D4R: COMMUNICATION-EFFICIENT ROBUST REDUCED-RANK REGRESSION FOR DISTRIBUTED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale multivariate data are often distributed across multiple machines and contaminated by outliers, which complicates efficient and interpretable modeling of shared low-rank structure. To address both challenges, we propose a **D**istributed **R**obust **R**educed-**R**ank **R**egression (D4R) framework based on a refined surrogate loss. D4R separates local outlier correction from global low-rank estimation while reducing both computation and communication costs. Theoretically, we establish non-asymptotic error bounds showing that D4R reaches the target statistical rate after finite communication rounds. Simulations and real-data studies demonstrate the robustness and efficiency of our method.

1 INTRODUCTION

Multivariate response data are ubiquitous in modern statistical applications, ranging from gene expression analysis in genomics (Zhi et al., 2013) to asset allocation in finance (Senneret et al., 2016). In the big data era, such datasets are often partitioned across multiple local agents or devices due to constraints such as data privacy, transmission bandwidth, or storage limitations (Kairouz & McMahan, 2021), and frequently contain observations contaminated by measurement errors, outliers, or adversarial corruption. Examples include anomalous measurements in wireless sensor networks (López-Valcarce et al., 2016), data contamination in distributed multi-sensor systems (Jiang et al., 2024), and local anomalies in decentralized multivariate time-series data (Le et al., 2025). A classical and widely studied approach to model such contamination is the mean-shift formulation in robust multivariate regression (She & Chen, 2017). Specifically, we consider the following model

$$\boldsymbol{y}_i = \boldsymbol{B}^{*\top}\boldsymbol{x}_i + \sqrt{N}\boldsymbol{c}_i^* + \boldsymbol{\epsilon}_i,\ i = 1, \ldots, N \tag{1}$$

where $\boldsymbol{B}^* \in \mathbb{R}^{p\times q}$ is the true coefficient matrix, $\boldsymbol{c}_i^* \in \mathbb{R}^q$ captures possible adversarial contamination, and $\boldsymbol{\epsilon}_i \in \mathbb{R}^q$ is the random noise vector.

Let $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]^\top \in \mathbb{R}^{N\times q}$ and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N\times p}$. Define the contamination matrix $\boldsymbol{C}^* = [\boldsymbol{c}_1^*, \ldots, \boldsymbol{c}_N^*]^\top \in \mathbb{R}^{N\times q}$ and the noise matrix $\boldsymbol{\mathcal{E}} = [\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N]^\top \in \mathbb{R}^{N\times q}$. Model 1 can then be equivalently written in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^* + \sqrt{N}\boldsymbol{C}^* + \boldsymbol{\mathcal{E}} = [\boldsymbol{X}\ \sqrt{N}\boldsymbol{I}]\begin{pmatrix}\boldsymbol{B}^*\\ \boldsymbol{C}^*\end{pmatrix} + \boldsymbol{\mathcal{E}}, \tag{2}$$

where the scaling factor $\sqrt{N}$ normalizes the augmented design so that the columns of $[\boldsymbol{X}\ \sqrt{N}\boldsymbol{I}]$ have comparable scales. Due to the overparameterization of the above formulation, meaningful estimation requires structural assumptions. We assume that the coefficient matrix $\boldsymbol{B}^*$ is low-rank, reflecting that the q responses are driven by a small number of shared latent factors. The low-rank paradigm has found widespread application across many fields, including compressed sensing (Fan et al., 2019), image recovery (Lu et al., 2019), and sensor localization (Nguyen et al., 2019). In addition, we assume that $\boldsymbol{C}^*$ is row-sparse because contamination usually affects only a small fraction of observations (Liu et al., 2003; Huber, 2011). Similar row-sparse structures appear in robust principal component analysis (Hubert et al., 2019) and low-rank approximation with outlying observations (Bhaskara & Kumar, 2018).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Although the contamination model 2 has been widely used in robust multivariate regression, existing theoretical analyses have mainly focused on the joint estimation or prediction error of $(\boldsymbol{B}^*, \boldsymbol{C}^*)$; see, for example, She & Chen (2017). This is primarily because the regression signal $\boldsymbol{X}\boldsymbol{B}^*$ and the contamination component $\boldsymbol{C}^*$ are coupled, creating a nontrivial identification issue that makes analyzing them individually challenging. In many applications, however, the principal inferential target is the low-rank coefficient matrix $\boldsymbol{B}^*$, whereas $\boldsymbol{C}^*$ serves as a nuisance component that captures sample-wise outlying effects. Therefore, establishing estimation guarantees for $\boldsymbol{B}^*$ itself is both statistically more informative and practically more relevant. To the best of our knowledge, such guarantees for $\boldsymbol{B}^*$ under adversarial contamination remain largely unexplored.

Beyond the identification issue arising from the coupling between $\boldsymbol{B}^*$ and $\boldsymbol{C}^*$, the distributed setting introduces additional statistical and computational challenges. A naive solution is a one-shot divide-and-conquer strategy (Lee et al., 2017), which fits the model 2 separately on each machine and averages the resulting estimators. However, this strategy can be suboptimal: limited local samples and contamination may bias the local estimates, while averaging neither removes this bias nor preserves low-rank structure (Chen et al., 2022b). Another natural approach is to apply the communication-efficient surrogate likelihood (CSL) framework (Wang et al., 2017; Jordan et al., 2019). However, standard CSL is not directly applicable to the contamination model 2 because the local and global curvature terms associated with the contamination component have incompatible structures and scales. A detailed discussion is provided in Appendix D. Therefore, developing a communication-efficient distributed framework under data corruption remains a nontrivial and underexplored problem.

To address these challenges, we propose D4R, a unified framework for robust multivariate regression with distributed and contaminated data. Our main **contributions** are summarized as follows: 1) A communication-efficient distributed framework for robust reduced-rank regression is proposed. This approach exploits the underlying low-rank structure and accommodates outliers in decentralized settings. 2) We develop a refined surrogate-based distributed algorithm that separates local contamination adjustment from global low-rank coefficient estimation. This design overcomes the limitations of standard CSL under the contamination model and achieves communication- and computation-efficient estimation. 3) A rigorous non-asymptotic analysis is provided for the proposed estimator, demonstrating that it attains the desired statistical error rate within a finite number of communication rounds.

Notations. For a positive integer d, let $[d] = \{1, \ldots, d\}$. For nonnegative quantities a and b, we write $a \lesssim b$ (resp. $a \gtrsim b$) if $a \leq Cb$ (resp. $a \geq Cb$) for some positive constant C independent of the problem dimensions, and $a \asymp b$ if both $a \lesssim b$ and $a \gtrsim b$ hold. The notation $\mathcal{O}(b)$ denotes a quantity whose absolute value is bounded by cb for some universal constant $c > 0$. We write $a \vee b = \max\{a, b\}$. For a vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|_2$ denotes its Euclidean norm. For a matrix $\boldsymbol{A}$, $\boldsymbol{A}_{i\cdot}$ denotes the i-th row of $\boldsymbol{A}$, and for an index set $\mathcal{J}$, $\boldsymbol{A}_{\mathcal{J}\cdot}$ denotes the submatrix formed by the rows indexed by $\mathcal{J}$. Moreover, $\|\boldsymbol{A}\|_F$, $\mathrm{rank}(\boldsymbol{A})$, and $\mathrm{tr}(\boldsymbol{A})$ denote its Frobenius norm, rank, and trace, respectively. Let $\|\boldsymbol{A}\|_{2,0} = |\{i : \|\boldsymbol{A}_{i\cdot}\|_2 \neq 0\}|$ be the number of nonzero rows of $\boldsymbol{A}$. For conformable matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, let $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B})$. Let $\boldsymbol{I}_d$ and $\boldsymbol{0}_{a\times b}$ denote the $d \times d$ identity matrix and the $a \times b$ zero matrix, respectively. For a full-column-rank matrix $\boldsymbol{A}$, define $P_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top$. For a symmetric matrix $\boldsymbol{A}$, $\Lambda_{\min}(\boldsymbol{A})$ and $\Lambda_{\max}(\boldsymbol{A})$ denote its smallest and largest eigenvalues. We use $\mathbf{1}(\cdot)$ for the indicator function and $\|\cdot\|_{\psi_2}$ for the sub-Gaussian Orlicz norm. Let c, c', and C be positive constants whose values may vary from time to time.

2 Related Work

Low-rank Modeling. The low-rank assumption is widely adopted in multivariate modeling to reduce model complexity and improve statistical efficiency (Negahban & Wainwright, 2012; Cho & Park, 2022; Zhu et al., 2022). Reduced-rank regression (Anderson, 1951) achieves this by restricting the rank of the coefficient matrix; see Trendafilov et al. (2025) for a recent overview. The nuclear norm and Schatten p-norms are also commonly employed to promote low-rank structure in the coefficient matrix (Nie et al., 2012; Wang et al., 2025). However, these methods can be sensitive to outliers, which may distort coefficient estimation and obscure the underlying low-rank structure (Guo & Lin, 2018).

Adversarial Contamination. In centralized settings, robust losses such as the Huber loss are widely used to mitigate the influence of contaminated observations (Huber, 1973). Motivated by the connection between robust losses and sparse mean-shift formulations (Donoho & Montanari, 2016), She & Chen (2017) developed a robust reduced-rank regression framework that represents contamination through a sparse additive response component, enabling simultaneous low-rank estimation and outlier detection. Related robust low-rank methods have been developed for trace regression, matrix completion, and sparse reduced-rank regression (Sasai & Fujisawa, 2020; Thompson, 2020; Tan et al., 2023). However, these methods rely on centralized data access and do not address settings in which contaminated observations are distributed across multiple machines.

Communication-Efficient Distributed Learning. Communication-efficient distributed learning aims to approximate full-sample estimators without pooling raw data on a central machine (Wang et al., 2017; Lee et al., 2017). A representative approach is communication-efficient surrogate likelihood, which constructs a central surrogate objective using local curvature information and aggregated gradients from distributed machines (Jordan et al., 2019). In the context of reduced-rank regression, Chen et al. (2022a) proposed an iterative distributed algorithm and established oracle-rate guarantees after finitely many communication rounds. Extending these surrogate-based methods to robust low-rank estimation, however, is nontrivial due to the simultaneous presence of low-rank structure and data contamination.

Robust Distributed Learning. Existing robust distributed learning methods mainly rely on robust losses or robust aggregation. For vector-valued models, distributed procedures have been developed under data contamination (Luo et al., 2022; Ma et al., 2025; Liu et al., 2026). For matrix-valued models, related methods employ quantile loss for reduced-rank regression (Chen et al., 2022b) or adaptive Huber loss for matrix regression (Wang et al., 2025), primarily targeting heavy-tailed errors. Another line of work addresses Byzantine machines through robust aggregation (Chen et al., 2017; Yin et al., 2018; Tu et al., 2023). These methods focus on heavy-tailed noise or machine-level corruption rather than sample-wise adversarial contamination, under which a small fraction of observations may be arbitrarily corrupted. Our work instead develops a distributed robust reduced-rank regression framework for adversarially contaminated data.

3 D4R FRAMEWORK

In this section, we develop the **D**istributed **R**obust **R**educed-**R**ank **R**egression (D4R) framework for large-scale multivariate regression under data contamination. Throughout this paper, we assume that the true coefficient matrix $\boldsymbol{B}^*$ satisfies $\mathrm{rank}(\boldsymbol{B}^*) \leq r$. The full dataset $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ is distributed across m machines, with the k-th machine storing $(\boldsymbol{X}_{\mathcal{I}_k\cdot}, \boldsymbol{Y}_{\mathcal{I}_k\cdot})$, where $|\mathcal{I}_k| = n$, $\cup_{k=1}^m \mathcal{I}_k = \{1, \ldots, N\}$ and $\mathcal{I}_{k_1} \cap \mathcal{I}_{k_2} = \emptyset$ for $k_1 \neq k_2$. Suppose the contamination at the k-th machine is o_k-row-sparse, meaning that $\|\boldsymbol{C}^*_{\mathcal{I}_k\cdot}\|_{2,0} = o_k$ for some $0 \leq o_k < n$. Let $O = \sum_{k=1}^m o_k$ be the total number of corrupted observations.

If all data were centrally available, the full-sample empirical loss would be

$$\mathcal{L}_{\mathrm{full}}(\bar{\boldsymbol{B}}) = \frac{1}{2N}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \sqrt{N}\boldsymbol{C}\|_F^2, \tag{3}$$

where $\bar{\boldsymbol{B}} = [\boldsymbol{B}^\top, \boldsymbol{C}^\top]^\top \in \mathbb{R}^{(p+N)\times q}$. Since pooling the raw data is costly, we construct a communication-efficient surrogate for this objective.

Without loss of generality, we designate machine 1 as the central machine. It exchanges summary information with the other $m-1$ machines. The model of the central machine is

$$\boldsymbol{Y}_{\mathcal{I}_1\cdot} = \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^* + \sqrt{N}\boldsymbol{C}^*_{\mathcal{I}_1\cdot} + \boldsymbol{\mathcal{E}}_{\mathcal{I}_1\cdot}, \tag{4}$$

with the corresponding local empirical loss

$$\mathcal{L}_1(\bar{\boldsymbol{B}}) = \frac{1}{2n}\|\boldsymbol{Y}_{\mathcal{I}_1\cdot} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B} - \sqrt{N}\boldsymbol{C}_{\mathcal{I}_1\cdot}\|_F^2. \tag{5}$$

It is natural to consider the CSL method (Jordan et al., 2019) to estimate $\bar{\boldsymbol{B}}^* = [\boldsymbol{B}^{*\top}\ \boldsymbol{C}^{*\top}]^\top$ by approximating the higher-order derivatives of $\mathcal{L}_{\mathrm{full}}$ with those of $\mathcal{L}_1$. However, for vector-valued

models under data contamination, the standard CSL construction suffers from fundamental structural and scaling mismatches between the high-order derivatives of $\mathcal{L}_{\text{full}}$ and $\mathcal{L}_1$ (Liu et al., 2026). The same problem arises here due to the coupled parameter structure of $\boldsymbol{B}$ and $\boldsymbol{C}$. Appendix D gives a detailed comparison. We therefore use the following refined block-diagonal surrogate Hessian, which removes the structural and scaling mismatches induced by the contamination component:

$$\bar{\boldsymbol{H}} = \begin{pmatrix} \frac{\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top \boldsymbol{X}_{\mathcal{I}_1\cdot}}{n} & \mathbf{0}_{p\times N} \\ \mathbf{0}_{N\times p} & \boldsymbol{I}_N \end{pmatrix}. \tag{6}$$

A detailed justification is provided in Appendix D. This gives the following surrogate loss

$$\begin{aligned} \tilde{\mathcal{L}}(\bar{\boldsymbol{B}}) =& \mathcal{L}_{\text{full}}(\bar{\boldsymbol{B}}^-) + \langle \nabla \mathcal{L}_{\text{full}}(\bar{\boldsymbol{B}}^-), \bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^- \rangle + \frac{1}{2}\text{tr}\{(\bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-)^\top \bar{\boldsymbol{H}} (\bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-)\} \\ =& -\frac{1}{N}\text{tr}\left((\boldsymbol{X}\boldsymbol{B} + \sqrt{N}\boldsymbol{C})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^- - \sqrt{N}\boldsymbol{C}^-)\right) \\ &+ \frac{1}{2n}\|\boldsymbol{X}_{\mathcal{I}_1\cdot}(\boldsymbol{B} - \boldsymbol{B}^-)\|_F^2 + \frac{1}{2}\|\boldsymbol{C} - \boldsymbol{C}^-\|_F^2 + \text{const}, \end{aligned} \tag{7}$$

where const collects all terms independent of the current parameter $\bar{\boldsymbol{B}}$. The block-diagonal matrix $\bar{\boldsymbol{H}}$ removes the quadratic cross-term between $\boldsymbol{B}$ and $\boldsymbol{C}$, making the surrogate separable with respect to the two parameter blocks. This separability enables each machine to update its local contamination parameters $\boldsymbol{C}_{\mathcal{I}_k\cdot}$ in parallel, while allowing the central machine to update $\boldsymbol{B}$ without requiring synchronous estimates of $\boldsymbol{C}$. In contrast to vector-valued models, where simple scalar thresholding suffices, optimizing $\boldsymbol{B}$ and $\boldsymbol{C}$ poses nontrivial algorithmic challenges, namely, enforcing the low-rank structure of $\boldsymbol{B}$ and the row-wise sparsity of $\boldsymbol{C}$.

Building on the surrogate formulation, we now introduce the corresponding distributed optimization framework and algorithmic procedure. The optimization with respect to the row-sparse contamination matrix $\boldsymbol{C}$ is fully local. To facilitate this, define the row-wise hard-thresholding operator $\mathcal{H}_\lambda : \mathbb{R}^{v\times q} \to \mathbb{R}^{v\times q}$

$$(\mathcal{H}_\lambda(\boldsymbol{A}))_{i\cdot} := \boldsymbol{A}_{i\cdot} \times \mathbf{1}\left(\|\boldsymbol{A}_{i\cdot}\|_2 \geq \lambda\right), \tag{8}$$

for $\boldsymbol{A} \in \mathbb{R}^{v\times q}$ and $i \in [v]$. Accordingly, each machine updates its local contamination matrix via

$$\boldsymbol{C}_{\mathcal{I}_k\cdot}^{(t-1)} = \mathcal{H}_{\lambda_k^{(t-1)}}\left(\frac{\boldsymbol{Y}_{\mathcal{I}_k\cdot} - \boldsymbol{X}_{\mathcal{I}_k\cdot}\boldsymbol{B}^{(t-1)}}{\sqrt{N}}\right). \tag{9}$$

From an optimization perspective, this hard-thresholding update can be interpreted through the $\boldsymbol{C}$-block of the surrogate loss in equation 7. Specifically, with $\boldsymbol{B}$ fixed at $\boldsymbol{B}^{(t-1)}$, the $\boldsymbol{C}$-dependent part of the surrogate in equation 7, together with the local $\ell_{2,0}$ penalty, give the following subproblem on the k-th machine

$$\min_{\boldsymbol{C}_{\mathcal{I}_k\cdot}} \frac{1}{2}\left\|\boldsymbol{C}_{\mathcal{I}_k\cdot} - \frac{\boldsymbol{Y}_{\mathcal{I}_k\cdot} - \boldsymbol{X}_{\mathcal{I}_k\cdot}\boldsymbol{B}^{(t-1)}}{\sqrt{N}}\right\|_F^2 + \frac{(\lambda_k^{(t-1)})^2}{2}\|\boldsymbol{C}_{\mathcal{I}_k\cdot}\|_{2,0}. \tag{10}$$

Hence, applying $\mathcal{H}_{\lambda_k^{(t-1)}}$ is not an ad hoc heuristic, but corresponds to an exact proximal update associated with this localized subproblem. The regularization parameters $\{\lambda_k^{(t-1)}\}_{k\in[m]}$ are adaptively updated across communication rounds, with their explicit expressions specified in Theorem 4. The practical choices are discussed in Section 5.

On the other hand, to efficiently solve for $\boldsymbol{B}^{(t)}$ on the central machine, we rewrite the surrogate loss $\tilde{\mathcal{L}}$ as a quadratic objective in $\boldsymbol{B}$, yielding the following equivalent optimization problem

$$\min_{\text{rank}(\boldsymbol{B})\leq r} Q(\boldsymbol{B}; \boldsymbol{B}^{(t-1)}) := \frac{1}{2n}\|\boldsymbol{M}_{\boldsymbol{B}}^{(t-1)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}\|_F^2, \tag{11}$$

where the pseudo-response matrix $\boldsymbol{M}_{\boldsymbol{B}}^{(t-1)}$ is defined as

$$\boldsymbol{M}_{\boldsymbol{B}}^{(t-1)} := \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t-1)} + \boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}\left[\frac{1}{N}\boldsymbol{X}^\top\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^{(t-1)} - \sqrt{N}\boldsymbol{C}^{(t-1)}\right)\right], \tag{12}$$

and $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n}\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top \boldsymbol{X}_{\mathcal{I}_1\cdot}$ represents the local sample covariance matrix of the central machine. Assuming that $\hat{\boldsymbol{\Sigma}}_1$ is invertible, a global minimizer of equation 11 admits a closed-form solution (Reinsel & Velu, 1998). Consequently, the update rule for $\boldsymbol{B}^{(t)}$ is given by

$$\boldsymbol{B}^{(t)} = \left(\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top \boldsymbol{X}_{\mathcal{I}_1\cdot}\right)^{-1} \boldsymbol{X}_{\mathcal{I}_1\cdot}^\top \boldsymbol{M}_{\boldsymbol{B}}^{(t-1)} P_{V(\boldsymbol{X}_{\mathcal{I}_1\cdot}, \boldsymbol{M}_{\boldsymbol{B}}^{(t-1)}, r)}, \tag{13}$$

where $V(\boldsymbol{X}_{\mathcal{I}_1\cdot}, \boldsymbol{M}_{\boldsymbol{B}}^{(t-1)}, r)$ is formed by the leading r eigenvectors of $\boldsymbol{M}_{\boldsymbol{B}}^{(\boldsymbol{t-1})\top} P_{\boldsymbol{X}_{\mathcal{I}_1\cdot}} \boldsymbol{M}_{\boldsymbol{B}}^{(\boldsymbol{t-1})}$.

The complete D4R procedure is provided in Algorithm 1 of Appendix A. As shown in Corollary 5, D4R attains the desired statistical accuracy within a constant number of communication rounds. Therefore, its computational complexity is $\mathcal{O}\left(Npq + n^2q + p^2q + p^3 + Np^2\right)$. This is substantially lower than the full-sample robust reduced-rank regression algorithm of She & Chen (2017), whose computational complexity is $\mathcal{O}\left(Npq + N^2q + p^2q + p^3 + Np^2\right)$. Moreover, when $p \asymp q$, the additional robustification cost $\mathcal{O}(Npq)$ is of the same order as $\mathcal{O}(Np^2)$. Hence, under this regime, our algorithm has the same computational complexity as the traditional distributed reduced-rank regression method (Chen et al., 2022a). These comparisons demonstrate that the proposed distributed framework achieves robustness without introducing additional leading-order computational cost, and is therefore computationally efficient.

4 Theoretical Analysis

In this section, we study the theoretical properties of $\boldsymbol{B}^{(T)}$. We begin by introducing the assumptions required for our analysis.

Assumption 1. *Assume the design matrix $\boldsymbol{X} \in \mathbb{R}^{N\times p}$ is rowwise independent and sub-Gaussian: each row $\boldsymbol{X}_{i\cdot} \overset{d}{=} \boldsymbol{Z}_{i\cdot}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{1\times p}$, where $\boldsymbol{Z}_{i\cdot} = (Z_{i1}, \cdots, Z_{ip}) \in \mathbb{R}^{1\times p}$ and each $\boldsymbol{Z}_{i\cdot}$ is i.i.d. centered 1-sub-Gaussian vectors such that $\mathbb{E}(\boldsymbol{Z}_{i\cdot}^\top \boldsymbol{Z}_{i\cdot}) = \boldsymbol{I}_p$. The covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ satisfies*

$$C_L \le \Lambda_{\min}(\boldsymbol{\Sigma}) \le \Lambda_{\max}(\boldsymbol{\Sigma}) \le C_U, \tag{14}$$

where $C_U \ge C_L > 0$ are two constants.

The noise matrix $\boldsymbol{\mathcal{E}} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N]^\top = \{e_{ij}\}_{i\in[N], j\in[q]} \in \mathbb{R}^{N\times q}$ has independent, centered, σ-sub-Gaussian entries. That is, for all $i \in [N], j \in [q]$, $\mathbb{E}[e_{ij}] = 0$ and $\|e_{ij}\|_{\psi_2} \le \sigma$. Furthermore, $\boldsymbol{\mathcal{E}}$ is independent of $\boldsymbol{X}$.

Assumption 2. *The full sample size $N = n \times m$ satisfies*

$$N \ge C\big(m(p + \log n) + r(p + q) + O\log N\big), \tag{15}$$

where $C > 0$ is a sufficiently large constant that depends only on C_L and C_U.

These assumptions are standard in high-dimensional distributed learning. Assumption 1 specifies sub-Gaussian tails for the design and noise and requires the population covariance to be well-conditioned (Wang et al., 2021; Chen et al., 2022a; Liu et al., 2026). Assumption 2 gives the sample-size requirement, which is mild (Jordan et al., 2019). Under these assumptions, we first derive a one-round error bound.

Proposition 3. *Suppose Assumptions 1 and 2 hold. Let $e_{\boldsymbol{B}}^{(0)} := \|\boldsymbol{B}^* - \boldsymbol{B}^{(0)}\|_F$. For each machine $k \in [m]$, choose*

$$\lambda_k^{(0)} \asymp \sqrt{\frac{p}{N(o_k \vee 1)} + \frac{\log N}{N}} e_{\boldsymbol{B}}^{(0)} + \sigma\frac{\sqrt{q} + \sqrt{\log N}}{\sqrt{N}}. \tag{16}$$

Then the estimator $\boldsymbol{B}^{(1)}$ after the first communication round satisfies

$$\|\boldsymbol{B}^{(1)} - \boldsymbol{B}^*\|_F \lesssim \Big(\sqrt{\frac{p + \log n}{n}} + \frac{O\log N}{N}\Big) e_{\boldsymbol{B}}^{(0)} + \sigma\Big(\sqrt{\frac{r(p + q + \log(2/\epsilon))}{N}} + \frac{qO + O\log N}{N}\Big), \tag{17}$$

with probability at least $1 - \epsilon - c'n^{-c}$, where $c, c' > 0$ are absolute constants.

Proposition 3 decomposes the one-round estimation error into two terms. The first term reflects the contraction of the initialization error $e_{\boldsymbol{B}}^{(0)}$, with the contraction factor determined by the local sample size and the contamination level. The second term represents the intrinsic statistical error, consisting of the low-rank estimation error and the additional error induced by data contamination. Therefore, as the number of communication rounds increases, the influence of the initialization error progressively diminishes, and the intrinsic statistical error becomes dominant.

Building on Proposition 3, we next extend the one-round result to multiple communication rounds and establish an estimation error bound for $\boldsymbol{B}^{(T)}$.

Theorem 4. *Suppose Assumptions 1 and 2 hold. For every machine $k \in [m]$ and round $t \geq 0$, choose*

$$\lambda_k^{(t)} \asymp \sqrt{\frac{p}{N(o_k \vee 1)} + \frac{\log N}{N}} \left(A \left(\sqrt{\frac{p + \log n}{n}} + \frac{O \log N}{N} \right) \right)^t e_{\boldsymbol{B}}^{(0)} + \sigma \frac{\sqrt{q} + \sqrt{\log N}}{\sqrt{N}} + \sigma \sqrt{\frac{p}{N(o_k \vee 1)} + \frac{\log N}{N}} \left(\sqrt{\frac{r(p+q+\log(2/\epsilon))}{N}} + \frac{O(q + \log N)}{N} \right), \tag{18}$$

where A is a constant depending only on C_L and C_U. Then, after $T \asymp \frac{\log(e_{\boldsymbol{B}}^{(0)}/\sigma) + \log N - \log(\sqrt{Nr(p+q)} + O \log N + qO)}{\log(N) - \log\left(\sqrt{Nm(p+\log n)} + O \log N\right)}$ communication rounds, the estimator satisfies

$$\|\boldsymbol{B}^{(T)} - \boldsymbol{B}^*\|_F \lesssim \sigma \left(\sqrt{\frac{r(p+q+\log(2/\epsilon))}{N}} + \frac{O(q + \log N)}{N} \right), \tag{19}$$

with probability at least $1 - \epsilon - c' n^{-c}$, where $c, c' > 0$ are absolute constants.

The error bound in Theorem 4 consists of two leading terms. The first term $\sigma\sqrt{r(p+q)/N}$ corresponds to the oracle statistical rate for the reduced-rank regression (Ma et al., 2020). The second term $\sigma O(q + \log N)/N$ is the additional error induced by adversarial contamination, which increases linearly with the contamination proportion O/N. The resulting dependence on both the rank and contamination proportion is consistent with the trends in Figure E.1 in Appendix E.1.

Corollary 5. *If there exist constants $\alpha_1 > 0$ and $\alpha_2 \in (0, 1)$ such that*

$$e_{\boldsymbol{B}}^{(0)}/\sigma \leq N^{\alpha_1}, \max\{mp, O\} \leq N^{\alpha_2}, \tag{20}$$

then the required number of communication rounds T is bounded by an absolute constant.

Thus, D4R attains the error rate in Theorem 4 after only a constant number of communication rounds.

5 Numerical Experiments

5.1 Simulation Studies

Experimental setup. We generate data from model 2 using $m = 20$ machines, each with $n = 200$ local samples, yielding a total sample size of $N = mn = 4000$. We set $p = 15$, $q = 40$, and $\mathrm{rank}(\boldsymbol{B}^*) = 3$. The coefficient matrix is generated by $\boldsymbol{B}^* = a\boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{p \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{q \times r}$ have i.i.d. $\mathcal{N}(0, 1)$ entries and $a = 1$. The rows of $\boldsymbol{X}$ are independently drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, with $\Sigma_{ij} = 0.5^{|i-j|}$. For contamination, we randomly select an index set $\mathcal{O} \subset [N]$ with $|\mathcal{O}|/N = 0.1$, set $\boldsymbol{C}^*_{\mathcal{O}^c,\cdot} = \boldsymbol{0}$, and draw the entries of $\boldsymbol{C}^*_{\mathcal{O},\cdot}$ independently from $\mathcal{N}(0, 0.5^2)$. We consider four noise distributions: (1) standard normal, $\mathcal{N}(0, 1)$; (2) Laplace with location 0 and scale $1/\sqrt{2}$; (3) scaled Student's t, $t_3/\sqrt{3}$; and (4) Cauchy with location 0 and scale 0.5. All results are reported as means, with standard errors (SEs) in parentheses, over 100 independent replications.

Competing methods and evaluation metrics. Here, we consider the following competing methods: 1) *Communication-Efficient Distributed Learning* (**Dist-CSL**, Jordan et al. (2019)): We apply CSL to the joint parameter $(\boldsymbol{B}, \boldsymbol{C})$, using nuclear-norm and $\ell_{2,1}$-norm penalties for the two blocks,

respectively. 2) *Distributed Robust Reduced Rank Regression with Quantile Loss* (**Dist-QRR**, Chen et al. (2022b)). 3) *Distributed Estimation in Heterogeneous Reduced Rank Regression* (**Dist-RRR**, Chen et al. (2022a)). 4) **DC-D4R**: A one-shot divide-and-conquer variant of D4R. Each machine alternates the $\boldsymbol{C}$-update in equation 9 and the closed-form reduced-rank update for $\boldsymbol{B}$ until convergence, then sends its local coefficient estimate to the central machine for averaging. 5) **Full-sample**: An oracle benchmark that pools all samples and knows the support of $\boldsymbol{C}^*$ and the rank of $\boldsymbol{B}^*$. It uses the quadratic loss under exponentially tailed noise and the Huber loss under heavy-tailed noise. Let $\hat{\boldsymbol{B}}$ denote the estimate from a given method and set $\boldsymbol{\Delta} = \hat{\boldsymbol{B}} - \boldsymbol{B}^*$. We report (i) the Frobenius error $\|\boldsymbol{\Delta}\|_F$, (ii) the maximum entrywise error $\|\boldsymbol{\Delta}\|_\infty = \max_{ij} |\hat{B}_{ij} - B^*_{ij}|$, (iii) the $\boldsymbol{\Sigma}$-norm error $\|\boldsymbol{\Delta}\|_\Sigma = \sqrt{\mathrm{tr}(\boldsymbol{\Delta}^\top \boldsymbol{\Sigma} \boldsymbol{\Delta})}$, and (iv) the estimated rank $\hat{r}$.

Tuning parameter selection. For each candidate rank $\bar{r} \in \{1, 2, \ldots, 8\}$, we fit D4R under $\mathrm{rank}(\boldsymbol{B}) \le \bar{r}$. At iteration t, machine k selects $\lambda \in \{0.002, 0.004, \ldots, 1.198, 2\}$ by minimizing a groupwise high-dimensional BIC (hBIC) adapted from Wang et al. (2013):

$$\mathrm{hBIC}(\boldsymbol{C}^{(t)}_{\mathcal{I}_k \cdot}) = \log\big(\|\boldsymbol{Y}_{\mathcal{I}_k \cdot} - \boldsymbol{X}_{\mathcal{I}_k \cdot} \boldsymbol{B}^{(t)} - \sqrt{N}\, \boldsymbol{C}^{(t)}_{\mathcal{I}_k \cdot}\|_F^2\big) + \frac{\gamma_{\boldsymbol{C}} q \log n}{n} \|\boldsymbol{C}^{(t)}_{\mathcal{I}_k \cdot}\|_{2,0}. \tag{21}$$

After fitting D4R for all candidate ranks, we select the final rank using the rank selection criterion (RSC) of Bunea et al. (2011):

$$\mathrm{RSC}(\bar{r}) = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}(\bar{r}) - \sqrt{N}\, \hat{\boldsymbol{C}}(\bar{r})\|_F^2 + \mu\, \bar{r}, \tag{22}$$

where $\mu = \gamma_{\boldsymbol{B}}(N + q)$. We set $\gamma_{\boldsymbol{C}} = 2$ and $\gamma_{\boldsymbol{B}} = 1$ and choose the rank with the smallest $\mathrm{RSC}(\bar{r})$. Fits for different candidate ranks are independent and can be run in parallel.

Effect of noise types. Table 1 reports the four metrics under normal and Cauchy noise; results for Laplace and t_3 noise are in Appendix E.2. Under Normal noise, D4R matches the Full-sample oracle across all metrics and correctly recovers the true rank. In contrast, Dist-CSL breaks down, Dist-QRR exhibits persistent bias due to nuclear-norm soft-thresholding, and Dist-RRR substantially overestimates the rank. Under Cauchy noise, D4R remains the best non-oracle method, followed by Dist-QRR, whereas Dist-CSL and Dist-RRR perform substantially worse. DC-D4R is competitive under Normal noise but less stable under Cauchy noise, highlighting the limitation of one-shot averaging. Additional results in Appendix E.2 show that D4R nearly perfectly detects contaminated rows under Normal, Laplace, and t_3 noise and retain near-perfect recall under Cauchy noise, with lower precision due to extreme Cauchy observations being detected as outliers.

Table 1: Estimation accuracy of all competing methods under normal and Cauchy noise. Entries are mean (SE) over 100 replications; bold marks the best non-oracle result.

	Normal				Cauchy			
Method	$\|\boldsymbol{\Delta}\|_F$	$\|\boldsymbol{\Delta}\|_\infty$	$\|\boldsymbol{\Delta}\|_\Sigma$	$\hat{r}$	$\|\boldsymbol{\Delta}\|_F$	$\|\boldsymbol{\Delta}\|_\infty$	$\|\boldsymbol{\Delta}\|_\Sigma$	$\hat{r}$
Full-sample	0.238(0.002)	0.037(0.001)	0.207(0.001)	3.0(0.0)	0.768(0.006)	0.120(0.002)	0.670(0.004)	3.0(0.0)
D4R	**0.240**(0.002)	**0.038**(0.001)	**0.209**(0.001)	**3.0**(0.0)	**1.552**(0.014)	**0.251**(0.004)	**1.351**(0.008)	**3.0**(0.0)
DC-D4R	0.394(0.011)	0.060(0.002)	0.320(0.008)	4.8(0.2)	2.482(0.089)	0.371(0.014)	2.037(0.067)	4.6(0.2)
Dist-CSL	3.283(0.244)	0.533(0.047)	2.743(0.197)	3.9(0.1)	3.953(0.203)	0.645(0.035)	3.446(0.162)	3.0(0.0)
Dist-QRR	2.342(0.131)	0.485(0.032)	1.863(0.103)	3.2(0.1)	2.779(0.128)	0.564(0.033)	2.210(0.100)	3.1(0.0)
Dist-RRR	5.411(0.143)	0.751(0.030)	4.244(0.115)	15.0(0.0)	8.002(0.199)	1.162(0.054)	6.348(0.160)	15.0(0.0)

Effect of number of machines. Figure 1 varies the number of machines m over $\{10, 15, 20, 35, 50, 75, 100\}$ with the local sample size $n = 200$ fixed. As $N = nm$ increases, D4R and DC-D4R both improve, with D4R approaching the Full-sample oracle more rapidly. In contrast, Dist-CSL and Dist-RRR show limited improvement as the number of machines increases. This advantage reflects the iterative refinement of the global estimate through multiple communication rounds, compared with one-shot averaging in DC-D4R.

For the above two experiments, paired Wilcoxon signed-rank tests reported in Appendix E.5 confirm that the improvements of D4R over the competing methods are statistically significant.

Table 2: Test MSE of all methods on the Airline SIR and Beijing PM2.5 datasets across different contamination proportions. Entries are mean (SE) over 100 train/test splits; bold marks the best reported value in each row.

Dataset	O/N	D4R	DC-D4R	Dist-CSL	Dist-QRR	Dist-RRR
Airline SIR	0.05	0.851(0.001)	**0.850**(0.001)	0.859(0.004)	0.991(0.007)	1.293(0.022)
($N_{\text{total}} = 20{,}000$	0.10	0.851(0.001)	**0.850**(0.001)	0.870(0.008)	0.978(0.005)	1.721(0.041)
$p = 10,\ q = 5$	0.15	0.851(0.001)	**0.850**(0.001)	2.182(0.051)	0.988(0.009)	2.106(0.052)
$m = 10$)	0.20	**0.851**(0.001)	0.851(0.001)	2.634(0.065)	0.982(0.008)	2.501(0.071)
Beijing PM2.5	0.05	**0.595**(0.003)	0.597(0.003)	1.115(0.024)	0.695(0.004)	1.467(0.033)
($N_{\text{total}} = 7137$	0.10	**0.595**(0.003)	0.599(0.003)	2.250(0.056)	0.697(0.004)	2.160(0.050)
$p = 17,\ q = 12$	0.15	**0.595**(0.003)	0.599(0.003)	3.268(0.114)	0.697(0.005)	3.082(0.122)
$m = 10$)	0.20	**0.596**(0.003)	0.599(0.003)	4.073(0.143)	0.697(0.005)	3.536(0.121)

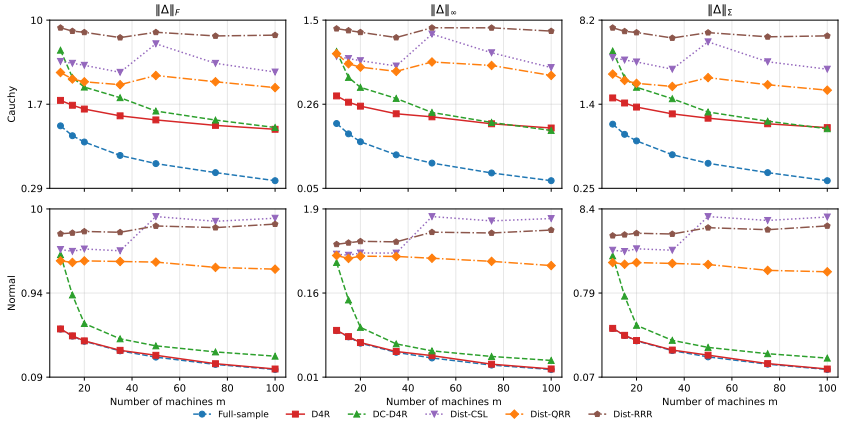

Figure 1: Effect of the number of machines m under Normal noise, with the local sample size fixed at $n = 200$.

Effect of communication rounds. Figure 2 reports the Frobenius estimation error of D4R over rounds $u = 1, \ldots, 8$ for three configurations with the same total sample size $N = 4000$: $(n, m) = (500, 8)$, $(800, 5)$, and $(1000, 4)$. We use $O/N = 0.10$ and Normal noise. The error decreases approximately geometrically and all three curves stabilize within 5 rounds, consistent with Corollary 5. Larger local sample sizes lead to faster convergence, as predicted by the dependence of the contraction factor on n in Proposition 3, with $(1000, 4)$ exhibiting the steepest initial error reduction.

Additional studies. Due to space constraints, simulations examining the effects of the contamination proportion O/N and signal strength a, together with a computational time comparison, are provided in Appendices E.3, E.4 and E.6.

5.2 Real Data Studies

We further evaluate D4R on four real datasets in distributed settings with row-wise response contamination.

Datasets and preprocessing. We consider four datasets. (i) *Airline SIR* ($N_{\text{total}} = 20{,}000$, $p = 10$, $q = 5$): 2008 U.S. domestic flight records, with arrival delay converted into five equal-frequency slice indicators following the SIR construction of Chen et al. (2022a). (ii) *Beijing PM2.5* ($N_{\text{total}} = 7137$, $p = 17$, $q = 12$): hourly measurements from twelve monitoring stations in 2016 (Zhang et al., 2017). The station-specific PM2.5 concentrations are responses, and the covariates include station-averaged meteorological variables, wind direction, and cyclical time features. (iii) *Facebook Live* ($N_{\text{total}} = 7050$, $p = 7$, $q = 8$): data from Thai Facebook Live sellers (Dehouche, 2020), with

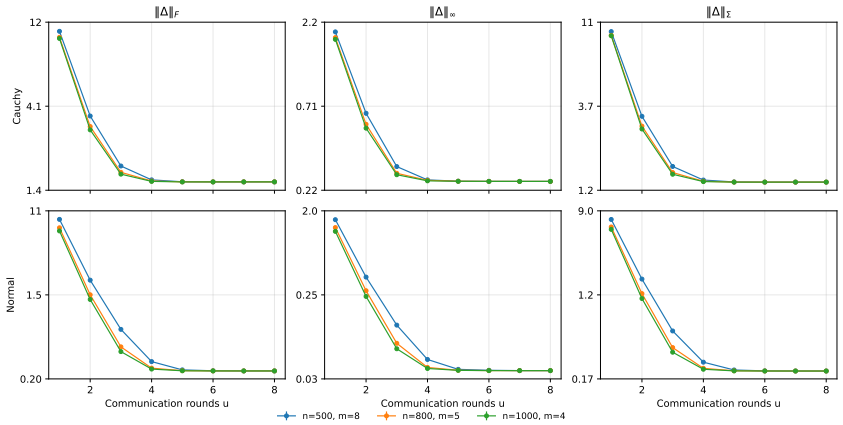

Figure 2: D4R convergence across communication rounds for three (n, m) configurations with fixed $N = 4000$. The estimation error decreases approximately geometrically and approaches its limiting level within five rounds, with larger local sample sizes yielding faster convergence.

post characteristics as covariates and `log1p`-transformed engagement measures as responses. (iv) *Facebook Metrics* ($N_{\text{total}} = 495$, $p = 9$, $q = 12$): data from a cosmetics brand's Facebook page (Moro et al., 2016), with post characteristics as covariates and engagement metrics as responses.

The Beijing PM2.5, Facebook Live, and Facebook Metrics datasets are available from the UCI Machine Learning Repository, whereas the Airline data are from the 2009 ASA Data Expo (Wickham, 2011). These datasets cover diverse sample sizes and multivariate response structures, providing complementary real-data settings for evaluating distributed estimation. For each dataset, we horizontally partition the training samples across machines to construct a distributed environment. Then we standardize $\boldsymbol{X}$ and $\boldsymbol{Y}$ and use a random 70%/30% train/test split.

Experimental setup. We divide the training data evenly across m machines. To compare the robustness of different methods, we randomly contaminate $O/N \in \{0.05, 0.10, 0.15, 0.20\}$ of the training observations. Following model 2, we add row-sparse contamination by setting $\boldsymbol{Y}_{i\cdot} \leftarrow \boldsymbol{Y}_{i\cdot} + \sqrt{N}\mathbf{1}_q$ for each selected observation. For each dataset and contamination level, we repeat the experiment over 100 random train/test splits and report mean test MSE and corresponding SEs.

Results. Table 2 reports results for Airline SIR and Beijing PM2.5. Results for the remaining datasets are in Appendix E.7 and show a similar pattern. D4R achieves the best or near-best performance across all contamination levels, with little change in test MSE as contamination increases. DC-D4R nearly matches D4R on Airline SIR but is slightly worse on Beijing PM2.5. This is consistent with the relatively large local sample sizes in our real-data settings, which yields accurate local estimates before averaging. In contrast, Dist-CSL and Dist-RRR deteriorate markedly as contamination ratio increases. Dist-QRR is less sensitive to contamination but remains less accurate than D4R.

6 Conclusion

In this paper, we proposed D4R, a distributed robust reduced-rank regression framework for large-scale multivariate data with anomalous observations. A row-sparse contamination model and a refined surrogate allow D4R to jointly estimate the low-rank coefficient matrix while detecting contaminated observations in a distributed setting. Under standard sub-Gaussian conditions, D4R attains the target nonasymptotic error rate after a constant number of communication rounds. Extensive numerical studies demonstrated the robustness and efficiency of D4R relative to existing alternatives. A limitation of the current analysis is the assumption of balanced data partitions across machines. Extending D4R to broader federated learning settings with heterogeneous client sample sizes and data distributions is a promising direction for future research.

AI USE STATEMENT

Generative AI tools were used only for language polishing and improving the clarity of author written text. All edits were reviewed by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

This study focuses on methodological development and evaluation using synthetic and publicly available datasets. We do not identify any material ethical concerns associated with the study.

REPRODUCIBILITY STATEMENT

The complete D4R algorithm and proofs of the main theoretical results are provided in Appendices A and B. Section 5 details the experimental settings, tuning procedures, datasets, and preprocessing. Additional experiments and the computing environment are also reported in Appendices E and F.

REFERENCES

Theodore Wilbur Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951.

Aditya Bhaskara and Srivatsan Kumar. Low rank approximation in the presence of outliers. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*, volume 116 of *Leibniz International Proceedings in Informatics*, pp. 4:1–4:16, 2018.

Florentina Bunea, Yiyuan She, and Marten H Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

Canyi Chen, Wangli Xu, and Liping Zhu. Distributed estimation in heterogeneous reduced rank regression: With application to order determination in sufficient dimension reduction. *Journal of Multivariate Analysis*, 190:104991, 2022a.

Xi Chen, Weidong Liu, and Xiaojun Mao. Robust reduced rank regression in a distributed setting. *Science China Mathematics*, 65(8):1707–1730, 2022b.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

Youngjin Cho and Seyoung Park. Multivariate response regression with low-rank and generalized sparsity. *Journal of the Korean Statistical Society*, 51(3):847–867, 2022.

Nassim Dehouche. Dataset on usage and engagement patterns for facebook live sellers in thailand. *Data in Brief*, 30:105661, 2020.

David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969, 2016.

Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 212(1):177–202, 2019.

Xiaojie Guo and Zhouchen Lin. Low-rank matrix recovery via robust outlier estimation. *IEEE Transactions on Image Processing*, 27(11):5316–5327, 2018.

Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pp. 1248–1251. Springer, 2011.

Mia Hubert, Peter J Rousseeuw, and Wannes Van den Bossche. Macropca: An all-in-one pca method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics*, 61(4):459–473, 2019.

Zihao Jiang, Weidong Zhou, Yu Guo, Chao Zhang, and Lu Tian. A novel robust multivariate laplace distribution-based distributed consensus information fusion. *Measurement*, 225:113986, 2024.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526), 2019.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2):1–210, 2021.

Longtan Le, Tung-Anh Nguyen, Han Shu, Suranga Seneviratne, Choong Seon Hong, and Nguyen H Tran. Federated koopman-reservoir learning for large-scale multivariate time-series anomaly detection. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*, pp. 61–70. SIAM, 2025.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Li Liu, Douglas M Hawkins, Sujoy Ghosh, and S Stanley Young. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100(23): 13167–13172, 2003.

Shixiang Liu, Delin Zhao, and Hanming Yang. Distributed learning with adversarial contamination: an efficient and optimal approach based on dynamic regularization. *Statistics and Computing*, 36 (4):150, 2026.

Roberto López-Valcarce, Daniel Romero, Josep Sala, and Alba Pages-Zamora. Distributed multivariate regression with unknown noise covariance in the presence of outliers: An mdl approach. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 1–5. IEEE, 2016.

Canyi Lu, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):925–938, 2019.

Jiyu Luo, Qiang Sun, and Wen-Xin Zhou. Distributed adaptive huber regression. *Computational Statistics & Data Analysis*, 169:107419, 2022.

Wei Ma, Junzhuo Gao, Lei Wang, and Heng Lian. Sparse and debiased adaptive huber regression in distributed data: Aggregated and communication-efficient approaches. *Statistica Sinica*, 35(4): 1971–1990, 2025.

Zhuang Ma, Zongming Ma, and Tingni Sun. Adaptive estimation in two-way sparse reduced-rank regression. *Statistica Sinica*, 30(4):2179–2201, 2020.

Sérgio Moro, Paulo Rita, and Bernardo Vala. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, 69(9):3341–3351, 2016.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Luong Trung Nguyen, Junhan Kim, Sangtae Kim, and Byonghyo Shim. Localization of iot networks via low-rank matrix completion. *IEEE Transactions on Communications*, 67(8):5833–5847, 2019.

Feiping Nie, Heng Huang, and Chris Ding. Low-rank matrix recovery via efficient schatten p-norm minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 655–661, 2012.

Gregory C Reinsel and Raja P Velu. *Multivariate reduced-rank regression*. Springer, 1998.

Takeyuki Sasai and Hironori Fujisawa. Adversarial robust low rank matrix estimation: Compressed sensing and matrix completion, 2020.

Marc Senneret, Yannick Malevergne, Patrice Abry, Gerald Perrin, and Laurent Jaffres. Covariance versus precision matrix estimation for efficient asset allocation. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):982–993, 2016.

Yiyuan She and Kun Chen. Robust reduced-rank regression. *Biometrika*, 104(3):633–647, 2017.

Kean Ming Tan, Qiang Sun, and Daniela Witten. Sparse reduced rank huber regression in high dimensions. *Journal of the American Statistical Association*, 118(544):2383–2393, 2023.

Philip Thompson. Outlier-robust sparse/low-rank least-squares regression and robust matrix completion, 2020.

N Trendafilov, M Gallo, V Simonacci, and V Todorov. Overview of reduced-rank regression with dense and sparse coefficients, and a new estimation procedure. *Behaviormetrika*, 52(2):593–625, 2025.

Jiyuan Tu, Weidong Liu, and Xiaojun Mao. Byzantine-robust distributed sparse learning for m-estimation. *Machine Learning*, 112(10):3773–3804, 2023.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Cheng Wang, Zhou Yu, and Liping Zhu. On cumulative slicing estimation for high dimensional data. *Statistica Sinica*, 31(1):223–242, 2021.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *International conference on machine learning*, pp. 3636–3645. PMLR, 2017.

Lan Wang, Yongdai Kim, and Runze Li. Calibrating non-convex penalized regression in ultra-high dimension. *Annals of statistics*, 41(5):2505, 2013.

Yue Wang, Wenqi Lu, Lei Wang, Zhongyi Zhu, Hongmei Lin, and Heng Lian. Regularized adaptive huber matrix regression and distributed learning. *Statistica Sinica*, 35(2):919–937, 2025.

Hadley Wickham. ASA 2009 data expo. *Journal of Computational and Graphical Statistics*, 20(2): 281–283, 2011.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pp. 5650–5659. PMLR, 2018.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017.

Wei Zhi, Jane Minturn, Eric Rappaport, Garrett Brodeur, and Hongzhe Li. Network-based analysis of multivariate gene expression data. In *Statistical Methods for Microarray Data Analysis: Methods and Protocols*, pp. 121–139. Springer, 2013.

Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2000–2031, 2022.

A Distributed Optimization Algorithm for D4R

Algorithm 1 Distributed Optimization Procedure for D4R

Require: Local data $\{(\boldsymbol{X}_{\mathcal{I}_k\cdot}, \boldsymbol{Y}_{\mathcal{I}_k\cdot})\}_{k=1}^m$, target rank r, total rounds T, $\boldsymbol{B}^{(0)} = \boldsymbol{0}_{p\times q}$, and $\{\lambda_k^{(t)}\}_{k\in[m],t=0,\ldots,T-1}$.

1: **for** $t = 1, 2, \ldots, T$ **do**

2: The central machine broadcasts the current $\boldsymbol{B}^{(t-1)}$ to all machines.

3: Each machine k updates its contamination matrix via

$$\boldsymbol{C}_{\mathcal{I}_k\cdot}^{(t-1)} = \mathcal{H}_{\lambda_k^{(t-1)}}\left(\frac{\boldsymbol{Y}_{\mathcal{I}_k\cdot} - \boldsymbol{X}_{\mathcal{I}_k\cdot}\boldsymbol{B}^{(t-1)}}{\sqrt{N}}\right).$$

4: Each machine k computes the adjusted local gradient and sends this $p\times q$ matrix to the central machine

$$\boldsymbol{G}_k^{(t-1)} = \boldsymbol{X}_{\mathcal{I}_k\cdot}^\top(\boldsymbol{Y}_{\mathcal{I}_k\cdot} - \boldsymbol{X}_{\mathcal{I}_k\cdot}\boldsymbol{B}^{(t-1)} - \sqrt{N}\boldsymbol{C}_{\mathcal{I}_k\cdot}^{(t-1)}).$$

5: The central machine collects adjusted gradients from all m machines and constructs the pseudo-response matrix

$$\boldsymbol{M}_{\boldsymbol{B}}^{(t-1)} = \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t-1)} + \boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}\frac{1}{N}\sum_{k=1}^m \boldsymbol{G}_k^{(t-1)}.$$

6: The central machine updates $\boldsymbol{B}^{(t)}$ via

$$\boldsymbol{B}^{(t)} = \frac{1}{n}\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top\boldsymbol{M}_{\boldsymbol{B}}^{(t-1)}P_{V(\boldsymbol{X}_{\mathcal{I}_1\cdot},\boldsymbol{M}_{\boldsymbol{B}}^{(t-1)},r)}.$$

7: **end for**

8: **Output:** $\hat{\boldsymbol{B}} = \boldsymbol{B}^{(T)}$, $\hat{\boldsymbol{C}} = (\boldsymbol{C}_{\mathcal{I}_1\cdot}^{(T-1)\top}, \ldots, \boldsymbol{C}_{\mathcal{I}_m\cdot}^{(T-1)\top})^\top$.

B Proofs of Main Results

The proof of our main results follows the iterative process detailed in Algorithm 1. Throughout this appendix, we denote the symmetric difference of sets A_1 and A_2 as $A_1 \triangle A_2 = (A_1\setminus A_2)\cup(A_2\setminus A_1)$. Let $\mathcal{O}_k^* = \{i \in \mathcal{I}_k : \boldsymbol{C}_{i\cdot}^* \neq \boldsymbol{0}\}$ denote the index set of contaminated observations on machine k and define the global contamination set as $\mathcal{O}^* = \bigcup_{k=1}^m \mathcal{O}_k^*$. Furthermore, to track the convergence at each iteration t, we define the estimation errors in the Frobenius norm as $e_{\boldsymbol{B}}^{(t)} := \|\boldsymbol{B}^* - \boldsymbol{B}^{(t)}\|_F$ and $e_{\boldsymbol{C}}^{(t)} := \|\boldsymbol{C}^* - \boldsymbol{C}^{(t)}\|_F$.To facilitate our theoretical analysis, particularly for the hard thresholding step, we introduce a complementary truncation operator $\mathcal{T}_\lambda : \mathbb{R}^{v\times q} \to \mathbb{R}^{v\times q}$:

$$(\mathcal{T}_\lambda(\boldsymbol{A}))_{i\cdot} := \boldsymbol{A}_{i\cdot} \times \boldsymbol{1}\left(\|\boldsymbol{A}_{i\cdot}\|_2 < \lambda\right), \quad \text{for } i \in [v].$$

This operator captures the components filtered out by the threshold. By definition, it trivially follows that $\mathcal{H}_\lambda(\boldsymbol{A}) = \boldsymbol{A} - \mathcal{T}_\lambda(\boldsymbol{A})$.

B.1 Step 1: Local contamination learning in each machine

In the first round, each machine k calculates the outlier estimator as

$$\begin{aligned}
\boldsymbol{C}_{\mathcal{I}_k\cdot}^{(0)} &= H_{\lambda_k^{(0)}}(\frac{\boldsymbol{Y}_{\mathcal{I}_k\cdot} - \boldsymbol{X}_{\mathcal{I}_k\cdot}\boldsymbol{B}^{(0)}}{\sqrt{N}}) \\
&= H_{\lambda_k^{(0)}}(\boldsymbol{C}_{\mathcal{I}_k\cdot}^* + \frac{1}{\sqrt{N}}\boldsymbol{X}_{\mathcal{I}_k}(\boldsymbol{B}^* - \boldsymbol{B}^{(0)}) + \frac{1}{\sqrt{N}}\boldsymbol{\mathcal{E}}_{\mathcal{I}_k\cdot}) \\
&= \boldsymbol{C}_{\mathcal{I}_k\cdot}^* + \frac{1}{\sqrt{N}}\boldsymbol{X}_{\mathcal{I}_k}(\boldsymbol{B}^* - \boldsymbol{B}^{(0)}) + \frac{1}{\sqrt{N}}\boldsymbol{\mathcal{E}}_{\mathcal{I}_k\cdot} \\
&\quad - \mathcal{T}_{\lambda_k^{(0)}}(\boldsymbol{C}_{\mathcal{I}_k\cdot}^* + \frac{1}{\sqrt{N}}\boldsymbol{X}_{\mathcal{I}_k}(\boldsymbol{B}^* - \boldsymbol{B}^{(0)}) + \frac{1}{\sqrt{N}}\boldsymbol{\mathcal{E}}_{\mathcal{I}_k\cdot})
\end{aligned} \tag{B.1}$$

We first establish the row-sparsity and the estimation error rate of $C^{(0)}_{\mathcal{I}_k\cdot}$. Let $\mathcal{O}^{(t)}_k$ denote the index set of detected outliers at the k-th machine in the t-th round, meaning that $\mathcal{O}^{(0)}_k = \{i \in \mathcal{I}_k : \boldsymbol{C}^{(0)}_{i\cdot} \neq \mathbf{0}\}$.

Now, we prove $|\mathcal{O}^{(0)}_k \setminus \mathcal{O}^*_k| < C_s o_k$ by contradiction, where $C_s \geq 1$ is a constant. Specifically, if $|\mathcal{O}^{(0)}_k \setminus \mathcal{O}^*_k| \geq C_s o_k$, we can always find a subset $\mathcal{O}'_k \subseteq (\mathcal{O}^{(0)}_k \setminus \mathcal{O}^*_k) \subset \mathcal{I}_k$ satisfying $|\mathcal{O}'_k| = C_s o_k$. Then, with Lemmas C.3 and C.8, there exist universal positive constants C', c such that

$$\begin{aligned}\sqrt{C_s o_k}\lambda^{(0)}_k &\leq \sqrt{\sum_{i\in\mathcal{O}'_k} \|(\boldsymbol{C}^{(0)})_{i\cdot}\|_2^2} = \|\boldsymbol{C}^{(0)}_{\mathcal{O}'_k\cdot}\|_F \\ &< \frac{1}{\sqrt{N}}\|\boldsymbol{X}_{\mathcal{O}'_k\cdot}\|_2\|\boldsymbol{B}^* - \boldsymbol{B}^{(0)}\|_F + \frac{1}{\sqrt{N}}\|\boldsymbol{\mathcal{E}}_{\mathcal{O}'_k\cdot}\|_F \\ &\leq C_{X,2}\sqrt{C_s}\sqrt{\frac{p+o_k\log N}{N}}\|\boldsymbol{B}^* - \boldsymbol{B}^{(0)}\|_F + C'\sigma\sqrt{C_s}\frac{\sqrt{qo_k}+\sqrt{o_k\log N}}{\sqrt{N}}\end{aligned} \tag{B.2}$$

holds with probability at least $1 - \mathcal{O}(e^{-c\log N})$, where the constant $C_{X,2}$ follows from equation C.3 in Lemma C.3.

Therefore, by taking

$$\lambda^{(0)}_k = C_{X,2}\sqrt{\frac{p+o_k\log N}{o_k N}}e^{(0)}_{\boldsymbol{B}} + C'\sigma\frac{\sqrt{q}+\sqrt{\log N}}{\sqrt{N}}, \tag{B.3}$$

for every $k \in [m]$, we show the contradiction and thus $|\mathcal{O}^{(0)}_k \setminus \mathcal{O}^*_k| < C_s o_k$. Then, for all $k \in [m]$, with probability at least $1 - \mathcal{O}(e^{-c\log N})$, we have

$$\begin{aligned}&\|\boldsymbol{C}^{(0)}_{\mathcal{I}_k\cdot} - \boldsymbol{C}^*_{\mathcal{I}_k\cdot}\|_F \\ \leq&\sqrt{\sum_{i\in\mathcal{O}^{(0)}_k\cup\mathcal{O}^*_k}\left\|\left(\frac{1}{\sqrt{N}}\boldsymbol{X}_{\mathcal{I}_k}(\boldsymbol{B}^* - \boldsymbol{B}^{(0)}) + \frac{1}{\sqrt{N}}\boldsymbol{\mathcal{E}}_{\mathcal{I}_k\cdot}\right)_{i\cdot}\right\|_2^2} \\ &+\sqrt{\sum_{i\in\mathcal{O}^*_k}\left\|\left(\frac{\boldsymbol{Y}_{\mathcal{I}_k\cdot} - \boldsymbol{X}_{\mathcal{I}_k\cdot}\boldsymbol{B}^{(0)}}{\sqrt{N}}\right)_{i\cdot}\right\|_2^2 \mathbf{1}\left(\left\|\left(\frac{\boldsymbol{Y}_{\mathcal{I}_k\cdot} - \boldsymbol{X}_{\mathcal{I}_k\cdot}\boldsymbol{B}^{(0)}}{\sqrt{N}}\right)_{i\cdot}\right\|_2 < \lambda^{(0)}_k\right)} \\ \leq&2\sqrt{C_s}C_{X,2}\sqrt{\frac{p+o_k\log N}{N}}\|\boldsymbol{B}^* - \boldsymbol{B}^{(0)}\|_F + 2C'\sigma\sqrt{C_s}\frac{\sqrt{qo_k}+\sqrt{o_k\log N}}{\sqrt{N}} + \sqrt{o_k}\lambda^{(0)}_k \\ \leq&\underbrace{C_{X,2}(2\sqrt{C_s}+1)}_{=:C_1}\sqrt{\frac{p+o_k\log N}{N}}e^{(0)}_{\boldsymbol{B}} + \underbrace{C'(2\sqrt{C_s}+1)}_{=:C_2}\sigma\frac{\sqrt{qo_k}+\sqrt{o_k\log N}}{\sqrt{N}},\end{aligned} \tag{B.4}$$

Summing over machines gives

$$\begin{aligned}e^{(0)}_{\boldsymbol{C}} = \|\boldsymbol{C}^{(0)} - \boldsymbol{C}^*\|_F &= \sqrt{\sum_{k\in[m]}\|\boldsymbol{C}^{(0)}_{\mathcal{I}_k\cdot} - \boldsymbol{C}^*_{\mathcal{I}_k\cdot}\|_F^2} \\ &\leq \sqrt{2C_1^2(e^{(0)}_{\boldsymbol{B}})^2\sum_{k\in[m]}\frac{p+o_k\log N}{N} + 4C_2^2\sigma^2\sum_{k\in[m]}\frac{qo_k+o_k\log N}{N}} \\ &\leq \sqrt{2}C_1\sqrt{\frac{p}{n}+\frac{O\log N}{N}}e^{(0)}_B + 2C_2\sigma\sqrt{\frac{qO+O\log N}{N}},\end{aligned} \tag{B.5}$$

holds with probability at least $1 - \mathcal{O}(e^{-c\log N})$, and $|\mathcal{O}^{(0)}\triangle\mathcal{O}^*| < (C_s+1)O$.

The same argument applies at every round $t \geq 1$. By taking

$$\lambda^{(t)}_k = C_{X,2}\sqrt{\frac{p+o_k\log N}{o_k N}}e^{(t)}_{\boldsymbol{B}} + C'\sigma\frac{\sqrt{q}+\sqrt{\log N}}{\sqrt{N}}, \tag{B.6}$$

the estimation error bound for the outlier matrix naturally extends to any t-th communication round ($t > 0$). Specifically, we have

$$e_{\boldsymbol{C}}^{(t)} \leq \sqrt{2}C_1\sqrt{\frac{p}{n} + \frac{O\log N}{N}}e_{\boldsymbol{B}}^{(t)} + 2C_2\sigma\sqrt{\frac{qO + O\log N}{N}} \tag{B.7}$$

which holds with probability at least $1 - \mathcal{O}(e^{-c\log N})$, alongside $|\mathcal{O}^{(t)} \triangle \mathcal{O}^*| < (C_s + 1)O$. The proof is omitted as it closely follows the initial round. We emphasize that this result holds not only in the case $o_k > 0$, but also in the case $o_k = 0$. In such an uncontaminated scenario, simply setting $\lambda_k^{(t)} := C_{X,2}\sqrt{\frac{p+\log N}{N}}e_{\boldsymbol{B}}^{(t)} + C'''\sigma\frac{\sqrt{q}+\sqrt{\log N}}{\sqrt{N}}$ theoretically guarantees that the detected outlier set is empty, i.e., $\mathcal{O}_k^{(t)} = \emptyset$.

B.2 Step 2: Low-rank learning in the central machine

Let $\boldsymbol{M}_e^{(t-1)} = \boldsymbol{M}_{\boldsymbol{B}}^{(t-1)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*$. By the optimality condition $Q(\boldsymbol{B}^{(t)}; \boldsymbol{B}^{(t-1)}) \leq Q(\boldsymbol{B}^*; \boldsymbol{B}^{(t-1)})$, we have

$$\frac{1}{2}\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\|_F^2 \leq \langle \boldsymbol{M}_{\boldsymbol{B}}^{(t-1)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*, \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\rangle. \tag{B.8}$$

Notice that the first term in the inner product can be decomposed as

$$\begin{aligned}
&\boldsymbol{M}_{\boldsymbol{B}}^{(t-1)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^* \\
=&\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t-1)} + \boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}[\frac{1}{N}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^{(t-1)} - \sqrt{N}\boldsymbol{C}^{(t-1)})] - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^* \\
=&\boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}\{\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{\mathcal{E}} + (\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})(\boldsymbol{B}^{(t-1)} - \boldsymbol{B}^*)\} + \boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}\frac{1}{\sqrt{N}}\boldsymbol{X}^\top(\boldsymbol{C}^* - \boldsymbol{C}^{(t-1)}).
\end{aligned} \tag{B.9}$$

Substituting this decomposition into equation B.8, and applying the von Neumann's trace duality inequality to the first term along with the Cauchy-Schwarz inequality to the second term, we obtain

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\|_F^2 \\
\leq&\|\frac{1}{N}\boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{X}^\top\boldsymbol{\mathcal{E}}\|_2\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\|_* \\
&+ \|\boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})(\boldsymbol{B}^{(t-1)} - \boldsymbol{B}^*)\|_F\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\|_F \\
&+ \|\boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}\frac{1}{\sqrt{N}}\boldsymbol{X}^\top(\boldsymbol{C}^* - \boldsymbol{C}^{(t-1)})\|_F\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)} - \boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\|_F.
\end{aligned} \tag{B.10}$$

It remains to deal with $I_1 := \|\frac{1}{N}\boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}\boldsymbol{X}^\top\boldsymbol{\mathcal{E}}\|_2$, $I_2 := \|\boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})(\boldsymbol{B}^{(t-1)} - \boldsymbol{B}^*)\|_F$ and $I_3 := \|\boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}\frac{1}{\sqrt{N}}\boldsymbol{X}^\top(\boldsymbol{C}^* - \boldsymbol{C}^{(t-1)})\|_F$.

By the triangle inequality and the sub-multiplicativity of the spectral norm, we can bound I_1 as follows. Applying Lemma C.9 for the spectral norm of $\boldsymbol{X}_{\mathcal{I}_1\cdot}$, Lemma C.6 for $\hat{\boldsymbol{\Sigma}}_1^{-1}$, Lemma C.1 for the gradient noise term, and Lemma C.7 for the covariance heterogeneity term, we obtain

$$\begin{aligned}
I_1 \leq&\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}(\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{\mathcal{E}})\|_2 \\
\leq&\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\|_2\|\hat{\boldsymbol{\Sigma}}_1^{-1}\|_2\|\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{\mathcal{E}}\|_2 \\
\leq&C_{X,3}(\sqrt{n} + \sqrt{p} + \sqrt{\log n})\frac{2}{C_L}C_3\sigma\sqrt{\frac{p + q + \log(2/\epsilon)}{N}},
\end{aligned} \tag{B.11}$$

which holds with probability at least $1 - \epsilon - \mathcal{O}(e^{-c\log n})$.

Similarly, we have

$$\begin{aligned}
I_2 \leq&\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\hat{\boldsymbol{\Sigma}}_1^{-1}(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})(\boldsymbol{B}^{(t-1)} - \boldsymbol{B}^*)\|_F \\
\leq&\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\|_2\|\hat{\boldsymbol{\Sigma}}_1^{-1}\|_2\|\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}}\|_2\|\boldsymbol{B}^{(t-1)} - \boldsymbol{B}^*\|_F \\
\leq&C_{X,3}(\sqrt{n} + \sqrt{p} + \sqrt{\log n})\frac{2}{C_L}C_{X,1}\left(\sqrt{\frac{p + \log n}{n}} + \sqrt{\frac{p + \log n}{N}}\right)e_{\boldsymbol{B}}^{(t-1)},
\end{aligned} \tag{B.12}$$

which holds with probability at least $1-\mathcal{O}(e^{-c\log n})$.

For I_3, let $\mathcal{O}_{\text{full}}$ be the index set of the non-zero rows in $(\boldsymbol{C}^*-\boldsymbol{C}^{(t-1)})$ with cardinality $|\mathcal{O}_{\text{full}}|\leq (C_s+1)O$. We apply Lemma C.4 to bound $\|\boldsymbol{X}_{\mathcal{O}_{\text{full}}\cdot}\|_2$, which yields

$$\begin{aligned} I_3 \leq & \|\boldsymbol{X}_{\mathcal{I}_1\cdot}\|_2\|\hat{\boldsymbol{\Sigma}}_1^{-1}\|_2\frac{1}{\sqrt{N}}\|\boldsymbol{X}_{\mathcal{O}_{\text{full}}\cdot}\|_2\|(\boldsymbol{C}^*-\boldsymbol{C}^{(t-1)})\|_F \\ \leq & C_{X,3}(\sqrt{n}+\sqrt{p}+\sqrt{\log n})\frac{2}{C_L}\sqrt{C_s+1}C_{X,2}\sqrt{\frac{p+O\log N}{N}}e_{\boldsymbol{C}}^{(t-1)} \\ \leq & C_{X,3}(\sqrt{n}+\sqrt{p}+\sqrt{\log n})\frac{2}{C_L}\sqrt{C_s+1}C_{X,2}\sqrt{\frac{p+O\log N}{N}} \\ & \times\{\sqrt{2}C_1\sqrt{\frac{p}{n}+\frac{O\log N}{N}}e_{\boldsymbol{B}}^{(t-1)}+2C_2\sigma\sqrt{\frac{qO+O\log N}{N}}\}, \end{aligned} \tag{B.13}$$

which holds with probability at least $1-\mathcal{O}(e^{-c\log n})$.

By Cauchy–Schwarz inequality, we have

$$\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)}-\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\|_*\leq\sqrt{2r}\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)}-\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\|_F. \tag{B.14}$$

Note that $\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)}-\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\|_F^2\geq n\Lambda_{\min}(\hat{\boldsymbol{\Sigma}}_1)\|\boldsymbol{B}^{(t)}-\boldsymbol{B}^*\|_F^2\geq n\frac{C_L}{2}\|\boldsymbol{B}^{(t)}-\boldsymbol{B}^*\|_F^2$ holds with probability at least $1-\mathcal{O}(e^{-c\log n})$, we have

$$e_{\boldsymbol{B}}^{(t)}=\|\boldsymbol{B}^{(t)}-\boldsymbol{B}^*\|_F\leq\sqrt{\frac{2}{nC_L}}\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^{(t)}-\boldsymbol{X}_{\mathcal{I}_1\cdot}\boldsymbol{B}^*\|_F. \tag{B.15}$$

Combining the preceding bounds gives

$$e_{\boldsymbol{B}}^{(t)}\leq\kappa\cdot e_{\boldsymbol{B}}^{(t-1)}+\Delta, \tag{B.16}$$

where

$$\kappa=3\{\frac{8\sqrt{2}C_{X,3}C_{X,1}+16\sqrt{C_s+1}C_{X,3}C_{X,2}C_1}{C_L^{3/2}}(\sqrt{\frac{p+\log n}{n}}+\frac{O\log N}{N})\} \tag{B.17}$$

and

$$\begin{aligned} \Delta=3\sigma\Bigg\{ & \left(\frac{8C_{X,3}C_3}{C_L^{3/2}}+\frac{4\sqrt{2C_s+2}C_{X,3}C_{X,2}C_2}{C_L^{3/2}}\right)\sqrt{\frac{r(p+q+\log(2/\epsilon))}{N}} \\ & +\frac{12\sqrt{2C_s+2}C_{X,3}C_{X,2}C_2}{C_L^{3/2}}\frac{O(q+\log N)}{N}\Bigg\}. \end{aligned} \tag{B.18}$$

Let $C=4(1\vee A^2)$ with $A:=\frac{24\sqrt{2}C_{X,3}C_{X,1}+48\sqrt{C_s+1}C_{X,3}C_{X,2}C_1}{C_L^{3/2}}$ in Assumption 2. It is easy to verify that $\kappa\leq 1/2$.

Unrolling the recurrence inequality B.16 to the T-th iteration yields $e_{\boldsymbol{B}}^{(T)}\leq\kappa^T e_{\boldsymbol{B}}^{(0)}+2\Delta$. To ensure the optimization error is dominated by the statistical error, we require $\kappa^T e_{\boldsymbol{B}}^{(0)}\asymp\Delta$. Moreover, we have $\Delta\asymp\sigma(\sqrt{\frac{r(p+q)}{N}}+\frac{qO+O\log N}{N})$. Consequently, assuming there exist constants $\alpha_1>0$ and $\alpha_2\in(0,1)$ such that $e_{\boldsymbol{B}}^{(0)}/\sigma\leq N^{\alpha_1}$ and $\max\{mp,O\}\leq N^{\alpha_2}$, then the required number of communication rounds T can be explicitly specified as

$$\begin{aligned} T= & \frac{\log(C''\sigma/e_{\boldsymbol{B}}^{(0)})-\log N+\log(\sqrt{Nr(p+q)}+O\log N+qO)}{\log(A/N)+\log\left(\sqrt{Nm(p+\log n)}+O\log N\right)} \\ \asymp & \frac{\log(e_{\boldsymbol{B}}^{(0)}/\sigma)+\log N-\log(\sqrt{Nr(p+q)}+O\log N+qO)}{\log(N)-\log\left(\sqrt{Nm(p+\log n)}+O\log N\right)} \\ \lesssim & \frac{(1+\alpha_1)\log N}{\log N-\frac{1+\alpha_2}{2}\log N-\log\log N-\log 3}\asymp 1. \end{aligned} \tag{B.19}$$

Since $T \asymp 1$, we conclude that the desired estimation error rate is achieved within a finite number of communication rounds, with probability at least $1 - \epsilon - \mathcal{O}(e^{-c \log n})$.

C TECHNICAL LEMMAS

Lemma C.1. *Under Assumptions 1 and 2, for any $\epsilon \in (0, 1)$, there exists an absolute constant $C_3 > 0$ such that*

$$\Pr\left(\left\|\frac{1}{N}\boldsymbol{X}^\top \boldsymbol{\mathcal{E}}\right\|_2 \geq C_3 \sigma \sqrt{\frac{p + q + \log(2/\epsilon)}{N}}\right) \leq \epsilon. \tag{C.1}$$

Proof. We aim to bound the spectral norm of $\boldsymbol{A} := \frac{1}{N}\boldsymbol{X}^\top \boldsymbol{\mathcal{E}} \in \mathbb{R}^{p \times q}$.

Step 1: Discretization via ϵ-nets. The spectral norm of $\boldsymbol{A}$ is defined as the supremum over the unit spheres:

$$\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{\alpha} \in S^{p-1},\, \boldsymbol{\beta} \in S^{q-1}} \boldsymbol{\alpha}^\top \boldsymbol{A} \boldsymbol{\beta},$$

where S^{p-1} and S^{q-1} are the unit spheres in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively.

By Corollary 4.2.13 of Vershynin (2018), there exist a $1/4$-net $\mathcal{M}$ of S^{p-1} and a $1/4$-net $\mathcal{N}$ of S^{q-1} with cardinalities $|\mathcal{M}| \leq 9^p$ and $|\mathcal{N}| \leq 9^q$. Following Exercise 4.4.3 of Vershynin (2018), we can upper bound the continuous supremum using these discrete nets:

$$\|\boldsymbol{A}\|_2 \leq 2 \sup_{\boldsymbol{\alpha} \in \mathcal{M},\, \boldsymbol{\beta} \in \mathcal{N}} \boldsymbol{\alpha}^\top \boldsymbol{A} \boldsymbol{\beta}.$$

This reduces the problem to bounding the quadratic form for a fixed pair of vectors, followed by taking a union bound over $\mathcal{M} \times \mathcal{N}$.

Step 2: Sub-exponential Concentration for a Fixed Pair. Fix an arbitrary pair $\boldsymbol{\alpha} \in \mathcal{M}$ and $\boldsymbol{\beta} \in \mathcal{N}$. We can write:

$$\boldsymbol{\alpha}^\top \boldsymbol{A} \boldsymbol{\beta} = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{\alpha}^\top \boldsymbol{x}_i)(\boldsymbol{\beta}^\top \boldsymbol{e}_i).$$

Let $\xi_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) := (\boldsymbol{\alpha}^\top \boldsymbol{x}_i)(\boldsymbol{\beta}^\top \boldsymbol{e}_i)$. Since $\boldsymbol{x}_i$ and $\boldsymbol{e}_i$ are independent, $\mathbb{E}[\xi_i(\boldsymbol{\alpha}, \boldsymbol{\beta})] = 0$.

Next, we bound the sub-exponential norm of $\xi_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Since $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2} \boldsymbol{z}_i$ and the components of $\boldsymbol{z}_i$ are independent 1-sub-Gaussian, $\boldsymbol{\alpha}^\top \boldsymbol{x}_i$ is sub-Gaussian with $\|\boldsymbol{\alpha}^\top \boldsymbol{x}_i\|_{\psi_2} \leq C' \sqrt{C_U}$. Similarly, the independence assumption on $\boldsymbol{\mathcal{E}}$ and $\|e_{ij}\|_{\psi_2} \leq \sigma$ yields $\|\boldsymbol{\beta}^\top \boldsymbol{e}_i\|_{\psi_2} \leq C' \sigma$.

The product of two independent sub-Gaussian random variables is a sub-exponential random variable. Thus, there exists an absolute constant $C'' > 0$ such that:

$$\|\xi_i(\boldsymbol{\alpha}, \boldsymbol{\beta})\|_{\psi_1} \leq \|\boldsymbol{\alpha}^\top \boldsymbol{x}_i\|_{\psi_2} \|\boldsymbol{\beta}^\top \boldsymbol{e}_i\|_{\psi_2} \leq C'' \sigma \sqrt{C_U} := K.$$

Step 3: Applying Bernstein's Inequality. Since $\xi_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are independent, mean-zero, sub-exponential random variables, Bernstein's inequality implies that for any $t > 0$:

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} \xi_i(\boldsymbol{\alpha}, \boldsymbol{\beta})\right| \geq t\right) \leq 2 \exp\left[-c_0 N \min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right].$$

Let $t = C_0 \sigma \sqrt{C_U} \sqrt{\frac{p+q+\log(2/\epsilon)}{N}}$, where $C_0 > 0$ is a sufficiently large absolute constant to be determined. Following the sample size assumption 2, we have

$$\Pr\left(|\boldsymbol{\alpha}^\top \boldsymbol{A} \boldsymbol{\beta}| \geq t\right) \leq 2 \exp\left(-c_1 \left(p + q + \log(2/\epsilon)\right)\right),$$

where $c_1 = c_0 C_0^2 / (C'')^2$.

Step 4: Applying Union Bound. Finally, we take a union bound over all pairs in $\mathcal{M} \times \mathcal{N}$:

$$\begin{aligned}\Pr\left(\|\boldsymbol{A}\|_2 \geq 2t\right) &\leq |\mathcal{M}||\mathcal{N}| \max_{\boldsymbol{\alpha}\in\mathcal{M},\,\boldsymbol{\beta}\in\mathcal{N}} \Pr\left(|\boldsymbol{\alpha}^\top \boldsymbol{A}\boldsymbol{\beta}| \geq t\right) \\ &\leq \exp\left((p+q)\log 9\right) \cdot 2\exp\left(-c_1(p+q+\log(2/\epsilon))\right) \\ &= 2\exp\left((p+q)(\log 9 - c_1) - c_1\log(2/\epsilon)\right).\end{aligned}$$

By choosing C_0 large enough such that $c_1 > \log 9$, we have

$$\Pr\left(\|\boldsymbol{A}\|_2 \geq \underbrace{2C_0\sqrt{C_U}}_{:=C_3}\sigma\sqrt{\frac{p+q+\log(2/\epsilon)}{N}}\right) \leq 2\exp\left(-\log(2/\epsilon)\right) = \epsilon.$$

The proof is completed. □

Lemma C.2. *Under Assumptions 1 and 2, for the central machine $\mathcal{I}_1$ with local sample size n, the local sample covariance matrix satisfies the following concentration inequality:*

$$P\left(\left\|\frac{1}{n}\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top\boldsymbol{X}_{\mathcal{I}_1\cdot} - \boldsymbol{\Sigma}\right\|_2 \geq \underbrace{\left(C+\sqrt{2/c}\right)}_{:=C_{X,1}}\sqrt{\frac{p+\log n}{n}}\right) \leq \mathcal{O}(e^{-c\log n}), \tag{C.2}$$

where C, c are two fixed universal positive constants.

Proof. By Remark 5.40 of Vershynin (2010), for any $t > 0$, we have

$$P\left(\left\|\frac{1}{n}\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top\boldsymbol{X}_{\mathcal{I}_1\cdot} - \Sigma\right\|_2 > \max(\iota, \iota^2)\right) \leq 2e^{-ct^2},$$

where $\iota = C\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}$ and C, c are two fixed universal positive constants.

Let $t = \sqrt{2/c}\sqrt{p+\log n}$. Then,

$$\iota = C\sqrt{\frac{p}{n}} + \sqrt{2/c}\sqrt{\frac{p+\log n}{n}} \leq \left(C+\sqrt{2/c}\right)\sqrt{\frac{p+\log n}{n}}.$$

Under the sample size condition in Assumption 2, we obtain

$$P\left(\left\|\frac{1}{n}\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top\boldsymbol{X}_{\mathcal{I}_1\cdot} - \Sigma\right\|_2 \geq \left(C+\sqrt{2/c}\right)\sqrt{\frac{p+\log n}{n}}\right) \leq \mathcal{O}(e^{-c\log n}).$$

This completes the proof. □

Lemma C.3. *Under Assumptions 1 and 2, for a given constant $C_s \geq 1$, there exists an absolute constant $c > 0$ such that the local sub-matrices bounded by the contamination size satisfy:*

$$P\left(\sup_{k\in[m]}\sup_{\mathcal{O}_k\subset\mathcal{I}_k:|\mathcal{O}_k|\leq C_s o_k}\|\boldsymbol{X}_{\mathcal{O}_k\cdot}\|_2 > \underbrace{\sqrt{C_U+2C}}_{:=C_{X,2}}\sqrt{C_s}\sqrt{p+o_k\log N}\right) \leq \mathcal{O}(e^{-c\log N}), \tag{C.3}$$

where C is a fixed universal positive constant.

Proof. For a fixed machine $k \in [m]$ and a fixed row subset $\mathcal{O}_k \subset \mathcal{I}_k$ with cardinality $|\mathcal{O}_k| = C_s o_k$, we consider the sample covariance matrix $\frac{1}{|\mathcal{O}_k|}\boldsymbol{X}_{\mathcal{O}_k\cdot}^\top\boldsymbol{X}_{\mathcal{O}_k\cdot}$. According to Exercise 9.2.5 of Vershynin (2018), for any $u_k > 0$, there exist absolute constants $C, c > 0$ such that

$$P\left(\left\|\frac{1}{|\mathcal{O}_k|}\boldsymbol{X}_{\mathcal{O}_k\cdot}^\top\boldsymbol{X}_{\mathcal{O}_k\cdot} - \Sigma\right\|_2 > C\left(\frac{p}{|\mathcal{O}_k|} \vee \sqrt{\frac{p}{|\mathcal{O}_k|}} \vee \frac{u_k}{|\mathcal{O}_k|} \vee \sqrt{\frac{u_k}{|\mathcal{O}_k|}}\right)\right) \leq 2e^{-u_k}.$$

By setting $u_k = 2C_s(o_k \log N + p)$, we then apply the union bound over all machines $k \in [m]$ and all subsets $\mathcal{O}_k \subset \mathcal{I}_k$ satisfying $|\mathcal{O}_k| \leq C_s o_k$. The total number of such subsets is bounded by $\sum_{k=1}^{m} \binom{n}{C_s o_k}$. Consequently, the uniform failure probability is controlled by

$$\sum_{k=1}^{m} \binom{n}{C_s o_k} 2e^{-u_k} \leq \sum_{k=1}^{m} \left(\frac{en}{C_s o_k}\right)^{C_s o_k} 2\exp\left(-2C_s(o_k \log N + p)\right) \leq \mathcal{O}(e^{-c \log N}).$$

Therefore, with probability at least $1 - \mathcal{O}(e^{-c \log N})$, the following inequality holds uniformly after multiplying by $|\mathcal{O}_k|$:

$$\left\| \boldsymbol{X}_{\mathcal{O}_k \cdot}^\top \boldsymbol{X}_{\mathcal{O}_k \cdot} - |\mathcal{O}_k|\Sigma \right\|_2 \leq 2C \cdot C_s(p + o_k \log N).$$

To bound the spectral norm of $\boldsymbol{X}_{\mathcal{O}_k \cdot}$, we utilize the triangle inequality on the squared spectral norm:

$$\|\boldsymbol{X}_{\mathcal{O}_k \cdot}\|_2^2 = \left\| \boldsymbol{X}_{\mathcal{O}_k \cdot}^\top \boldsymbol{X}_{\mathcal{O}_k \cdot} \right\|_2 \leq \||\mathcal{O}_k|\Sigma\|_2 + \left\| \boldsymbol{X}_{\mathcal{O}_k \cdot}^\top \boldsymbol{X}_{\mathcal{O}_k \cdot} - |\mathcal{O}_k|\Sigma \right\|_2.$$

Given the bounded eigenvalue assumption $\|\Sigma\|_2 \leq C_U$, the expected covariance component is bounded by $\||\mathcal{O}_k|\Sigma\|_2 \leq C_s o_k C_U$. Combining this with the uniform deviation bound yields

$$\|\boldsymbol{X}_{\mathcal{O}_k \cdot}\|_2^2 \leq C_s o_k C_U + 2C \cdot C_s(p + o_k \log N) \leq C_s(C_U + 2C)(p + o_k \log N).$$

Therefore,

$$\|\boldsymbol{X}_{\mathcal{O}_k \cdot}\|_2 \leq \sqrt{C_s}\sqrt{C_U + 2C}\sqrt{p + o_k \log N}.$$

This completes the proof. □

Following the similar argument in Lemma C.3, we immediately given the following lemma.

Lemma C.4. *Under Assumptions 1 and 2, for a given constant $C_s \geq 1$, there exists an absolute constant $c > 0$ such that the local sub-matrices bounded by the contamination size satisfy:*

$$P\left(\sup_{\mathcal{O}_{\text{full}} \subset [N]: |\mathcal{O}_{\text{full}}| \leq (C_s+1)O} \|\boldsymbol{X}_{\mathcal{O}_{\text{full}} \cdot}\|_2 > C_{X,2}\sqrt{C_s + 1}\sqrt{p + O \log N} \right) \leq \mathcal{O}(e^{-c \log N}), \tag{C.4}$$

where C is a fixed universal positive constant.

Lemma C.5. *Under Assumptions 1 and 2, the minimum eigenvalue of the local sample covariance matrix $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n}\boldsymbol{X}_{\mathcal{I}_1 \cdot}^\top \boldsymbol{X}_{\mathcal{I}_1 \cdot}$ satisfies the following tail bound:*

$$P\left(\Lambda_{\min}(\hat{\boldsymbol{\Sigma}}_1) \leq \frac{C_L}{2} \right) \leq \mathcal{O}(e^{-c \log n}). \tag{C.5}$$

Proof. Since both $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n}\boldsymbol{X}_{\mathcal{I}_1 \cdot}^\top \boldsymbol{X}_{\mathcal{I}_1 \cdot}$ are symmetric positive semi-definite matrices, their eigenvalues coincide with their singular values. By Weyl's inequality, we have:

$$\Lambda_{\min}(\hat{\boldsymbol{\Sigma}}_1) \geq \Lambda_{\min}(\boldsymbol{\Sigma}) - \|\hat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}\|_2.$$

Consequently, by Lemma C.2, the spectral norm deviation is bounded by $C_{X,1}\sqrt{\frac{p+\log n}{n}}$ with high probability. Following the sample size assumption 2, we have

$$C_{X,1}\sqrt{\frac{p + \log n}{n}} \leq \frac{C_L}{2}.$$

Then, conditionally on the "good" event

$$\mathcal{A} = \left\{ \|\hat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}\|_2 \leq C_{X,1}\sqrt{\frac{p + \log n}{n}} \right\},$$

Weyl's inequality directly implies:

$$\Lambda_{\min}(\hat{\boldsymbol{\Sigma}}_1) \geq \Lambda_{\min}(\boldsymbol{\Sigma}) - \frac{C_L}{2} \geq C_L - \frac{C_L}{2} = \frac{C_L}{2}.$$

Hence, the occurrence of the "bad" event implies that the event $\mathcal{A}$ must have failed:

$$\left\{\Lambda_{\min}(\hat{\mathbf{\Sigma}}_1) \leq \frac{C_L}{2}\right\} \subseteq \mathcal{A}^c.$$

Therefore, utilizing the concentration probability established in Lemma C.2, we obtain the final bound:

$$P\left(\Lambda_{\min}(\hat{\mathbf{\Sigma}}_1) \leq \frac{C_L}{2}\right) \leq P(\mathcal{A}^c) \leq \mathcal{O}(e^{-c\log n}).$$

This completes the proof. □

Lemma C.6. *Under Assumptions 1 and 2, when the local sample size n is sufficiently large, the spectral norm of the inverse local sample covariance matrix $\hat{\mathbf{\Sigma}}_1^{-1} = (n^{-1}\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top \boldsymbol{X}_{\mathcal{I}_1\cdot})^{-1}$ satisfies the following tail bound:*

$$P\left(\|\hat{\mathbf{\Sigma}}_1^{-1}\|_2 \leq \frac{2}{C_L}\right) \geq 1 - \mathcal{O}(e^{-c\log n}). \tag{C.6}$$

Proof. For any symmetric positive definite matrix, the spectral norm of its inverse is equal to the reciprocal of its smallest eigenvalue:

$$\|\hat{\mathbf{\Sigma}}_1^{-1}\|_2 = \Lambda_{\max}(\hat{\mathbf{\Sigma}}_1^{-1}) = \frac{1}{\Lambda_{\min}(\hat{\mathbf{\Sigma}}_1)}.$$

By Lemma C.5, we have established that with probability at least $1 - \mathcal{O}(e^{-c\log n})$, the minimum eigenvalue is bounded below:

$$\Lambda_{\min}(\hat{\mathbf{\Sigma}}_1) \geq \frac{C_L}{2}.$$

Conditioned on this high-probability event, we have:

$$\|\hat{\mathbf{\Sigma}}_1^{-1}\|_2 = \frac{1}{\Lambda_{\min}(\hat{\mathbf{\Sigma}}_1)} \leq \frac{1}{C_L/2} = \frac{2}{C_L}.$$

This completes the proof. □

Lemma C.7. *Under Assumptions 1 and 2, we have*

$$P\left(\|\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}\|_2 > C_{X,1}\left(\sqrt{\frac{p+\log n}{n}} + \sqrt{\frac{p+\log n}{N}}\right)\right) \leq \mathcal{O}(e^{-c\log n}). \tag{C.7}$$

Proof. By the triangle inequality, the spectral norm of the difference between the local and global sample covariance matrices can be bounded by their respective deviations from the population covariance matrix $\mathbf{\Sigma}$:

$$\|\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}\|_2 = \|\hat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma} + \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_2 \leq \|\hat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}\|_2 + \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2. \tag{C.8}$$

Let $\mathcal{A}_1$ denote this "good" event

$$\mathcal{A}_1 = \left\{\|\hat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}\|_2 \leq C_{X,1}\sqrt{\frac{p+\log n}{n}}\right\}.$$

By Lemma C.2, $\mathcal{A}_1$ holds with probability at least $1 - \mathcal{O}(e^{-c\log n})$.

Let the global "good" event $\mathcal{A}_N$:

$$\mathcal{A}_N = \left\{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 \leq C_{X,1}\sqrt{\frac{p+\log n}{N}}\right\}.$$

Similarly, $\mathcal{A}_N$ holds with probability at least $1 - \mathcal{O}(e^{-c\log n})$.

Conditioned on the intersection of these two events $\mathcal{A}_1 \cap \mathcal{A}_N$, substituting the upper bounds into (C.8) directly yields:

$$\|\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}}\|_2 \leq C_{X,1} \left(\sqrt{\frac{p + \log n}{n}} + \sqrt{\frac{p + \log n}{N}} \right).$$

To bound the probability of the complementary (failure) event, we apply the union bound:

$$P\left((\mathcal{A}_1 \cap \mathcal{A}_N)^c\right) = P(\mathcal{A}_1^c \cup \mathcal{A}_N^c) \leq P(\mathcal{A}_1^c) + P(\mathcal{A}_N^c) \leq \mathcal{O}(e^{-c \log n}).$$

Consequently, the stated concentration inequality holds with probability at least $1 - \mathcal{O}(e^{-c \log n})$. This completes the proof. □

Lemma C.8. *Suppose $\mathcal{I}_k \subset [N]$ for blocks $k \in [m]$. Then, there exist absolute constants $C', c > 0$ such that:*

$$P\left(\sup_{k \in [m]} \sup_{\substack{\mathcal{O}_k \subset \mathcal{I}_k \\ |\mathcal{O}_k| \leq C_s o_k}} \|\boldsymbol{\mathcal{E}}_{\mathcal{O}_k \cdot}\|_F > C' \sigma \sqrt{C_s o_k} (\sqrt{q} + \sqrt{\log N}) \right) \leq \mathcal{O}(e^{-c \log N}). \tag{C.9}$$

Proof. For any $k \in [m]$ and $\mathcal{O}_k \subset \mathcal{I}_k$ with $|\mathcal{O}_k| \leq C_s o_k$, we have

$$\|\boldsymbol{\mathcal{E}}_{\mathcal{O}_k \cdot}\|_F^2 = \sum_{i \in \mathcal{O}_k} \|\boldsymbol{e}_i\|_2^2 \leq |\mathcal{O}_k| \max_{i \in [N]} \|\boldsymbol{e}_i\|_2^2 \leq C_s o_k \max_{i \in [N]} \|\boldsymbol{e}_i\|_2^2.$$

Taking square roots on both sides yields

$$\|\boldsymbol{\mathcal{E}}_{\mathcal{O}_k \cdot}\|_F \leq \sqrt{C_s o_k} \max_{i \in [N]} \|\boldsymbol{e}_i\|_2. \tag{C.10}$$

Therefore,

$$\sup_{k \in [m]} \sup_{\substack{\mathcal{O}_k \subset \mathcal{I}_k \\ |\mathcal{O}_k| \leq C_s o_k}} \|\boldsymbol{\mathcal{E}}_{\mathcal{O}_k \cdot}\|_F \leq \sqrt{C_s o_k} \max_{i \in [N]} \|\boldsymbol{e}_i\|_2.$$

By Theorem 3.1.1 in Vershynin (2018), there exists an absolute constant $C_1 > 0$ such that for any $i \in [N]$ and $t > 0$:

$$P\left(\|\boldsymbol{e}_i\|_2 > C_1 \sigma (\sqrt{q} + t)\right) \leq 2 \exp(-t^2).$$

Applying a union bound over $i \in [N]$:

$$P\left(\max_{i \in [N]} \|\boldsymbol{e}_i\|_2 > C_1 \sigma (\sqrt{q} + t) \right) \leq 2N \exp(-t^2).$$

Choosing $t = C_2 \sqrt{\log N}$ for some $C_2 > 1$, the failure probability satisfies

$$2N \exp(-C_2^2 \log N) = 2 \exp((1 - C_2^2) \log N) = \mathcal{O}(e^{-c \log N}),$$

where $c = C_2^2 - 1 > 0$. With probability at least $1 - \mathcal{O}(e^{-c \log N})$,

$$\max_{i \in [N]} \|\boldsymbol{e}_i\|_2 \leq C_1 \sigma (\sqrt{q} + C_2 \sqrt{\log N}) \leq C' \sigma (\sqrt{q} + \sqrt{\log N}), \tag{C.11}$$

where $C' = C_1 \max(1, C_2)$. Combining equation C.10 and equation C.11, we conclude that, with probability at least $1 - \mathcal{O}(e^{-c \log N})$,

$$\sup_{k \in [m]} \sup_{\substack{\mathcal{O}_k \subset \mathcal{I}_k \\ |\mathcal{O}_k| \leq C_s o_k}} \|\boldsymbol{\mathcal{E}}_{\mathcal{O}_k \cdot}\|_F \leq C' \sigma \sqrt{C_s o_k} (\sqrt{q} + \sqrt{\log N}).$$

The proof is complete. □

Lemma C.9. *Under Assumption 1, there exist absolute constants $C, c > 0$ such that*

$$\Pr\left(\|\boldsymbol{X}_{\mathcal{I}_1 \cdot}\|_2 > \underbrace{C \sqrt{C_U}}_{:= C_{X,3}} \left(\sqrt{n} + \sqrt{p} + \sqrt{\log n} \right) \right) \leq \mathcal{O}(e^{-c \log n}). \tag{C.12}$$

Proof. Since $\boldsymbol{X}_{\mathcal{I}_1\cdot} \overset{d}{=} \boldsymbol{Z}_{\mathcal{I}_1\cdot}\boldsymbol{\Sigma}^{1/2}$, we have

$$\|\boldsymbol{Z}_{\mathcal{I}_1\cdot}\boldsymbol{\Sigma}^{1/2}\|_2 \le \|\boldsymbol{Z}_{\mathcal{I}_1\cdot}\|_2 \, \|\boldsymbol{\Sigma}^{1/2}\|_2 \le \sqrt{C_U}\, \|\boldsymbol{Z}_{\mathcal{I}_1\cdot}\|_2.$$

By Theorem 4.6.1 in Vershynin (2018), there exist absolute constants $C, c > 0$ such that for any $t > 0$,

$$\Pr\big(\|\boldsymbol{Z}_{\mathcal{I}_1\cdot}\|_2 > C(\sqrt{n} + \sqrt{p} + t)\big) \le 2e^{-ct^2}.$$

Therefore,

$$\Pr\Big(\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\|_2 > C\sqrt{C_U}\,\big(\sqrt{n} + \sqrt{p} + t\big)\Big) \le 2e^{-ct^2}.$$

Setting $t = \sqrt{\log n}$ yields

$$\Pr\Big(\|\boldsymbol{X}_{\mathcal{I}_1\cdot}\|_2 > C\sqrt{C_U}\,\Big(\sqrt{n} + \sqrt{p} + \sqrt{\log n}\Big)\Big) \le 2n^{-c}.$$

This completes the proof. □

D Comparison of Full-Sample and Local Hessians and Motivation for the Surrogate Hessian

In the contaminated-data setting, the higher-order derivatives of $\mathcal{L}_{\text{full}}$ cannot be well approximated by those of $\mathcal{L}_1$. Specifically, for $\mathcal{L}_{\text{full}}$, its high-order derivatives can be written as

$$\begin{aligned}&\mathcal{L}_{\text{full}}(\bar{\boldsymbol{B}}) - \mathcal{L}_{\text{full}}(\bar{\boldsymbol{B}}^-) - \langle \nabla_{\bar{\boldsymbol{B}}}\mathcal{L}_{\text{full}}(\bar{\boldsymbol{B}}^-), \bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-\rangle \\ &\quad = \frac{1}{2}\text{tr}\{(\bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-)^\top \boldsymbol{H}_{\text{full}}(\bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-)\},\end{aligned} \tag{D.1}$$

where

$$\boldsymbol{H}_{\text{full}} = \begin{pmatrix} \frac{\boldsymbol{X}^\top\boldsymbol{X}}{N} & \frac{\boldsymbol{X}^\top}{\sqrt{N}} \\ \frac{\boldsymbol{X}}{\sqrt{N}} & \boldsymbol{I}_N \end{pmatrix}. \tag{D.2}$$

In contrast, the high-order derivatives of $\mathcal{L}_1$ is

$$\begin{aligned}&\mathcal{L}_1(\bar{\boldsymbol{B}}) - \mathcal{L}_1(\bar{\boldsymbol{B}}^-) - \langle \nabla_{\bar{\boldsymbol{B}}}\mathcal{L}_1(\bar{\boldsymbol{B}}^-), \bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-\rangle \\ &= \frac{1}{2}\text{tr}\{(\bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-)^\top \boldsymbol{H}(\bar{\boldsymbol{B}} - \bar{\boldsymbol{B}}^-)\},\end{aligned} \tag{D.3}$$

where

$$\boldsymbol{H} = \begin{pmatrix} \frac{\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top\boldsymbol{X}_{\mathcal{I}_1\cdot}}{n} & \frac{\sqrt{N}\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top}{n} & \boldsymbol{0} \\ \frac{\sqrt{N}\boldsymbol{X}_{\mathcal{I}_1\cdot}}{n} & m\boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}. \tag{D.4}$$

A direct comparison of equation D.2 and equation D.4 highlights notable structural and scaling differences. Specifically, the non-zero blocks in $\boldsymbol{H}$ exhibit inconsistent scaling compared to the full sample (e.g., $m\boldsymbol{I}_n$ versus $\boldsymbol{I}_N$), and the zero block entirely misses curvature information regarding the contamination parameters $C^*_{\mathcal{I}_k\cdot}, k = 2, 3, \ldots, m$.

Notice that if we assume the rows of $\boldsymbol{X}$ are i.i.d. mean-zero with covariance $\boldsymbol{\Sigma}$, the expectation of the full-sample Hessian is a block-diagonal matrix

$$\mathbb{E}(\boldsymbol{H}_{\text{full}}) = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_N \end{pmatrix}. \tag{D.5}$$

Motivated by this observation, we consider the following block-diagonal matrix,

$$\bar{\boldsymbol{H}} = \begin{pmatrix} \frac{\boldsymbol{X}_{\mathcal{I}_1\cdot}^\top\boldsymbol{X}_{\mathcal{I}_1\cdot}}{n} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_N \end{pmatrix}, \tag{D.6}$$

which coincides with the expectation of the full-sample Hessian.

E ADDITIONAL EXPERIMENTS

This section supplements the main text with empirical validation of the theoretical dependence on the rank and contamination proportion, results under Laplace and t_3 noise, evaluations of outlier-row detection, experiments varying the signal strength, statistical significance tests, computational time comparisons, and results on the remaining real datasets.

E.1 EMPIRICAL VALIDATION OF THEORETICAL DEPENDENCE

Figure E.1 shows that the estimation error increases with the rank and contamination proportion, consistent with the dependence characterized by our theoretical results (Theorem 4).

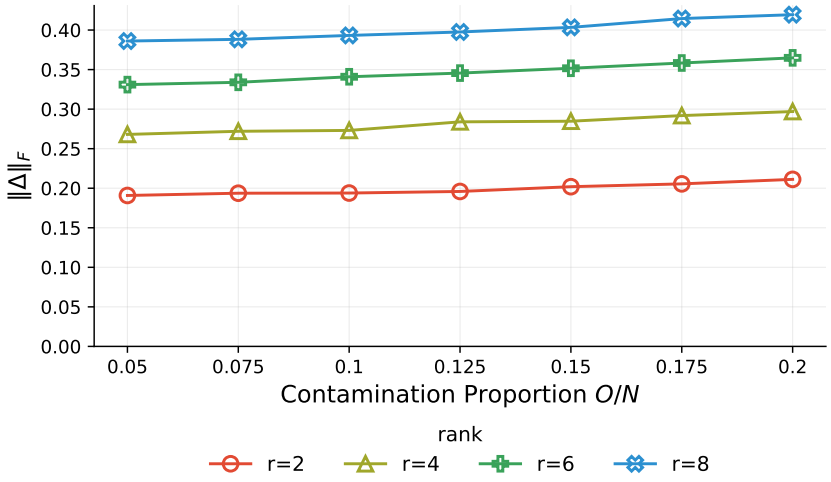

Figure E.1: Estimation error of D4R as rank and contamination increase. We use the basic simulation setup with $m = 20$, $n = 200$, $p = 15$, $q = 40$, $a = 1$, and Gaussian noise. Each point averages 100 replications.

E.2 EFFECT OF NOISE TYPES

Table E.1 reports the estimation results under Laplace and t_3 noise, complementing the Normal and Cauchy results in the main text. D4R matches the Full-sample oracle up to rounding across all metrics and correctly recovers the true rank, whereas the competing methods are substantially less accurate.

Table E.1: Estimation accuracy of all competing methods under Laplace and t_3 noise. Entries are mean (SE) over 100 replications; bold marks the best non-oracle result.

	Laplace				t_3			
Method	$\|\mathbf{\Delta}\|_F$	$\|\mathbf{\Delta}\|_\infty$	$\|\mathbf{\Delta}\|_\Sigma$	$\hat{r}$	$\|\mathbf{\Delta}\|_F$	$\|\mathbf{\Delta}\|_\infty$	$\|\mathbf{\Delta}\|_\Sigma$	$\hat{r}$
Full-sample	0.238(0.002)	0.038(0.001)	0.208(0.001)	3.0(0.0)	0.213(0.002)	0.034(0.001)	0.186(0.001)	3.0(0.0)
D4R	**0.240**(0.002)	**0.038**(0.001)	**0.210**(0.001)	**3.0**(0.0)	**0.242**(0.002)	**0.039**(0.001)	**0.210**(0.001)	**3.0**(0.0)
DC-D4R	0.384(0.012)	0.058(0.002)	0.313(0.008)	4.7(0.2)	0.378(0.011)	0.060(0.002)	0.311(0.008)	4.4(0.2)
Dist-CSL	3.117(0.244)	0.509(0.047)	2.624(0.198)	3.5(0.1)	3.242(0.242)	0.525(0.046)	2.706(0.196)	3.7(0.1)
Dist-QRR	2.522(0.127)	0.533(0.037)	2.004(0.100)	3.2(0.1)	2.354(0.121)	0.484(0.029)	1.876(0.095)	3.2(0.1)
Dist-RRR	5.411(0.141)	0.749(0.029)	4.242(0.113)	15.0(0.0)	5.411(0.145)	0.748(0.029)	4.243(0.116)	15.0(0.0)

Beyond estimation accuracy, we also examine how well each method recovers the support of the row-sparse contamination matrix. Table E.2 reports the precision, recall, and F_1 scores for outlier detection by the three methods that estimate the contamination matrix $\boldsymbol{C}$ (D4R, DC-D4R, and Dist-CSL). Under Normal, Laplace, and t_3 noise, D4R achieves near-perfect outlier detection, with precision, recall, and F_1 scores all close to 1. DC-D4R performs similarly, indicating that the local

hard-thresholding step reliably identifies contaminated observations. Under heavy-tailed Cauchy noise, D4R and DC-D4R retain near-perfect recall, but their precision decreases to approximately 0.53–0.55. This result suggests that both methods identify nearly all contaminated observations but also flag some uncontaminated observations with extreme Cauchy noise as outliers. In contrast, Dist-CSL detects only a small proportion of the contaminated observations across all noise distributions.

Table E.2: Precision, recall, and F_1 score (mean(SE) over 100 replications) of outlier-row detection under four noise distributions, for the three methods that estimate the contamination matrix $\boldsymbol{C}$; bold marks the best value per metric within each noise block.

Noise	Method	Precision	Recall	F_1
Normal	D4R	**1.0000**(0.0000)	**0.9997**(0.0001)	**0.9998**(0.0000)
	DC-D4R	1.0000(0.0000)	0.9960(0.0003)	0.9980(0.0001)
	Dist-CSL	0.9012(0.0153)	0.0294(0.0008)	0.0568(0.0016)
Laplace	D4R	**1.0000**(0.0000)	**0.9997**(0.0001)	**0.9998**(0.0000)
	DC-D4R	1.0000(0.0000)	0.9962(0.0003)	0.9981(0.0001)
	Dist-CSL	0.8816(0.0167)	0.0293(0.0008)	0.0565(0.0016)
t_3	D4R	**1.0000**(0.0000)	**0.9997**(0.0001)	**0.9998**(0.0000)
	DC-D4R	1.0000(0.0000)	0.9960(0.0003)	0.9980(0.0001)
	Dist-CSL	0.8954(0.0157)	0.0294(0.0008)	0.0568(0.0016)
Cauchy	D4R	0.5255(0.0012)	**0.9997**(0.0001)	0.6888(0.0010)
	DC-D4R	**0.5538**(0.0012)	0.9972(0.0002)	**0.7120**(0.0010)
	Dist-CSL	0.2543(0.0071)	0.0281(0.0008)	0.0503(0.0014)

E.3 Effect of contamination proportion

We vary the contamination proportion O/N over $\{0.05, 0.10, 0.15, 0.20\}$ under Normal and Cauchy noise. Table E.3 reports the Frobenius estimation error $\|\boldsymbol{\Delta}\|_F$ of all competing methods. Under Normal noise, D4R remains close to the Full-sample oracle, with its error increasing only slightly from 0.233 to 0.260 as O/N increases from 0.05 to 0.20, whereas the competing methods deteriorate substantially. Under Cauchy noise, D4R remains comparatively stable, with errors ranging from 1.49 to 1.66. In contrast, the errors of Dist-CSL and Dist-RRR exceed 5 and 9, respectively, at $O/N = 0.20$. D4R also correctly recovers the true rank $\hat{r} = 3$ at every contamination proportion.

Table E.3: Frobenius estimation error $\|\boldsymbol{\Delta}\|_F$ of all methods across different contamination proportions under Normal and Cauchy noise. Entries are mean (SE) over 100 replications; bold marks the best non-oracle result.

Noise	O/N	Full-sample	D4R	DC-D4R	Dist-CSL	Dist-QRR	Dist-RRR
Normal	0.05	0.230(0.002)	**0.233**(0.002)	0.369(0.010)	1.698(0.041)	2.174(0.141)	3.661(0.040)
	0.10	0.238(0.002)	**0.240**(0.002)	0.394(0.011)	3.283(0.244)	2.342(0.131)	5.411(0.143)
	0.15	0.246(0.002)	**0.249**(0.002)	0.405(0.011)	4.261(0.269)	2.488(0.174)	6.345(0.106)
	0.20	0.254(0.002)	**0.260**(0.002)	0.436(0.014)	5.634(0.840)	2.629(0.140)	7.447(0.311)
Cauchy	0.05	0.744(0.006)	**1.489**(0.011)	3.222(0.012)	8.553(4.846)	2.742(0.178)	7.255(0.523)
	0.10	0.768(0.006)	**1.552**(0.014)	2.482(0.089)	3.953(0.203)	2.779(0.128)	8.002(0.199)
	0.15	0.784(0.006)	**1.604**(0.015)	2.171(0.052)	4.360(0.249)	2.805(0.191)	8.625(0.204)
	0.20	0.810(0.007)	**1.661**(0.015)	2.186(0.052)	5.342(0.845)	3.149(0.170)	9.526(0.400)

E.4 Effect of signal strength

Figure E.2 reports the absolute (top) and relative (bottom) estimation errors as the signal strength a increases from 0.1 to 0.6, under both Normal and Cauchy noise. D4R remains close to the Full-sample oracle across the entire signal range in both absolute and relative terms. Dist-QRR exhibits nearly constant or slightly decreasing relative error, but its absolute error increases with a across all three metrics. Dist-CSL and Dist-RRR remain substantially less accurate throughout.

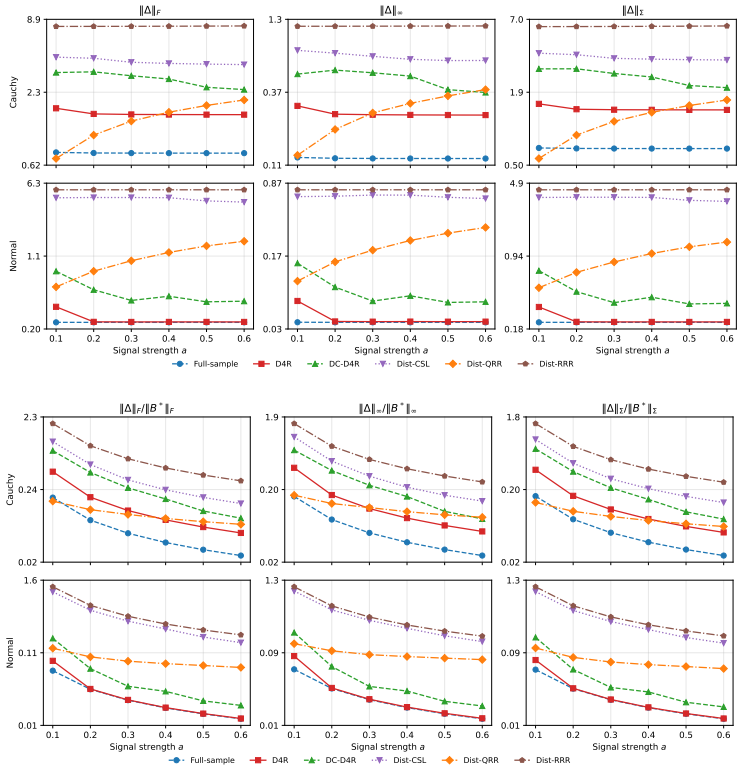

Figure E.2: Absolute and relative estimation errors of all methods across different signal strengths a under Normal and Cauchy noise. The upper panels report $\|\mathbf{\Delta}\|_F$, $\|\mathbf{\Delta}\|_\infty$ and $\|\mathbf{\Delta}\|_\Sigma$, whereas the bottom panels report their corresponding relative errors.

E.5 Statistical significance

To assess whether D4R's improvements are statistically significant, we apply the two-sided Wilcoxon signed-rank test to the paired per-replication errors of D4R versus each competitor (100 paired observations per cell). Tables E.4 and E.5 report the p-values. All improvements are significant at the $p < 0.001$ level except a single cell (D4R v.s. DC-D4R under Cauchy noise at $m = 100$, $p = 0.196$), where the two methods converge to comparable accuracy as the total sample size grows large.

Table E.4: Wilcoxon signed-rank test p-values comparing D4R with each competing method under four noise distributions, based on $\|\mathbf{\Delta}\|_F$ over 100 paired replications.

Noise	DC-D4R	Dist-CSL	Dist-QRR	Dist-RRR
Normal	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
Laplace	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
t_3	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
Cauchy	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-14}$	$< 10^{-17}$

E.6 Computational time comparison

Table E.6 reports the total runtime of all methods as the number of machines increases under Normal and Cauchy noise.

Table E.5: Wilcoxon signed-rank test p-values comparing D4R with each competing method across different numbers of machines under Normal and Cauchy noise, based on $\|\mathbf{\Delta}\|_F$ over 100 paired replications.

Noise	m	DC-D4R	Dist-CSL	Dist-QRR	Dist-RRR
Normal	10	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
	15	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
	20	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
	35	$< 10^{-15}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
	50	$< 10^{-11}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
	75	$< 10^{-15}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
	100	$< 10^{-15}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
Cauchy	10	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-15}$	$< 10^{-17}$
	15	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-15}$	$< 10^{-17}$
	20	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-14}$	$< 10^{-17}$
	35	$< 10^{-13}$	$< 10^{-17}$	$< 10^{-12}$	$< 10^{-17}$
	50	$< 10^{-8}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
	75	$< 10^{-4}$	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$
	100	0.196	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-17}$

Table E.6: Computational time (in seconds) of all methods across different numbers of machines under Normal and Cauchy noise.

	Normal				Cauchy			
Method	$m = 10$	$m = 20$	$m = 50$	$m = 100$	$m = 10$	$m = 20$	$m = 50$	$m = 100$
Full-sample	0.027(0.003)	0.052(0.003)	0.206(0.004)	0.739(0.005)	0.116(0.003)	0.268(0.003)	1.284(0.009)	6.050(0.053)
D4R	0.560(0.015)	0.844(0.027)	1.751(0.065)	3.229(0.123)	0.527(0.005)	0.700(0.014)	1.358(0.041)	2.358(0.070)
DC-D4R	0.937(0.022)	2.024(0.047)	5.341(0.120)	11.060(0.253)	0.649(0.079)	1.176(0.147)	2.448(0.299)	4.388(0.529)
Dist-CSL	1.166(0.036)	3.078(0.059)	7.397(0.238)	15.694(0.507)	3.460(0.034)	6.361(0.038)	11.440(0.077)	20.034(0.109)
Dist-QRR	0.834(0.008)	1.094(0.033)	1.285(0.017)	2.044(0.016)	1.147(0.014)	1.125(0.014)	1.398(0.019)	2.300(0.039)
Dist-RRR	0.161(0.015)	0.228(0.016)	0.254(0.009)	0.618(0.034)	0.175(0.013)	0.201(0.012)	0.321(0.012)	3.369(0.524)

E.7 Additional real-data results

Table E.7 reports the test MSE of all competing methods on Facebook Live and Facebook Metrics across different contamination proportions. D4R has the lowest or tied-lowest MSE at every contamination level and remains stable as O/N increases.

Table E.7: Test MSE of all competing methods on Facebook Live and Facebook Metrics across different contamination proportions. Entries are mean (SE) over 100 train/test splits; bold marks the best reported value in each row.

Dataset	O/N	D4R	DC-D4R	Dist-CSL	Dist-QRR	Dist-RRR
Facebook Live	0.05	**0.836**(0.003)	0.836(0.003)	0.839(0.005)	0.867(0.003)	1.171(0.019)
($N_{\text{total}} = 7050$	0.10	**0.836**(0.003)	0.836(0.003)	1.526(0.032)	0.868(0.003)	1.496(0.033)
$p = 7,\ q = 8$	0.15	**0.835**(0.003)	0.835(0.003)	1.831(0.055)	0.868(0.003)	1.737(0.048)
$m = 5$)	0.20	**0.836**(0.003)	0.836(0.003)	2.466(0.120)	0.875(0.004)	2.154(0.078)
Facebook Metrics	0.05	**0.839**(0.011)	0.849(0.009)	0.884(0.048)	0.860(0.009)	1.309(0.048)
($N_{\text{total}} = 495$	0.10	**0.837**(0.010)	0.854(0.011)	0.851(0.016)	0.861(0.009)	1.912(0.114)
$p = 9,\ q = 12$	0.15	**0.834**(0.009)	0.856(0.010)	0.880(0.041)	0.865(0.009)	2.174(0.119)
$m = 3$)	0.20	**0.839**(0.009)	0.866(0.011)	0.846(0.011)	0.866(0.009)	2.546(0.120)

F COMPUTING ENVIRONMENT

All experiments were conducted on a personal computer running macOS 15.7 (64-bit, Apple silicon), equipped with an Apple M2 Pro chip (10-core CPU) and 16 GB of memory. All computations were performed on the CPU. The implementation was developed in Python 3.9 using NumPy (≥ 1.26), SciPy (≥ 1.13), and Matplotlib (≥ 3.9).