DasGupta and Gupta (2003), it follows from results therein that for all $\mu>0$

$$\mathbf{P}\left(\max_{1\leq i\leq n}\|\mathcal{S}\,\mathtt{P}_X^\perp\mathfrak{e}_i\|_2^2\geq\frac{n-d}{n}(1+\mu)\wedge 1\right)\leq n\exp\left(-(n-d)\frac{\eta^2}{12}\right).$$

In particular, with the choice $\mu=\sqrt{24\frac{\log n}{n-d}}=:c_1$,

$$\mathbf{P}\left(\max_{1\leq i\leq n}\|\mathcal{S}\,\mathtt{P}_X^\perp\mathfrak{e}_i\|_2^2\geq\mu_{n,d}\right)\leq 1/n,\quad \mu_{n,d}:=\left(\tfrac{n-d}{n}+\sqrt{24\tfrac{\log n}{n}}\right)\wedge 1.$$

Combining this result with Lemma E.2 with $r=m$, $L=n$, $\max_{1\leq\ell\leq L}\sigma_\ell=\mu_{n,d}$, we have

$$\|\mathtt{P}_X^{\perp\otimes}\sigma\widetilde{e}\|_{2,\infty}\leq\mu_{n,d}\,\sigma\{\sqrt{m}+2\sqrt{\log n}\}$$

with probability at least $1-2/n$. This finally yields the choice

$$\lambda_0=\frac{\mu_{n,d}\,\sigma}{\sqrt{n\cdot m}}\left(1+\sqrt{\frac{4\log n}{m}}\right).$$

■

(II) *Bound on $\|B^*-B\|_F$.*

Let $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the minimum and maximum singular value functional, respectively. Invoking Lemma A.1, we bound

$$\begin{aligned}\|\widehat{B}-B^*\|_F &\leq \frac{\left\|\left(\frac{X^\top X}{n}\right)^{-1}\frac{X^\top}{\sqrt{n}}(\sigma\mathcal{S}E+\sqrt{n}(\Xi^*-\widehat{\Xi})\right\|_F}{\sqrt{n}}\\ &\leq \sigma\frac{\left\|\left(\frac{X^\top X}{n}\right)^{-1}\frac{X^\top}{\sqrt{n}}\mathcal{S}E\right\|_F}{\sqrt{n}}+\frac{\|\widehat{\Xi}-\Xi^*\|_F}{\sigma_{\min}(X/\sqrt{n})},\end{aligned}\tag{39}$$

where we have used that $\left(\frac{X^\top X}{n}\right)^{-1}\frac{X^\top}{\sqrt{n}} = \left(\frac{X}{\sqrt{n}}\right)^\dagger$, with † denoting the Moore-Penrose pseudo-inverse, and $\sigma_{\max}\left(\left(\frac{X}{\sqrt{n}}\right)^\dagger\right) = \sigma_{\min}^{-1}((X/\sqrt{n})^\dagger)$. Consider $\Gamma = \mathcal{S}\frac{X}{\sqrt{n}}\left(\frac{X^\top X}{n}\right)^{-2}\frac{X^\top}{\sqrt{n}}\mathcal{S}$, and let $\Gamma^\otimes = I_m \otimes \Gamma$. We then can write

$$\left\|\left(\tfrac{X^\top X}{n}\right)^{-1}\frac{X^\top}{\sqrt{n}}\mathcal{S}E\right\|_F^2 = \|\Gamma^\otimes e\|_2^2,$$

where e is a standard Gaussian random vector of dimension $n \cdot m$. By straightforward adaptations of Lemma 3 in Slawski and Ben-David (2019) that is based on a concentration result for quadratic forms in Hsu et al. (2012), we obtain that

$$\mathbf{P}\left(\left\|\left(\tfrac{X^\top X}{n}\right)^{-1}\frac{X^\top}{\sqrt{n}}E\right\|_F > \frac{\sqrt{5(d\cdot m \vee \log(n\cdot m))}}{\sigma_{\min}(X/\sqrt{n})}\,\middle|\, X\right) \le \exp(-(d\cdot m)\vee\log(n\cdot m)).$$

The proof is completed by appealing to concentration results (e.g., Corollary 5.35 in Vershynin (2012)) to lower bound $\sigma_{\min}(X/\sqrt{n})$ with X having i.i.d. standard Gaussian entries.

Appendix B. Proofs of Lemmas 1 and 2

Lemma 1 is an immediate consequence of the following result.

Lemma B.1 *(Proposition 2.6 in Latala et al. (2007))*
Let $g \sim N(0, I_d)$. There exist universal constants $\alpha_0 \in (0,1)$ and $\kappa > 0$ such that for any $\alpha \in (0, \alpha_0)$

$$\sup_{\mu\in\mathbb{R}^m}\mathbf{P}\left(\|\mu - B^{*\top}g\|_2 \le \alpha\|B^*\|_F\right) \le \exp\left(\kappa\log(\alpha)\,\mathrm{srank}(B^*)\right).$$

Lemma 1 is obtained by applying Lemma B.1 with $\mu = 0$, $g = \frac{\mathbf{x}_i - \mathbf{x}_j}{\sqrt{2}}$, and then using a union bound over pairs, i.e., $\{\min_{i<j}\|B^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2 \le \delta\} \subseteq \bigcup_{i<j}\{\|B^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2 \le \delta\}$ for any $\delta > 0$. We then choose α as the term inside the curly brackets in (19) to conclude the result.

Remark 1. Lemma B.1 immediately implies that the quantity γ_0^2 (20) exhibits qualitatively the same lower bound as γ^2 according to Lemma 1: since it is assumed that $\{\mathbf{y}_i : i \in \mathcal{N}\}$ and $\{\mathbf{x}_j : 1 \le j \le n\}$ are independent, we have

$$\begin{aligned}\mathbf{P}\left(\min_{\substack{i\in\mathcal{N}\\1\le j\le n}}\|\mathbf{y}_i - B^{*\top}\mathbf{x}_j\|_2 \le \delta\right) &\le \sum_{i\in\mathcal{N}}\sum_{j=1}^n \mathbf{E}_{\mathbf{y}_i}\left[\mathbf{P}(\|\mathbf{y}_i - B^{*\top}\mathbf{x}_j\|_2 \le \delta\,|\mathbf{y}_i)\right]\\ &\le |\mathcal{N}|n\sup_{\mu\in\mathbb{R}^m}\mathbf{P}(\|\mu - B^{*\top}\mathbf{x}_j\|_2 \le \delta),\end{aligned}$$

and thus Lemma B.1 can be applied as in the proof of Lemma 1. Since $|\mathcal{N}|n \lesssim \binom{n}{2}$, the lower bound (19) also holds true for γ_0^2 up to a constant factor, i.e., $\gamma_0^2 \gtrsim \gamma^2$.

Remark 2. A similar albeit slightly weaker result than Lemma B.1 holds true if the entries of g are independent, unit variance *sub*-Gaussian random variables (see, e.g., §2.5 in Vershynin (2018)). Specifically, Theorem 2.5 in Latala et al. (2007) implies that

$$\sup_{\mu\in\mathbb{R}^m} \mathbf{P}\left(\|\mu - B^{*\top} g\|_2 \le \frac{1}{2}\|B^*\|_F\right) \le 2\exp\left(-c\cdot \mathrm{srank}(B^*)\right),$$

for some constant $c > 0$. The main difference of the above result and that of Lemma B.1 is that the tail bound in the latter can still be driven to zero even if $\mathrm{srank}(B^*) = O(1)$ by choosing the parameter α appropriately. On the other hand, if α is chosen as a constant bounded away from zero, the two results yield qualitatively the same conclusions.

Regarding Lemma 2, we first prove the lower bound. We observe that under the assumption of B^* having constant non-zero singular values, $\|B^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \sim 2b_*^2\chi^2(r)$, where $\chi^2(\nu)$ denotes the Chi-Square distribution with $\nu \in \{1, 2, \ldots\}$ degrees of freedom. It is easy to verify that for $r = 2(q+1)$, $q \in \{0, 1, \ldots\}$,

$$\mathbf{P}(\chi^2(r) \le z) = 1 - \exp(-z/2)\sum_{s=0}^{q}\frac{(z/2)^s}{s!}, \quad z \ge 0. \tag{40}$$

Combining (40) with a union bound over pairs $i < j$, we obtain

$$\mathbf{P}\left(\min_{i<j}\|B^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \le 2b_*^2 z\right) \le \binom{n}{2}\left(1 - \exp(-z/2)\sum_{s=0}^{q}\frac{(z/2)^s}{s!}\right). \tag{41}$$

Below, z is chosen s.t. the r.h.s. of the above inequality is upper bounded by δ. We have

$$\binom{n}{2}\left(1 - \exp(-z/2)\sum_{s=0}^{q}\frac{(z/2)^s}{s!}\right) = \binom{n}{2}\left(\exp(-z/2)\sum_{s=q+1}^{\infty}\frac{(z/2)^s}{s!}\right) \le \binom{n}{2}\frac{(z/2)^{q+1}}{(q+1)!}, \tag{42}$$

where the inequality follows from a Taylor expansion with Lagrange form of the remainder:

$$\begin{aligned}
&\exp(z/2) = \sum_{s=0}^{q}\frac{(z/2)^s}{s!} + \frac{\exp(\xi)}{(q+1)!}(z/2)^{q+1} \ \text{ for some } \xi \in [0, z/2] \\
\Rightarrow\ &\exp(z/2) - \sum_{s=0}^{q}\frac{(z/2)^s}{s!} = \sum_{s=q+1}^{\infty}\frac{(z/2)^s}{s!} = \frac{\exp(\xi)}{(q+1)!}(z/2)^{q+1} \le \exp(z/2)\frac{(z/2)^{q+1}}{(q+1)!}.
\end{aligned}$$

Using that $\frac{1}{(q+1)!} \le ((q+1)/e)^{-(q+1)}$, (42) can be upper bounded as

$$\binom{n}{2}\exp\left(-(q+1)\log\left(\frac{2(q+1)}{z\cdot e}\right)\right) \le \frac{n^2}{2}\left(\frac{2(q+1)}{z\cdot e}\right)^{-(q+1)}.$$

Choosing $z = \frac{2}{e}(q+1)\cdot(n^{-2}\delta)^{1/(q+1)}$ ensures that the probability in (41) is bounded by $\frac{\delta}{2}$.

We turn to the upper bound in Lemma 2. Let $n_2 = \lfloor \frac{n}{2} \rfloor$. We first use that for any $z \geq 0$

$$\begin{aligned}\mathbf{P}\left(\min_{i<j}\|\mathbf{B}^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 < z\right) &\geq \mathbf{P}\left(\min_{1\leq i\leq n/2}\|(\mathbf{B}^*)^\top(\mathbf{x}_{2i} - \mathbf{x}_{2i-1})\|_2^2 < z\right) \\ &= 1 - \mathbf{P}(\chi^2(r) > z/2b_*^2)^{n_2}, \end{aligned} \tag{43}$$

where we have used that $\{\|\mathbf{B}^{*\top}(\mathbf{x}_{2i} - \mathbf{x}_{2i-1})\|_2^2\}_{i=1}^{n_2} \overset{\text{i.i.d.}}{\sim} 2b_*^2\chi^2(r)$. Using (40) and setting $z = c \cdot 4b_*^2$ in (43) for $c > 0$ to be determined below, we obtain that

$$\begin{aligned}\mathbf{P}\left(\min_{i<j}\|\mathbf{B}^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 < z\right) &\geq 1 - \left(\sum_{s=0}^{q}\frac{c^s}{s!}\exp(-c)\right)^{n_2} \\ &= 1 - \left(1 - \sum_{s=q+1}^{\infty}\frac{c^s}{s!}\exp(-c)\right)^{n_2} \\ &\geq 1 - \left(1 - \tfrac{c^{q+1}}{(q+1)!}\exp(-c)\right)^{n_2}. \end{aligned} \tag{44}$$

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 (44)

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

as long as $n \geq \theta$. Setting $\theta = 8$, the above probability is lower bounded by 0.75 if $n > 8(q+1)^{q+1}$. Combining this with the choice of $z = c \cdot 4b_*^2$ in (43) yields the assertion.

Appendix C. Proof of Theorem 2

We first show that $\widehat{\Theta}(\widehat{B})_{i,:} = \Theta^*_{i,:} = 0$ for $i \in \mathcal{N} = \{1 \leq i \leq n : \ \theta^*(i) = 0\}$. For this purpose, it needs to be established that $\min_{i\in\mathcal{N}}\min_{1\leq j\leq n}\|\mathbf{y}_i - \widehat{B}^\top\mathbf{x}_j\|_2 > \tau$. We have

$$\begin{aligned}\min_{i\in\mathcal{N}}\min_{1\leq j\leq n}\|\mathbf{y}_i - \widehat{B}^\top\mathbf{x}_j\|_2 &\geq \min_{i\in\mathcal{N}}\min_{1\leq j\leq n}\|\mathbf{y}_i - B^{*\top}\mathbf{x}_j\|_2 - \max_{1\leq j\leq n}\|\mathbf{x}_j\|_2\|B^* - \widehat{B}\|_2 \\ &\geq \gamma_0\sigma\sqrt{m}\mathsf{SNR}^{1/2} - \max_{1\leq j\leq n}\|\mathbf{x}_j\|_2\|B^* - \widehat{B}\|_2 \\ &> 2\max\left\{\max_{1\leq j\leq n}\|\mathbf{x}_j\|_2\|B^* - \widehat{B}\|_2, \tau\right\} - \max_{1\leq j\leq n}\|\mathbf{x}_j\|_2\|B^* - \widehat{B}\|_2 > \tau, \end{aligned}$$

in view of the event $\mathcal{B}$ defined in the theorem.

Next, we show that $\widehat{\Theta}(\widehat{B})_{i,:} \neq 0$ if $i \in \mathcal{N}^c$. This is implied by demonstrating that $\max_{i\in\mathcal{N}^c}\|\mathbf{y}_i - \widehat{B}^\top\mathbf{x}_{\theta^*(i)}\|_2 \leq \tau$. We have

$$\begin{aligned}\max_{i\in\mathcal{N}^c}\|\mathbf{y}_i - \widehat{B}^\top\mathbf{x}_{\theta^*(i)}\|_2 &\leq \max_{i\in\mathcal{N}^c}\|\mathbf{y}_i - B^{*\top}\mathbf{x}_{\theta^*(i)}\|_2 + \max_{1\leq j\leq n}\|\mathbf{x}_j\|_2\|B^* - \widehat{B}\|_2 \\ &\leq \sigma\max_{1\leq i\leq n}\|\boldsymbol{\epsilon}_i\|_2 + \max_{1\leq j\leq n}\|\mathbf{x}_j\|_2\|B^* - \widehat{B}\|_2. \end{aligned}$$

Consider the event

$$\left\{\sigma\max_{1\leq i\leq n}\|\boldsymbol{\epsilon}_i\|_2 \leq \sigma\sqrt{m} + 2\sqrt{\log n}\right\}. \tag{45}$$

By Lemma E.2, event (45) holds with probability at least $1-1/n$. Observe that conditional on the event (45), $\max_{i\in\mathcal{N}^c}\|\mathbf{y}_i-\widehat{B}^\top\mathbf{x}_{\theta^*(i)}\|_2\le\tau_0<\tau$ with τ_0 as defined in Theorem 2.

Finally, we show that for $i\in\mathcal{N}^c$, it holds that $\widehat{\Theta}(\widehat{B})_{i\theta^*(i)}=1$ which then in conjunction with the two previous results implies that $\widehat{\Theta}(\widehat{B})=\Theta^*$. For this purpose, we consider

$$
\begin{aligned}
&\bigcap_{i\in\mathcal{N}^c}\bigcap_{\substack{1\le j\le n\\ j\neq\theta^*(i)}}\Big\{\|\mathbf{y}_i-\widehat{B}^\top\mathbf{x}_{\theta^*(i)}\|_2^2\le\|\mathbf{y}_i-\widehat{B}^\top\mathbf{x}_j\|_2^2\Big\}\\
&=\bigcap_{i\in\mathcal{N}^c}\bigcap_{\substack{1\le j\le n\\ j\neq\theta^*(i)}}\Big\{\|(B^*-\widehat{B})^\top\mathbf{x}_{\theta^*(i)}+\sigma\boldsymbol{\epsilon}_i\|_2^2\le\|B^{*\top}\mathbf{x}_{\theta^*(i)}-\widehat{B}^\top\mathbf{x}_j+\sigma\boldsymbol{\epsilon}_i\|_2^2\Big\}\\
&=\bigcap_{i\in\mathcal{N}^c}\bigcap_{\substack{1\le j\le n\\ j\neq\theta^*(i)}}\Big\{\|(B^*-\widehat{B})^\top\mathbf{x}_{\theta^*(i)}\|_2^2+2\langle(B^*-\widehat{B})^\top\mathbf{x}_{\theta^*(i)},\sigma\boldsymbol{\epsilon}_i\rangle\\
&\qquad\le\|B^{*\top}\mathbf{x}_{\theta^*(i)}-\widehat{B}^\top\mathbf{x}_j\|_2^2+2\langle B^{*\top}\mathbf{x}_{\theta^*(i)}-\widehat{B}^\top\mathbf{x}_j,\sigma\boldsymbol{\epsilon}_i\rangle\Big\}\\
&=\bigcap_{i\in\mathcal{N}^c}\bigcap_{\substack{1\le j\le n\\ j\neq\theta^*(i)}}\Big\{\|(B^*-\widehat{B})^\top\mathbf{x}_{\theta^*(i)}\|_2^2+2\langle(B^*-\widehat{B})^\top\mathbf{x}_{\theta^*(i)},\sigma\boldsymbol{\epsilon}_i\rangle\\
&\qquad\le\|B^{*\top}(\mathbf{x}_{\theta^*(i)}-\mathbf{x}_j)\|_2^2+\|(\widehat{B}-B^*)^\top\mathbf{x}_j\|_2^2+\\
&\qquad\quad+2\langle B^{*\top}(\mathbf{x}_{\theta^*(i)}-\mathbf{x}_j),(B^{*\top}-\widehat{B}^\top)\mathbf{x}_j\rangle+2\langle B^{*\top}\mathbf{x}_{\theta^*(i)}-\widehat{B}^\top\mathbf{x}_j,\sigma\boldsymbol{\epsilon}_i\rangle\Big\}\\
&=\bigcap_{i\in\mathcal{N}^c}\bigcap_{\substack{1\le j\le n\\ j\neq\theta^*(i)}}\Big\{\|(B^*-\widehat{B})^\top\mathbf{x}_{\theta^*(i)}\|_2^2-\|(B^*-\widehat{B})^\top\mathbf{x}_j\|_2^2+\\
&\qquad+2\langle(\widehat{B}-B^*)^\top(\mathbf{x}_j-\mathbf{x}_{\theta^*(i)}),\sigma\boldsymbol{\epsilon}_i\rangle+2\langle B^{*\top}(\mathbf{x}_j-\mathbf{x}_{\theta^*(i)}),\sigma\boldsymbol{\epsilon}_i\rangle+\\
&\qquad+2\langle B^{*\top}(\mathbf{x}_{\theta^*(i)}-\mathbf{x}_j),(B^{*\top}-\widehat{B}^\top)\mathbf{x}_j\rangle\le\|B^{*\top}(\mathbf{x}_{\theta^*(i)}-\mathbf{x}_j)\|_2^2\Big\}\\
&\supseteq\bigcap_{i\in\mathcal{N}^c}\bigcap_{\substack{1\le j\le n\\ j\neq\theta^*(i)}}\Bigg\{\frac{\|(B^*-\widehat{B})^\top\mathbf{x}_{\theta^*(i)}\|_2^2}{\|B^{*\top}(\mathbf{x}_{\theta^*(i)}-\mathbf{x}_j)\|_2^2}+\frac{2\|\sigma\boldsymbol{\epsilon}_i\|_2}{\|B^{*\top}(\mathbf{x}_{\theta^*(i)}-\mathbf{x}_j)\|_2}+\\
&\qquad+\frac{2\|(B^{*\top}-\widehat{B}^\top)\mathbf{x}_j\|_2}{\|B^{*\top}(\mathbf{x}_{\theta^*(i)}-\mathbf{x}_j)\|_2}+\frac{2\|(\widehat{B}-B^*)^\top(\mathbf{x}_j-\mathbf{x}_{\theta^*(i)})\|_2\|\sigma\boldsymbol{\epsilon}_i\|_2}{\|B^{*\top}(\mathbf{x}_{\theta^*(i)}-\mathbf{x}_j)\|_2^2}\le 1\Bigg\}\\
&\supseteq\Bigg\{\left(\frac{\|B^*-\widehat{B}\|_2\max\limits_{1\le i\le n}\|\mathbf{x}_i\|_2}{\min_{i<j}\|B^{*\top}(\mathbf{x}_i-\mathbf{x}_j)\|_2}\right)^2+\frac{2\sigma\max\limits_{1\le i\le n}\|\boldsymbol{\epsilon}_i\|_2}{\min_{i<j}\|B^{*\top}(\mathbf{x}_i-\mathbf{x}_j)\|_2}+\frac{2\|B^*-\widehat{B}\|_2\max\limits_{1\le i\le n}\|\mathbf{x}_i\|_2}{\min_{i<j}\|B^{*\top}(\mathbf{x}_i-\mathbf{x}_j)\|_2}\\
&\qquad+\frac{2\sigma\max_{1\le i\le n}\|\boldsymbol{\epsilon}_i\|_2}{\min_{i<j}\|B^{*\top}(\mathbf{x}_i-\mathbf{x}_j)\|_2}\cdot\frac{2\|B^*-\widehat{B}\|_2\max_{1\le i\le n}\|\mathbf{x}_i\|_2}{\min_{i<j}\|B^{*\top}(\mathbf{x}_i-\mathbf{x}_j)\|_2}\le 1\Bigg\}
\end{aligned}
\tag{46}
$$

Given the event $\mathcal{B}$, we have that

$$
\min_{i<j}\|B^{*\top}(\mathbf{x}_j-\mathbf{x}_i)\|_2=\gamma\|B^*\|_F=\gamma\sigma\sqrt{m}\mathsf{SNR}^{1/2}. \tag{47}
$$

Plugging (47) into (46) and (45), it is easy to verify that under the conditions of the theorem the left hand side of the event in (47) is upper bounded by $1/36+1/3+1/3+1/9<1$ with the stated probability.

We now turn to the converse statement in the regime $m = O(1)$ (second bullet); the converse statement without restriction on m is given subsequently. Let (i_0, j_0) denote the pair of indices such that

$$\|B^{*\top}(\mathbf{x}_{i_0} - \mathbf{x}_{j_0})\|_2^2 = \min_{i<j} \|B^{*\top}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 = \gamma^2 \|B^*\|_F^2,$$

and suppose that $i_0' = \theta^{*-1}(i_0) \neq \emptyset$. For the event $\{\widehat{\Theta}(B^*) = \Theta^*\}$ to hold it is required that

$$\begin{aligned}
&\|\mathbf{y}_{i_0'} - B^{*\top}\mathbf{x}_{i_0}\|_2^2 \leq \|\mathbf{y}_{i_0'} - B^{*\top}\mathbf{x}_{j_0}\|_2^2 \\
\Leftrightarrow &2\langle \sigma \boldsymbol{\epsilon}_{i_0'}, B^{*\top}(\mathbf{x}_{j_0} - \mathbf{x}_{i_0})\rangle \leq \|B^{*\top}(\mathbf{x}_{i_0} - \mathbf{x}_{j_0})\|_2^2 \\
\Leftrightarrow &2\left\langle \sigma \boldsymbol{\epsilon}_{i_0'}, \frac{B^{*\top}(\mathbf{x}_{j_0} - \mathbf{x}_{i_0})}{\|B^{*\top}(\mathbf{x}_{i_0} - \mathbf{x}_{j_0})\|_2} \right\rangle \leq \|B^{*\top}(\mathbf{x}_{i_0} - \mathbf{x}_{j_0})\|_2 \\
\Leftrightarrow &2\left\langle \sigma \boldsymbol{\epsilon}_{i_0'}, \frac{B^{*\top}(\mathbf{x}_{j_0} - \mathbf{x}_{i_0})}{\|B^{*\top}(\mathbf{x}_{i_0} - \mathbf{x}_{j_0})\|_2} \right\rangle \leq \gamma \|B^*\|_F \\
\Leftrightarrow &2\left\langle \sigma \boldsymbol{\epsilon}_{i_0'}, \frac{B^{*\top}(\mathbf{x}_{j_0} - \mathbf{x}_{i_0})}{\|B^{*\top}(\mathbf{x}_{i_0} - \mathbf{x}_{j_0})\|_2} \right\rangle \leq \gamma \sigma \sqrt{m} \mathsf{SNR}^{1/2}
\end{aligned}$$

Note that conditional on $\mathbf{x}_{i_0}, \mathbf{x}_{j_0}$ the left hand side follows a $N(0, 4\sigma^2)$-distribution. It is easy to show that if $g \sim N(0,1)$, $\mathbf{P}(|g| \leq \delta) \leq \delta$ and thus $\mathbf{P}(g > \delta) \geq \frac{1}{2}(1-\delta)$for all $\delta > 0$. Hence if

$$\gamma \mathsf{SNR}^{1/2} < \frac{2}{3}\frac{1}{\sqrt{m}} \Leftrightarrow \gamma^2 \mathsf{SNR} < \frac{4}{9m} =: c, \tag{48}$$

$\widehat{\Theta}(B^*) \neq \Theta^*$ with probability at least $1/3$.

We now turn to the converse statement without restriction on m (first bullet). Note that the event $\{\widehat{\Theta}(B^*) = \Theta^*\}$ implies the event

$$\begin{aligned}
&\bigcap_{i=1}^n \left\{ \|\mathbf{y}_i - B^{*\top}\mathbf{x}_{\theta^*(i)}\|_2^2 \leq \min_{j \neq \theta^*(i)} \|\mathbf{y}_i - B^{*\top}\mathbf{x}_j\|_2^2 \right\} \\
=&\bigcap_{i=1}^n \bigcap_{j \neq \theta^*(i)} \left\{ 2\sigma \left\langle \boldsymbol{\epsilon}_i, B^{*\top}(\mathbf{x}_j - \mathbf{x}_{\theta^*(i)})/\|B^{*\top}(\mathbf{x}_j - \mathbf{x}_{\theta^*(i)})\|_2 \right\rangle \leq \|B^{*\top}\mathbf{x}_{\theta^*(i)} - B^{*\top}\mathbf{x}_j\|_2 \right\} \\
\subseteq&\bigcap_{i=1}^n \left\{ 2\sigma \left\langle \boldsymbol{\epsilon}_i, B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})/\|B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})\|_2 \right\rangle \leq \|B^{*\top}\mathbf{x}_{\theta^*(i)} - B^{*\top}\mathbf{x}_{\eta(i)}\|_2 \right\},
\end{aligned} \tag{49}$$

where $\eta(i) = \theta^*(i) - 1$ if $\theta^*(i) \geq 2$ and $\eta(i) = \theta^*(i) + 1$ otherwise. Now note that conditional on the $\{\mathbf{x}_i\}_{i=1}^n$, the collection

$$\{\langle \boldsymbol{\epsilon}_i, B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})/\|B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})\|_2 \rangle,\ 1 \leq i \leq n\}$$

are i.i.d. $N(0,1)$ random variables. By standard concentration arguments for the maximum of a collection of Gaussian random variables (cf. Ledoux and Talagrand (1991), p. 79), we thus have

$$\mathbf{P}\left(\max_{1 \leq i \leq n} 2\sigma \left\langle \boldsymbol{\epsilon}_i, \frac{B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})}{\|B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})\|_2} \right\rangle < 2\sigma c_0 \sqrt{\log n} \ \middle| \{\mathbf{x}_i\}_{i=1}^n \right) \leq 2/5, \tag{50}$$

for a constant $c_0 > 0$. At the same time, concentration of Lipschitz functions of Gaussian random variables yields

$$\begin{aligned}\mathbf{P}(\|B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})\|_2^2 \geq (1+t)^2 2\|B^*\|_F^2) &\leq \exp\left(-\frac{t^2\|B^*\|_F^2}{2\|B^*\|_2^2}\right) \\ &\leq \exp\left(-\frac{t^2}{2}\right),\ t \geq 0,\ 1 \leq i \leq n.\end{aligned} \tag{51}$$

Let $i_{\max}$ be the index such that

$$\left\langle \boldsymbol{\epsilon}_{i_{\max}}, \frac{B^{*\top}(\mathbf{x}_{\eta(i_{\max})} - \mathbf{x}_{\theta^*(i_{\max})})}{\|B^{*\top}(\mathbf{x}_{\eta(i_{\max})} - \mathbf{x}_{\theta^*(i_{\max})})\|_2} \right\rangle = \max_{1\leq i\leq n} \left\langle \boldsymbol{\epsilon}_i, \frac{B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})}{\|B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})\|_2} \right\rangle$$

Since $\{(B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})/\|B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})\|_2, \|B^{*\top}(\mathbf{x}_{\eta(i)} - \mathbf{x}_{\theta^*(i)})\|_2\}_{i=1}^n$ are pairs of independent random variables, we combine (50) and (51) to conclude that the event $\mathcal{A}_1 \cap \mathcal{A}_2$ occurs with probability at least $1/3$, where

$$\begin{aligned}\mathcal{A}_1 &= \left\{ 2\sigma \left\langle \boldsymbol{\epsilon}_{i_{\max}}, \frac{B^{*\top}(\mathbf{x}_{\eta(i_{\max})} - \mathbf{x}_{\theta^*(i_{\max})})}{\|B^{*\top}(\mathbf{x}_{\eta(i_{\max})} - \mathbf{x}_{\theta^*(i_{\max})})\|_2} \right\rangle > 2c_0\sigma\sqrt{\log n} \right\} \\ \mathcal{A}_2 &= \left\{ \|B^{*\top}(\mathbf{x}_{\eta(i_{\max})} - \mathbf{x}_{\theta^*(i_{\max})})\|_2 \leq \sqrt{18}\|B^*\|_F = \sigma\sqrt{18m}\mathsf{SNR}^{1/2} \right\}.\end{aligned}$$

Combining (49) and the previous display then yields that $\widehat{\Theta}(B^*) \neq \Theta^*$ with the stated probability if

$$\mathsf{SNR} < \frac{4}{18}c_0^2\frac{\log n}{m} =: c'\frac{\log n}{m}.$$

Appendix D. Proof of Proposition 1

By the triangle inequality and the fact that $\Xi^*_{i,:} = 0$ for all $i \in S_*^c$, we have

$$\begin{aligned}\min_{i\in S_*}\|\widehat{\Xi}_{i,:}\|_2 - \max_{i\in S_*^c}\|\widehat{\Xi}_{i,:}\|_2 &\geq \min_{i\in S_*}\|\Xi^*_{i,:}\|_2 - 2\max_{1\leq i\leq n}\|\widehat{\Xi}_{i,:} - \Xi^*_{i,:}\|_2 \\ &\geq \min_{i\in S_*}\|\Xi^*_{i,:}\|_2 - 2\|\widehat{\Xi} - \Xi^*\|_F.\end{aligned} \tag{52}$$

In the sequel, we derive a lower bound on $\min_{i\in S_*}\|\Xi^*_{i,:}\|_2$ in a fashion similar to the previous proof. For any i with $\theta^*(i) = 0$, we have

$$\|\sqrt{n}\Xi^*_{i,:}\|_2 = \|\mathbf{y}_i - B^{*\top}\mathbf{x}_i\|_2 \geq \gamma_0\|B^*\|_F = \gamma_0 \cdot \sigma\sqrt{\mathsf{SNR}}\sqrt{m}. \tag{53}$$

On the other hand, for any i with $\theta^*(i) \notin \{0, i\}$, we have

$$\|\sqrt{n}\Xi^*_{i,:}\|_2 = \|B^{*\top}\mathbf{x}_{\theta^*(i)} - B^{*\top}\mathbf{x}_i\|_2 \geq \gamma\|B^*\|_F = \gamma \cdot \sigma\sqrt{\mathsf{SNR}}\sqrt{m} \tag{54}$$

Combining (52), (53) and (54) yields the assertion.

Appendix E. Auxiliary Results

Lemma E.1 *For any $r \geq 1$, we have the inclusion*

$$\{v \in \mathbb{R}^{n\cdot m} : \ \|v\|_2 \leq 1, \ \|v\|_{2,1} \leq \sqrt{r}\} \subset 2\operatorname{conv} B_0(r), \tag{55}$$

with $\|\cdot\|_{2,1}$ and $B_0(r)$ are defined in (29) *and Lemma A.2, respectively.*

Proof The proof is an adaptation of a standard argument in the sparsity literature, cf. Lemma 3.1 in Plan and Vershynin (2013b). Pick an arbitrary element v contained in the left hand side in (55), and consider subsets $T_\ell \subset \{1, \ldots, n\}$, $|T_\ell| \leq r$, and corresponding vectors $v(T_\ell) \in B_0(r)$ such that

$$(v(T_\ell))_j := \begin{cases} v_j & \text{if } j \in \bigcup_{i\in T_\ell} G_i, \\ 0 & \text{else.} \end{cases}$$

and such that T_1 contains the r indices of $\{1, \ldots, n\}$ corresponding to the r largest norms among $\{\|v^{[i]}\|_2\}_{i=1}^n$, T_2 contains the r indices corresponding to the next r largest norms among $\{\|v^{[i]}\|_2\}_{i=1}^n$, and so forth. Observe that $v = \sum_\ell v(T_\ell)$ and that for any ℓ

$$\|v(T_{\ell+1})\|_{2,\infty} = \max_{i\in T_{\ell+1}} \|v^{[i]}\|_2 \leq \frac{1}{r}\sum_{i\in T_\ell} \|v^{[i]}\|_2 = \frac{1}{r}\|v(T_\ell)\|_{2,1}$$

As a result,

$$\|v(T_{\ell+1})\|_2 \leq \sqrt{r}\|v(T_{\ell+1})\|_{2,\infty} = \frac{1}{\sqrt{r}}\|v(T_\ell)\|_{2,1}.$$

Consequently,

$$\begin{aligned} \sum_\ell \|v(T_\ell)\|_2 &= \|v(T_1)\|_2 + \sum_{\ell\geq 2} \|v(T_\ell)\|_2 \\ &\leq 1 + \frac{1}{\sqrt{r}} \sum_{\ell\geq 1} \|v(T_\ell)\|_{2,1} \\ &\leq 1 + \frac{1}{\sqrt{r}} \sum_{\ell\geq 1} \sum_{i\in T_\ell} \|v_{G_i}\|_2 \\ &\leq 1 + \frac{1}{\sqrt{r}} \|v\|_{2,1} \leq 2. \end{aligned}$$

In conclusion, we have demonstrated that

$$v = \sum_\ell \underbrace{\frac{v(T_\ell)}{\|v(T_\ell)\|_2}}_{\in B_0(r)} \underbrace{\|v(T_\ell)\|_2}_{\lambda_\ell}, \qquad \sum_\ell \lambda_\ell \leq 2,$$

and thus $v \in 2\operatorname{conv} B_0(r)$. Since v was an arbitrary element of the left hand side in (55), the proof is complete. ■

Lemma E.2 *Let $g_\ell \sim N(0, \sigma_\ell^2 I_r)$, $1 \le \ell \le L$, be isotropic Gaussian random vectors. Then:*

$$\mathbf{P}\left(\max_{1\le\ell\le L}\|g_\ell\|_2 > \max_{1\le\ell\le L}\sigma_\ell\{\sqrt{r}+2\sqrt{\log L}\}\right) \le 1/L.$$

Proof We note that $\mathbf{E}[\|g\|_\ell] \le \sigma_\ell\sqrt{r}$, $\ell = 1,\dots,L$, and that the map $x \mapsto \|x\|_2$ is 1-Lipschitz. By concentration of measure of Lipschitz functions of Gaussian random vectors, we hence have

$$\mathbf{P}(\|g_\ell\|_2 \ge \sigma_\ell(\sqrt{r}+2\sqrt{\log L})) \le \exp(-2\log L),\ \ell = 1,\dots,L.$$

The result then follows from a union bound over $\{1,\dots,L\}$. ■

Lemma E.3 (**Gordon's Escape Theorem (Gordon, 1988)**) *Let K be a closed subset of the unit sphere in $\mathbb{R}^p$, let $\nu_r = \mathbf{E}_{g\sim N(0,I_r)}[\|g\|_2]$, and let $\varepsilon \in (0,1)$. If the Gaussian width (cf. §7.5 in Vershynin (2018)) of K obeys $w(K) < (1-\varepsilon)\nu_q - \varepsilon\nu_p$, then a $(p-q)$-dimensional subspace V drawn uniformly from the Grassmannian $\mathsf{G}(p,p-q)$ satisfies*

$$\mathbf{P}(\mathrm{dist}(K,V) > \varepsilon) \ge 1 - \frac{7}{2}\exp\left(-\frac{1}{2}\left(\frac{(1-\varepsilon)\nu_q - \varepsilon\nu_p - w(K)}{3+\varepsilon+\varepsilon\nu_p/\nu_q}\right)^2\right).$$

Appendix F. From Gaussian to sub-Gaussian

In this section, we state and prove a result analogous to Lemma E.3 above for random subspaces V generated by a p-by-$(p-q)$ matrix A with i.i.d. isotropic *sub*-Gaussian rows, i.e., $\mathbf{E}[\langle A_{i,:}, v\rangle^2] = 1$ and $\|\langle A_{i,:}, v\rangle\|_{\psi_2} \le L < \infty$ for all $v \in \mathbb{R}^{p-q}$, $1 \le i \le n$, where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm of a random variable (see, e.g., §2.5 in Vershynin (2018)).

Lemma F.1 *Let $V = \mathrm{range}(A)$ with A as above, and let K be a closed subset of the unit sphere in $\mathbb{R}^p$. For any $\varepsilon, \alpha \in (0,1)$, if*

$$p > \frac{1}{1-\varepsilon^2}\frac{2(p-q)+C_1L^4\cdot w^2(K)}{(1-\alpha)^2} \vee \frac{C}{\alpha^2}\{(p-q)\vee\log p\} \tag{56}$$

then $\mathbf{P}(\mathrm{dist}(K,V) > \varepsilon) \ge 1-2\big(\exp(-w^2(K))+\exp(-c\{(p-q)\vee\log p\})\big)$, where $C_1, C_2, c > 0$ are universal constants depending only on L.

It is worth noting that the condition (56) is comparable to the condition in Lemma E.3 which after term simplifications becomes $p \gtrsim \frac{1}{1-\varepsilon^2}((p-q)+w^2(K))$, which corresponds to the first (and leading) term on the right hand side of (56).

Proof Let $V^\perp$ denote the orthogonal complement of V in $\mathbb{R}^p$, respectively. Accordingly, denote by P_V and $\mathsf{P}_{V^\perp}$ the orthoprojectors on V and $V^\perp$, respectively. Note that

$$\mathrm{dist}^2(K,V) = \inf_{\xi\in K}\|\mathsf{P}_{V^\perp}\xi\|_2^2 = 1 - \sup_{\xi\in K}\|\mathsf{P}_V\xi\|_2^2. \tag{57}$$

Hence in order to lower bound $\mathrm{dist}^2(K,V)$, it suffices to upper bound $\sup_{\xi\in K}\|\mathsf{P}_V\xi\|_2^2$. Assuming for now that A is non-singular, we have

$$\begin{aligned}\sup_{\xi\in K}\|\mathsf{P}_V\xi\|_2^2 &= \sup_{\xi\in K}\xi^\top A(A^\top A)^{-1}A^\top\xi \\ &\le \sup_{\xi\in K}\|(A^\top A)^{-1/2}A^\top\xi\|_2^2 \\ &\le \|(A^\top A)^{-1/2}\|_2^2\ \sup_{\xi\in K}\|A^\top\xi\|_2^2 \le \frac{1}{\sigma_{\min}(A)^2}\sup_{\xi\in K}\|A^\top\xi\|_2^2. \end{aligned} \tag{58}$$

In order to bound the second factor on the right hand side, we invoke the following result:

Lemma F.2 *(cf. Exercise 9.1.8 in Vershynin (2018)). Let A, L, and K be as above. Then for any $u\ge 0$, the following event occurs with probability at least $1-2\exp(-u^2)$:*

$$\sup_{\xi\in K}\left|\|A^\top\xi\|_2-\sqrt{p-q}\right|\le CL^2(w(K)+u).$$

Invoking the above lemma with the choice $u=w(K)$, we obtain that

$$\mathbf{P}\left(\sup_{\xi\in K}\|A^\top\xi\|_2\le\sqrt{p-q}+C'L^2w(K)\right)\ge 1-2\exp(-w^2(K)). \tag{59}$$

At the same time, concentration results (Vershynin, 2012, Theorem 5.35) on the minimum singular value of random matrices with sub-Gaussian rows yield that for any $\alpha\in(0,1)$

$$\mathbf{P}(\sigma_{\min}(A)^2\ge(1-\alpha)^2p)\ge 1-2\exp(-c\{(p-q)\vee\log p\}) \tag{60}$$

provided that $p\ge\frac{C}{\alpha^2}\{(p-q)\vee\log p\}$ for positive constants $c=c_L$ and $C=C_L$ depending only on the sub-Gaussian norm L of the rows of A. Combining (57), (58), (59) and (60), we obtain that with the probability stated in the theorem, it holds that

$$\inf_{\xi\in K}\|\mathsf{P}_{V^\perp}\xi\|_2^2\ge 1-\frac{2(p-q)+C''L^4w^2(K)}{(1-\alpha)^2p}\ge\varepsilon^2$$

as long as $p>\frac{1}{1-\varepsilon^2}\frac{2(p-q)+C''L^4\cdot w^2(K)}{(1-\alpha)^2}$ for any $\varepsilon\in(0,1)$, which concludes the proof. ■

Appendix G. Conditional gradient method for optimization of (13) & (14)

We start with optimization problem (14). Let

$$f(\Theta):=\frac{1}{2n\cdot m}\|\mathsf{P}_X^\perp\Theta Y\|_F^2,\qquad \nabla f(\Theta)=\frac{1}{n\cdot m}\mathsf{P}_X^\perp\Theta YY^\top$$

be the objective and gradient, respectively, of (14). Following Algorithm 1 in Jaggi (2013), the conditional gradient (Frank-Wolfe) updates for minimizing f over $\mathcal{C}_k:=\{\Theta\in\mathcal{C}:\sum_{i=1}^n\Theta_{ii}\ge n-k\}$ with $\mathcal{C}$ defined in (12) are given as follows.

Algorithm 2 Frank-Wolfe method for minimizing (14)

Initialize $\Theta^{(0)} = I_n$.
Repeat for $t = 0, 1, \ldots$

$$D^{(t)} \leftarrow \operatorname*{argmin}_{\Theta \in \mathcal{C}_k} \operatorname{tr}(\Theta^\top \nabla f(\Theta^{(t)})), \qquad \Theta^{(t+1)} \leftarrow (1-\alpha^{(t)})\Theta^{(t)} + \alpha^{(t)} D^{(t)},$$

where $\alpha^{(t)} = \operatorname{argmin}_{\alpha > 0} f((1-\alpha^{(t)})\Theta^{(t)} + \alpha D^{(t)}) = -\frac{\operatorname{tr}(\mathsf{P}_X^\perp D^{(t)} Y Y^\top \Theta^{(t)\top})}{\operatorname{tr}(\mathsf{P}_X^\perp D^{(t)} Y Y^\top D^{(t)\top})}$.

The dominant computational cost in the above algorithm is incurred for the argmin over $\mathcal{C}_k$, which requires the solution of a linear program with n^2 variables and $O(n)$ linear constraints.

A similar algorithm can be applied for optimization problem (13). An additional complication arises from the penalty in (13) which renders the objective non-smooth. As a workaround, we apply the above Frank-Wolfe scheme to a successively smoothed objective (Nesterov, 2005).