Reduced-Rank L1-Norm Principal-Component Analysis With Performance Guarantees

Hossein Kamrani, Alireza Zolghadr Asli, *Member, IEEE*, Panos P. Markopoulos, *Member, IEEE*, Michael Langberg, *Senior Member, IEEE*, Dimitris A. Pados, *Senior Member, IEEE*, and George N. Karystinos, *Member, IEEE*

***Abstract*—Standard Principal-Component Analysis (PCA) is known to be sensitive to outliers among the processed data. On the other hand, L1-norm-based PCA (L1-PCA) exhibits sturdy resistance against outliers, while it performs similar to standard PCA when applied to nominal or smoothly corrupted data [1]. Exact calculation of the K L1-norm Principal Components (L1-PCs) of a rank-r data matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ costs $\mathcal{O}(N^{(r-1)K+1})$ [1], [2]. In this work, we present reduced-rank L1-PCA (RR L1-PCA): a hybrid approach that approximates the K L1-PCs of $\mathbf{X}$ by the L1-PCs of its L2-norm-based rank-d approximation ($d \leq r$), calculable exactly with reduced complexity $\mathcal{O}(N^{(d-1)K+1})$. The proposed method combines the denoising capabilities and low computation cost of standard PCA with the outlier-resistance of L1-PCA. RR L1-PCA is accompanied by formal performance guarantees as well as thorough numerical studies that corroborate its computational and corruption resistance merits.**

***Index Terms*—Faulty data, L1-norm, matrix analysis, PCA, outliers.**

Manuscript received September 6, 2019; revised February 16, 2020, May 27, 2020, and September 28, 2020; accepted November 3, 2020. Date of publication November 23, 2020; date of current version January 5, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Justin Dauwels. This work was supported in part by the U.S. National Science Foundation under Grants 1808582, 1462341, 1526771, 1704813, and 1828181. The work of Panos P. Markopoulos was supported by the U.S. National Science Foundation under Grant 1808582. The work of Michael Langberg was supported by the U.S. National Science Foundation under Grants 1526771 and 1462341. The work of Dimitris A. Pados was supported by the U.S. National Science Foundation under Grants 1704813 and 1828181. *(Corresponding author: Panos P. Markopoulos.)*

Hossein Kamrani was with the University at Buffalo, SUNY, Buffalo, NY 14260 USA. He is now with the Department of Communication and Electronics Engineering, Shiraz University, Shiraz, Iran (e-mail: kamrani@shirazu.ac.ir).

Alireza Zolghadr Asli is with the Department of Communication and Electronics Engineering, Shiraz University, Shiraz 7134851154, Iran (e-mail: zolghadr@shirazu.ac.ir).

Panos P. Markopoulos is with the Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY 14623 USA (e-mail: pxmeee@rit.edu).

Michael Langberg is with the Department of Electrical Engineering, State University of New York at Buffalo, University at Buffalo, Buffalo, NY 14260 USA (e-mail: mikel@buffalo.edu).

Dimitris A. Pados is with the Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431 USA (e-mail: dpados@fau.edu).

George N. Karystinos is with the School of Electrical and Computer Engineering, Technical Univerisity of Crete, Chania 73100, Greece (e-mail: karystinos@telecom.tuc.gr).

Digital Object Identifier 10.1109/TSP.2020.3039599

I. Introduction

PRINCIPAL-COMPONENT Analysis (PCA) finds numerous applications in the fields of signal processing, image processing, communications, computer vision, machine learning, and bio-informatics/genomics, to name a few [3]–[7]. PCA seeks a low-dimensional linear subspace that maximizes data presence, traditionally measured by data variance and estimated via the L2-norm of the projected data. Therefore, PCA is also known as L2-PCA. In an array of applications, PCA has been shown to attain high denoising performance when low-rank data are corrupted by benign (e.g., Gaussian) noise. Another advantage of PCA is its low-cost implementation by means of standard singular-value decomposition (SVD) and faster practical variants [8]. Owning to these merits, PCA is widely used today for general denoising, data compression, visualization, clustering, detection, and classification [6].

On the negative side, PCA is well-known to suffer significant performance degradation when the processed data include irregular points that lie far from the true/nominal subspace (i.e., the low-rank data subspace prior to any contamination) [9]. Such points, commonly referred to as *outliers*, may appear due to sensor malfunctions, errors in data transmission/transcription, or heavy-tail noise corruption [10], [11]. This sensitivity can be emphasized when outliers form a high-variance low-rank subspace [12] –a case of particular interest in this work. The outlier sensitivity of PCA derives from its L2-norm-maximization definition, by which it places squared emphasis on every datum and benefits unfavorably peripheral entries.

In view of the above, in the past decade there has been increased research interest for the development of L1-norm-based Principal Component (L1-PC) calculators. Existing L1-norm-based PCA formulations seek to either minimize the absolute subspace representation error (L1-error-minimization) or maximize the absolute magnitude of the subspace-projected data points (L1-projection-maximization). For the first approach (error minimization), there has been thorough theoretical analysis in the literature [13]–[19] and efficient/iterative algorithms have been presented [20]–[24]. However, for a general number of principal components, no exact solution exists to date. For the second approach (projection maximization), henceforth referred to simply as *L1-PCA*, the exact solution is known for any number of components [1], [2]. In addition, an array of approximate algorithms have been proposed in the literature with varying performance/cost trade-offs [25]–[34]. The outlier resistance of L1-PCA has already been leveraged in many applications, such as direction-of-arrival (DoA) estimation, interference-suppressive filtering, image restoration, face recognition, and

video surveillance, to name a few [35]–[41]. More recently, L1-PCA has also been extended to tensor processing [42]–[46].

Authors in [1] showed that L1-PCA can be equivalently formulated as a combinatorial optimization problem, in the form of nuclear norm maximization over antipodal-binary variables. In addition, [1] offered the first optimal algorithms that compute the K L1-PCs of matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$, with complexity $\mathcal{O}(2^{NK})$, in the general case and $\mathcal{O}(N^{(r-1)K+1})$ when $r := \text{rank}(\mathbf{X})$ is a constant with respect to N. The complexity of these optimal algorithms often renders them unsuitable for several real-world applications where N and/or D are large. The complexity of low-cost approximate solvers (such as those proposed in [30] and [28]) also depends on r. Therefore, we are motivated to design an L1-PCA solver that operates on a low-rank approximation of the data matrix.

In fact, one could conjecture that if $\mathbf{X}_d$ is a rank-d matrix that constitutes a close approximation to $\mathbf{X}$, for some $d \in \{K+1, \ldots, r\}$, then its L1-PCs would also be similar to those of $\mathbf{X}$. In this work, we examine this conjecture for the first time and prove its truthfulness when $\mathbf{X}_d$ is the projection of $\mathbf{X}$ on its d-dimensional L2-norm principal subspace –i.e., the subspace spanned by the d most significant PCs of $\mathbf{X}$, obtain by standard SVD. Accordingly, the K L1-PCs of $\mathbf{X}_d$ can be computed exactly with reduced cost $\mathcal{O}(N^{(d-1)K+1})$ [1] or approximately by any of the algorithms in [25]–[28], [30] with cost as low as $\mathcal{O}(N^2 dK)$. We refer to the proposed method as *Reduced-Rank L1-PCA* (RR L1-PCA). Our algorithmic developments are accompanied by performance bounds that solely depend on the size and singular-value profile of $\mathbf{X}$ –and are, thus, calculable by means of SVD, prior to any L1-PCA computation.

The remainder of this paper is organized as follows. In Section II, we briefly review the L1-PCA theory. In Section III, we present the proposed RR L1-PCA approach. In Section IV, we present formal performance bounds for the proposed method. Section V holds an array of experimental results from the fields of data analysis, image reconstruction, and direction-of-arrival estimation. Finally, concluding remarks and acknowledgements are presented in Sections VI and VII, respectively.

This work is a significantly extended version of the conference publication [47] by P.P.M, D.A.P., G.N.K., and M.L. Major extensions include improved performance bounds, presented in Section IV, and extended numerical studies, presented in Section V.

II. L1-PCA THEORY

We consider matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$ of rank $r \leq \min\{D, N\}$ and denote by $\sigma_i(\mathbf{X})$ its i-th highest singular value. The L2-norm, L1-norm, and nuclear norm of $\mathbf{X}$ are given by

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{D}\sum_{j=1}^{N} [\mathbf{X}]_{i,j}^2} = \sqrt{\text{Tr}(\mathbf{X}^\top \mathbf{X})} = \sqrt{\sum_{i=1}^{r} \sigma_i^2(\mathbf{X})},$$

$$\|\mathbf{X}\|_1 = \sum_{i=1}^{D}\sum_{j=1}^{N} |[\mathbf{X}]_{i,j}|, \text{ and}$$

$$\|\mathbf{X}\|_* = \text{Tr}(\sqrt{\mathbf{X}^\top \mathbf{X}}) = \sum_{i=1}^{r} \sigma_i(\mathbf{X}), \tag{1}$$

respectively. Given a desired number of $K < r$ PCs, the L1-PCA of $\mathbf{X}$ is defined as the pursuit of an orthonormal basis $\mathbf{Q}^* \in \mathbb{S}_{D,K} := \{\mathbf{U} \in \mathbb{R}^{D \times K};\ \mathbf{U}^\top \mathbf{U} = \mathbf{I}_K\}$ that solves

$$\underset{\mathbf{Q} \in \mathbb{S}_{D,K}}{\text{maximize}} \ \|\mathbf{X}^\top \mathbf{Q}\|_1. \tag{2}$$

Authors in [1] showed that L1-PCA in the form of (2) is equivalent to the combinatorial optimization over NK antipodal binary variables. In addition, [1] offered the first two exact algorithms for solving (2). Prior to these exact algorithms, several approximate calculators had been proposed in the literature [25]–[28]. In the sequel, we briefly present the optimal solution to L1-PCA, in the form of Theorem 1, as derived in [1].

Theorem 1: (L1-PCA Theorem [1]) Let $\mathbf{B}^*$ be a solution to

$$\underset{\mathbf{B} \in \{\pm 1\}^{N \times K}}{\text{maximize}} \ \|\mathbf{X}\mathbf{B}\|_*. \tag{3}$$

If $\mathbf{X}\mathbf{B}^*$ admits SVD

$$\mathbf{X}\mathbf{B}^* \overset{\text{svd}}{=} \mathbf{Y}\mathbf{\Sigma}\mathbf{Z}^\top, \tag{4}$$

then an exact solution to (2) is given by

$$\mathbf{Q}^* = \mathbf{Y}\mathbf{Z}^\top = \Phi(\mathbf{X}\mathbf{B}^*) := \underset{\mathbf{Q} \in \mathbb{S}_{D,K}}{\text{argmin}} \ \|\mathbf{X}\mathbf{B}^* - \mathbf{Q}\|_F. \tag{5}$$

Moreover, it holds that $\|\mathbf{X}^\top \mathbf{Q}^*\|_1 = \|\mathbf{X}\mathbf{B}^*\|_*$ and $\mathbf{B}^* = \text{sgn}(\mathbf{X}^\top \mathbf{Q}^*)$, where $\text{sgn}(\cdot)$ returns the $\{\pm 1\}$-sign matrix of its argument.

Accordingly, for $K = 1$, (3) takes the form[1]

$$\underset{\mathbf{b} \in \{\pm 1\}^N}{\text{maximize}} \ \|\mathbf{X}\mathbf{b}\|_2 \tag{6}$$

and, given the optimal solution $\mathbf{b}^*$ to (6), the L1-PC of $\mathbf{X}$ is

$$\mathbf{q}^* = \mathbf{X}\mathbf{b}^* \|\mathbf{X}\mathbf{b}^*\|_2^{-1}. \tag{7}$$

In addition, $\|\mathbf{X}^\top \mathbf{q}^*\|_1 = \|\mathbf{X}\mathbf{b}^*\|_2$ and $\mathbf{b}^* = \text{sgn}(\mathbf{X}^\top \mathbf{q}^*)$.

In view of the above connection of L1-PCA to combinatorial optimization, a conceptually simple optimal algorithm[1] searches exhaustively the size-2^{NK} feasibility set $\{\pm 1\}^{N \times K}$ of (3) for $\mathbf{B}^*$ and then returns $\mathbf{Q}^*$ by (5).[2] Authors in [1] offered an alternative polynomial-time algorithm that searches for $\mathbf{B}^*$ in a subset of $\{\pm 1\}^{N \times K}$ (wherein $\mathbf{B}^*$ is guaranteed to exist) with overall cost $\mathcal{O}(N^{(r-1)K+1})$, when r is considered fixed with respect to N, which is a reasonable assumption for most cases of engineering interest where D is the number of sensors or data-features and, thus, constant with respect to the number of measurements, N.

III. REDUCED-RANK L1-PCA

A. Proposed Method

Consider SVD $\mathbf{X} \overset{\text{svd}}{=} \mathbf{U}\mathbf{D}_{r \times r}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{S}_{D,r}$, $\mathbf{V} \in \mathbb{S}_{N,r}$, and $\mathbf{D} = \text{diag}([\sigma_1(\mathbf{X}), \sigma_2(\mathbf{X}), \ldots, \sigma_r(\mathbf{X})]^\top)$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors of $\mathbf{X}$, respectively. $\mathbf{D}$ contains the r positive singular values of $\mathbf{X}$. Moreover, for any $i, j \in [r] := \{1, 2, \ldots, r\}$, with $i < j$, we define $\mathbf{U}_{i \to j} := [\mathbf{U}]_{:,i:j}$, $\mathbf{V}_{i \to j} := [\mathbf{V}]_{:,i:j}$, $\mathbf{D}_{i \to j} := [\mathbf{D}]_{i:j,i:j}$,

[1] The nuclear norm, $\|\cdot\|_*$, and Euclidean norm, $\|\cdot\|_2$, operators coincide for vector arguments.

[2] In practice, this algorithm can take advantage of the nuclear-norm invariability to negations and permutations of the columns of its matrix argument and search exhaustively in a size-$\binom{2^{N-1}+K-1}{K}$ subset of $\{\pm 1\}^{N \times K}$, wherein a solution to (3) is guaranteed to exist. Still, this reduced set has asymptotic size $\mathcal{O}(2^{NK})$.

Algorithm 1: Proposed Framework for RR L1-PCA. `L1-PCA`$(\mathbf{X}_d, K)$ Returns the K Exact [2] (or Approximate –e.g., [25]–[30]) L1-PCs of $\mathbf{X}_d$.

Data: Measurement matrix $\mathbf{X}$; number of PCs K; reduced rank $d \geq K$
Result: Basis $\mathbf{Q}_d^*$
$\mathbf{U}\mathbf{D}\mathbf{V}^\top \overset{\text{SVD}}{\leftarrow} \mathbf{X}$;
$\mathbf{X}_d \leftarrow \mathbf{U}_{1\to d}\mathbf{U}_{1\to d}^\top \mathbf{X}$;
$\mathbf{Q}_d^* \leftarrow$ `L1-PCA`$(\mathbf{X}_d, K)$

and $\mathbf{D}_j := \text{diag}([\sigma_1(\mathbf{X}), \sigma_2(\mathbf{X}), \ldots, \sigma_j(\mathbf{X}), \ \mathbf{0}_{r-j}^\top]^\top)$. The rank-$d$ L2-norm approximation of matrix $\mathbf{X}$ is then given by $\mathbf{X}_d := \mathbf{U}_{1\to d}\mathbf{U}_{1\to d}^\top \mathbf{X} = \mathbf{U}_{1\to d}\mathbf{U}_{1\to d}^\top \mathbf{U}\mathbf{D}\mathbf{V}^\top = \mathbf{U}_{1\to d}\mathbf{D}_{1\to d}\mathbf{V}_{1\to d}^\top = \mathbf{U}\mathbf{D}_d\mathbf{V}^\top$. In this work, we propose to approximate the exact L1-PCs of $\mathbf{X}$, $\mathbf{Q}^*$, by the L1-PCs of its rank-d approximation, $\mathbf{X}_d$, defined as

$$\mathbf{Q}_d^* = \underset{\mathbf{Q} \in \mathbb{S}_{D,K}}{\text{argmax}} \ \|\mathbf{X}_d^\top \mathbf{Q}\|_1, \tag{8}$$

for $d \geq K$. The proposed RR L1-PCA framework is presented in Algorithm 1.

The rationale behind the proposed framework is the following: if $\mathbf{X}_d$ is close to $\mathbf{X}$, depending on the magnitude of the omitted singular values $\sigma_{d+1}(\mathbf{X}), \ldots, \sigma_r(\mathbf{X})$, $\mathbf{Q}_d^*$ will be a close approximation to $\mathbf{Q}^*$. The polynomial-time algorithm of [1] can solve (8) with reduced cost $\mathcal{O}(N^{(d-1)K+1})$. The near-optimal algorithms of [30] can solve (8) with cost $\mathcal{O}(N^2K^2(K^2+d))$. The algorithms of [26] and [28] can approximate the solution to (8) with cost $\mathcal{O}(N^2dK)$.[3] In practice, we can set d to be greater-or-equal to the dimensionality of the dominant subspace in $\mathbf{X}$ (i.e., the number of significantly large singular values), capturing both inliers and outliers –which can be then suppressed by L1-PCA.

The following Lemma 1 shows that the L1-PCA of a matrix that has rank lower than its number of rows can be recast as the L1-PCA of a shorter matrix of full row rank. A proof for the lemma is provided in the appendix.

Lemma 1: Consider matrix $\mathbf{X} \in \mathbb{R}^{D\times N}$ with rank r and $\mathbf{X} \overset{\text{svd}}{=} \mathbf{U}\mathbf{D}_{r\times r}\mathbf{V}^\top$. If $\bar{\mathbf{Q}}^*$ are the $K < r$ L1-PCs of $\mathbf{D}\mathbf{V}^\top \in \mathbb{R}^{r\times N}$, then $\mathbf{U}\bar{\mathbf{Q}}^*$ are the K L1-PCs of $\mathbf{X}$.

According to Lemma 1, we can obtain the K L1-PCs of $\mathbf{X}_d$ in (8) (recall $K \leq d \leq r$) by those of $\mathbf{D}_{1\to d}\mathbf{V}_{1\to d}^\top$. That is, $\mathbf{Q}_d^* = \mathbf{U}_{1\to d}\bar{\mathbf{Q}}_d^*$, where $\bar{\mathbf{Q}}_d^*$ are the K L1-PCs of $\mathbf{D}_{1\to d}\mathbf{V}_{1\to d}^\top$.

B. Further Refinement

The proposed RR L1-PCA method approximates the solution to the L1-PCA problem in (2) $\mathbf{Q}^*$, by the solution to (8), $\mathbf{Q}_d^*$, for some $d \in \{K, K+1, \ldots, r\}$.

As an optional *refinement*, we can run on $\mathbf{Q}_d^*$ the alternating optimization of [28], for any desired number of iterations, with practically negligible additional cost. Specifically, consider the reduced-rank approximate solution $\mathbf{Q}_d^* \in \mathbb{S}_{D,K}$, denote it by $\mathbf{Q}_{\text{ref}}^{(0)}$, for the purpose of this refinement step, and define

$$\mathbf{B}_{\text{ref}}^{(1)} = \text{sgn}(\mathbf{X}^\top \mathbf{Q}_{\text{ref}}^{(0)}). \tag{9}$$

[3]The presented complexities for the iterative algorithms of [26] and [28] consider $N \geq d \geq K$ and that the number of iterations is upper bounded by a linear function of N, which is corroborated by our numerical experiments.

Algorithm 2: Alternating-Optimization Iterations for the Refinement of $\mathbf{Q}_d^*$ [28].

Data: Approximate RR solution $\mathbf{Q}_d^*$ and $\mathbf{X}$
Result: Refined basis $\mathbf{Q}_{\texttt{ref}}^{(i)}$
Initialize: $\mathbf{B}_{\texttt{ref}}^{(1)} \leftarrow \texttt{sgn}(\mathbf{X}^\top \mathbf{Q}_d^*)$, $i \leftarrow 1$;
while *not termination criterion* **do**
  $i \leftarrow i+1$;
  $(\mathbf{Y}^{(i)}, \mathbf{\Sigma}_{K\times K}^{(i)}, \mathbf{Z}^{(i)}) \leftarrow \texttt{svd}(\mathbf{X}\mathbf{B}_{\texttt{ref}}^{(i-1)})$;
  $\mathbf{Q}_{\texttt{ref}}^{(i)} \leftarrow \mathbf{Y}^{(i)}(\mathbf{Z}^{(i)})^\top$;
  $\mathbf{B}_{\texttt{ref}}^{(i)} \leftarrow \texttt{sgn}(\mathbf{X}^\top \mathbf{Q}_{\texttt{ref}}^{(i)})$;
end

Then, for $i = 1, 2, \ldots$, until the termination condition is met, obtain iteratively

$$\mathbf{Q}_{\text{ref}}^{(i)} = \mathbf{\Phi}(\mathbf{X}\mathbf{B}_{\text{ref}}^{(i)}). \tag{10}$$

For this iterative refinement, two important properties hold: (i) for any $i \geq 1$, $\mathbf{Q}_{\text{ref}}^{(i)}$ will offer a greater (or equal) value to (2), than $\mathbf{Q}_{\text{ref}}^{(i-1)}$ and (ii) the iterations converge. To prove the increase to (2), we calculate

$$\begin{aligned}
\|\mathbf{X}^\top \mathbf{Q}_{\text{ref}}^{(i)}\|_1 &= \max_{\mathbf{B}\in\{\pm 1\}^{N\times K}} \text{Tr}\left(\mathbf{B}^\top \mathbf{X}^\top \mathbf{Q}_{\text{ref}}^{(i)}\right) \\
&\geq \text{Tr}\left((\mathbf{B}_{\text{ref}}^{(i)})^\top \mathbf{X}^\top \mathbf{Q}_{\text{ref}}^{(i)}\right) = \max_{\mathbf{Q}\in\mathbb{S}_{D,K}} \text{Tr}\left((\mathbf{B}_{\text{ref}}^{(i)})^\top \mathbf{X}^\top \mathbf{Q}\right) \\
&\geq \text{Tr}\left((\mathbf{B}_{\text{ref}}^{(i)})^\top \mathbf{X}^\top \mathbf{Q}_{\text{ref}}^{(i-1)}\right) \\
&= \|\mathbf{X}^\top \mathbf{Q}_{\text{ref}}^{(i-1)}\|_1.
\end{aligned} \tag{11}$$

A proof similar to (11) was also presented in [28]. Regarding the convergence to the metric, it is evident since the metric is upper bounded by $\|\mathbf{X}^\top \mathbf{Q}^*\|_1$ and the refinement iterations increase it at every step. Thus, if the refinement iterations terminate at some step t, the algorithm returns $\mathbf{Q}_{\text{ref}}^{(t)}$ instead of $\mathbf{Q}_d^*$, for which it holds

$$\|\mathbf{X}^\top \mathbf{Q}_{\text{ref}}^{(t)}\|_1 \geq \|\mathbf{X}^\top \mathbf{Q}_d^*\|_1. \tag{12}$$

A pseudocode for this refinement is offered in Algorithm 2.

C. Performance Evaluation

Next, we focus on formally evaluating the proposed L1-PCA approximation by means of the following criteria.

1) Metric Approximation: For any $\mathbf{Q} \in \mathbb{S}_{D,K}$, we define the approximation error to the L1-PCA metric, when $\mathbf{X}_d$ is employed instead of $\mathbf{X}$, as
$$\gamma_d(\mathbf{Q}; \mathbf{X}) := \|\mathbf{X}^\top \mathbf{Q}\|_1 - \|\mathbf{X}_d^\top \mathbf{Q}\|_1. \tag{13}$$
Certainly, if $\gamma_d(\mathbf{Q}; \mathbf{X}) = 0$ for every $\mathbf{Q}$, (8) coincides with (2). If, for a given value of d, $\gamma_d(\mathbf{Q}; \mathbf{X})$ takes low values for every $\mathbf{Q}$, then we can also expect $\|\mathbf{X}^\top \mathbf{Q}_d^*\|_1$ to be close to $\|\mathbf{X}^\top \mathbf{Q}^*\|_1$.[4]
2) Solution Approximation: We define the approximation accuracy to the L1-PCA metric in (2), when the approximate

[4]Specifically, $\|\mathbf{X}^\top \mathbf{Q}^*\|_1 - \|\mathbf{X}^\top \mathbf{Q}_d^*\|_1 \leq \|\mathbf{X}^\top \mathbf{Q}^*\|_1 - \|\mathbf{X}^\top \mathbf{Q}_d^*\|_1 + \|\mathbf{X}_d^\top \mathbf{Q}_d^*\|_1 - \|\mathbf{X}_d^\top \mathbf{Q}^*\|_1 = \gamma_d(\mathbf{Q}^*; \mathbf{X}) - \gamma_d(\mathbf{Q}_d^*; \mathbf{X})$.

solution $\mathbf{Q}_d^*$ is employed instead of $\mathbf{Q}^*$, as

$$\rho_d(\mathbf{X}) := \frac{\|\mathbf{X}^\top \mathbf{Q}_d^*\|_1}{\|\mathbf{X}^\top \mathbf{Q}^*\|_1} \leq 1. \tag{14}$$

Understandably, if $\rho_d(\mathbf{X})$ is close to 1, then we say that $\mathbf{Q}_d^*$ is near-optimal. In addition, we define the performance gap to the L1-PCA metric in (2), when $\mathbf{Q}_d^*$ is employed instead of $\mathbf{Q}^*$, as

$$\xi_d(\mathbf{X}) := \|\mathbf{X}^\top \mathbf{Q}^*\|_1 - \|\mathbf{X}^\top \mathbf{Q}_d^*\|_1. \tag{15}$$

3) Computational Efficiency: The cost for obtaining $\mathbf{X}_d$ from $\mathbf{X}$ should not cancel out the computational savings claimed by solving (8) instead of (2).

IV. Bounds on Metric and Solution Approximation

A. Metric Approximation

For any $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $i \leq j \leq \min\{n,p\}$, we define $\lambda_{i\to j}(\mathbf{A}) := \sqrt{\sum_{k=i}^{j} \sigma_k^2(\mathbf{A})}$. The following new Theorem 2 applies to any $\mathbf{X} \in \mathbb{R}^{D\times N}$ and presents formal lower and upper bounds for $\gamma_d(\mathbf{Q};\mathbf{X})$, for any $\mathbf{Q} \in \mathbb{S}_{D,K}$. The proof is offered in the Appendix.

Theorem 2: For any $\mathbf{X} \in \mathbb{R}^{D\times N}$, $\mathbf{Q} \in \mathbb{S}_{D,K}$, and (d,K) such that $K \leq d \leq r = \text{rank}(\mathbf{X})$, it holds that

$$\lambda_{h+1\to r}(\mathbf{X}) - \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) \leq \gamma_d(\mathbf{Q};\mathbf{X}) \leq \min\left\{\sqrt{NK}\lambda_{\min(d+1,r)\to\min(K+d,r)}(\mathbf{X}), \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) - \lambda_{h+1\to d}(\mathbf{X})\right\}, \tag{16}$$

where $h := \min\{r, D-K\}$. Moreover, if $h+1 > r$, then $\lambda_{h+1\to r}(\mathbf{X}) = 0$, while if $h+1 > d$, then $\lambda_{h+1\to d}(\mathbf{X}) = 0$.

Importantly, Theorem 2 offers bounds for $\gamma_d(\mathbf{Q};\mathbf{X})$ that solely depend on the singular values of $\mathbf{X}$ and, thus, are computable by SVD of $\mathbf{X}$.

B. Solution Approximation

Below, we show that the solution approximation accuracy $\rho_d(\mathbf{X})$ may remain desirably high even for small values of d. Thus, the proposed reduced-rank approximation can claim significant computational benefits at a negligible performance cost. Moreover, in this section, we provide formal bounds for $\rho_d(\mathbf{X})$ and $\xi_d(\mathbf{X})$, calculable by SVD of $\mathbf{X}$.

a) Special cases $d = r$ and $d = K = 1$: It is clear that when $d = r$ (no rank reduction), $\mathbf{X}_d = \mathbf{X}$ and, thus, $\rho_d(\mathbf{X}) = 1$ and $\xi_d(\mathbf{X}) = 0$. On the other hand, when $d = K = 1$, $\mathbf{X}_1 = \sigma_1(\mathbf{X})\mathbf{u}_1\mathbf{v}_1^\top$, where $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$ are the right-hand-side singular vectors of $\mathbf{X}$ corresponding to the singular values $\sigma_1(\mathbf{X}) \geq \sigma_2(\mathbf{X}) \geq \cdots \geq \sigma_r(\mathbf{X})$. Accordingly, $\|\mathbf{X}_1^\top \mathbf{q}\|_1 = \sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1 |\mathbf{u}_1^\top \mathbf{q}|$ and the solution to

$$\underset{\mathbf{q}\in\mathbb{S}_{D,1}}{\text{maximize}} \ \left\|\mathbf{X}_1^\top \mathbf{q}\right\|_1 \tag{17}$$

is $\mathbf{q}_1^* = \mathbf{u}_1$. That is, $\mathbf{q}_1^*$ coincides with the standard PC of $\mathbf{X}$ and the maximum value attained in (17) is $\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1$. Accordingly, $\rho_d(\mathbf{X}) = 1$ and $\xi_d(\mathbf{X}) = 0$.

b) General case $K \leq d \leq r$: The following Theorem 3 provides novel bounds for the approximation accuracy parameters $\rho_d(\mathbf{X})$ and $\xi_d(\mathbf{X})$, attained by the proposed method. A proof is provided in the Appendix.

Theorem 3: It holds that

$$1 \geq \rho_d(\mathbf{X}) \geq \max\left\{\frac{\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1}{\sqrt{N}\lambda_{1\to K}(\mathbf{X})}, \frac{1}{1+\frac{\sqrt{\min\{(r-d),K\}N}\sigma_{d+1}(\mathbf{X})}{\lambda_{1\to d}(\mathbf{X})}}, \frac{1}{1+\frac{\sqrt{\min\{(r-d),K\}N}\sigma_{d+1}(\mathbf{X})}{\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1}}, \frac{\lambda_{1\to d}(\mathbf{X})}{\sqrt{N}\lambda_{1\to K}(\mathbf{X})}\right\} \tag{18}$$

and

$$0 \leq \xi_d(\mathbf{X}) \leq \min\left\{\sqrt{\min\{(r-d),K\}NK}\,\sigma_{d+1}(\mathbf{X}), \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) - \sqrt{K}\,\sigma_1(\mathbf{X})\,\|\mathbf{v}_1\|_1, \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) - \sqrt{K}\lambda_{1\to d}(\mathbf{X})\right\}. \tag{19}$$

Similar to Theorem 2, Theorem 3 offers bounds for the solution approximation that depend only on the singular values of $\mathbf{X}$ and d. Therefore, through SVD on $\mathbf{X}$, one can determine the reduced-dimension d that allows for acceptable bounds. It is also worth noting that (18) formalizes our initial conjecture that, if $\sigma_{d+1}(\mathbf{X})$ is negligibly small so that $\mathbf{X}_d$ practically coincides with $\mathbf{X}$, then $\rho_d(\mathbf{X})$ approaches 1 and the proposed RR L1-PCA tends to optimality.

Table I presents straightforwardly derived inequalities for the bounds in (18) and (19), for 5 selected conditions on the size and SVD profile of $\mathbf{X}$. In view of the bound inequalities for each condition, we can significantly simplify (18) and (19). For instance, for $\lambda_{1\to d}(\mathbf{X}) \leq \sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1$ (condition of first row of Table I), we can simplify the lower bound in (18) as

$$\max\left\{\frac{\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1}{\sqrt{N}\lambda_{1\to K}(\mathbf{X})}, \frac{1}{1+\frac{\sqrt{\min\{(r-d),K\}N}\sigma_{d+1}(\mathbf{X})}{\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1}}\right\}. \tag{20}$$

For ease in reading Table I, we denote the four terms in the max in (18) as $\{\rho_i\}_{i=1}^4$ (with index in the order of appearance) and the terms in the min in (19) as $\{\xi_i\}_{i=1}^3$ (with index in the order of appearance). For instance, $\xi_3 = \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) - \sqrt{K}\lambda_{1\to d}(\mathbf{X})$.

Table I offers some insights on the bounds and how they relate to the nature of the data matrix. To promote understanding, let us consider the following scenaria of interest.

Scenario 1: We consider that the nominal data have a single dominant dimension and that the noise variance is low. Then, $\sigma_1(\mathbf{X})$ will be close to $\lambda_{1\to d}(\mathbf{X})$. Moreover, since the dominant dimension is represented across all measurements, the right-hand singular vector $\mathbf{v}_1$ will be balanced (not sparse) and its L1 norm will be tending to its upper bound, $\sqrt{N}$.[5] Thus, we expect that the condition of row 1 in Table I will be active (instead of row 2), which in turn implies $\rho_1 \geq \rho_4$ and $\rho_2 \leq \rho_3$. Moreover, by row 5 of Table I, we can expect that $\rho_d(\mathbf{X})$ will be close to 1, for every value of d. Moreover, assuming negligibly low $\sigma_{d+1}(\mathbf{X})/\sigma_1(\mathbf{X})$ (which is valid for low noise variance) and $\|\mathbf{v}_1\|_1/\sqrt{N} < 1$, row 4 of Table I will be active (instead of row

[5]It holds that $\|\mathbf{v}_1\|_2 = 1 \leq \|\mathbf{v}_1\|_1 = \sqrt{N} = \sqrt{N}\|\mathbf{v}_1\|_2$.

TABLE I
COMPARISON OF THE BOUNDS FOR $\rho_d(\mathbf{X})$ AND $\xi_d(\mathbf{X})$ IN THEOREM 3, UNDER DISTINCT CONDITIONS ON THE SIZE AND SVD OF $\mathbf{X}$

Conditions on size and SVD of $\mathbf{X}$	Bounds for $\rho_d(\mathbf{X})$	Bounds for $\xi_d(\mathbf{X})$
$\lambda_{1\to d}(\mathbf{X}) \leq \sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1$	$\rho_1 \geq \rho_4,\ \rho_2 \leq \rho_3$	$\xi_2 \leq \xi_3$
$\lambda_{1\to d}(\mathbf{X}) > \sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1$	$\rho_1 < \rho_4,\ \rho_2 > \rho_3$	$\xi_2 > \xi_3$
$\lambda_{1\to K}(\mathbf{X}) \leq \frac{\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1}{\sqrt{N}} + \sqrt{\min\{(r-d), K\}}\ \sigma_{d+1}(\mathbf{X})$	$\rho_1 \geq \rho_3$	$\xi_1 \geq \xi_2$
$\lambda_{1\to K}(\mathbf{X}) > \frac{\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1}{\sqrt{N}} + \sqrt{\min\{(r-d), K\}}\ \sigma_{d+1}(\mathbf{X})$	$\rho_1 < \rho_3$	$\xi_1 < \xi_2$
$\lambda_{1\to K}(\mathbf{X}) = \sigma_1(\mathbf{X})$ and $\|\mathbf{v}_1\|_1 = \sqrt{N}$	$\rho_d(\mathbf{X}) = 1$	$\xi_d(\mathbf{X}) = 0$

3) and $\rho_1 < \rho_3$. Combining the conditions of row 1 and row 4, we find $\rho_3 > \rho_1 \geq \rho_4$ and $\rho_3 \geq \rho_2$. Thus, in this scenario, ρ_3 is expected to be the bound closest to $\rho_d(\mathbf{X})$, for any d.

Scenario 2: In this scenario, we consider the same nominal data as above, but high noise variance. Now the singular values of $\mathbf{X}$ are much more spread out and row 2 can be active instead of row 1 (the same will hold if the nominal data describe a high-rank subspace, even for low/intermediate noise variance). For $K = 1$, in accordance with the nominal data dimensionality, $\lambda_{1\to K}(\mathbf{X}) = \sigma_1(\mathbf{X})$. In addition, since both noise and data are uniformly represented across all measurements, $\|\mathbf{v}_1\|_1$ will again be near $\sqrt{N}$ and $\sigma_{d+1}(\mathbf{X})/\sigma_1(\mathbf{X})$ can be non-negligible. Thus, in this scenario, row 3 of Table I can be active instead of row 4.

Scenario 3: In this third scenario, we consider intermediate/low noise variance, nominal data of any rank, and an outlier of very high variance in one (or few) of the measurements. Determined by the outlier, $\sigma_1(\mathbf{X})$ will be dominant and comparable to $\lambda_{1\to d}(\mathbf{X})$. In addition, since the outlier variance is in just few measurements, $\mathbf{v}_1$ will be sparse and $\|\mathbf{v}_1\|_1$ can approach 1. Accordingly, by (18), ρ_1 can be close to ρ_4 and that ρ_2 will be close to ρ_3. Moreover, for low noise variance, $\sigma_{d+1}(\mathbf{X})/\sigma_1(\mathbf{X})$ can be low. Thus, row 4 of Table I can be active instead of row 3, so that $\rho_3 > \rho_1$. In this scenario, as the noise variance increases, $\sigma_{d+1}(\mathbf{X})/\sigma_1(\mathbf{X})$ and $\|\mathbf{v}_1\|_1$ can increase and, thus, ρ_2 and ρ_3 may fall under ρ_1.

In the three exemplary scenarios discussed above (corroborated by our numerical studies in Section V), we see how the bounds relate to the nature of the measurements at hand. Similar intuitions can be drawn from the bounds of $\xi_d(\mathbf{X})$.

V. EXPERIMENTAL STUDIES

A. Solution Approximation

In this study, we study the performance of RR methods on the L1-PCA solution approximation criterion, for varying number of data points, N. We consider measurement matrix $\mathbf{X} = \mathbf{X}_{\text{nom}} + \mathbf{O} + \mathbf{N} \in \mathbb{R}^{(D=6)\times N}$ such that: (i) the columns of $\mathbf{X}_{\text{nom}}$ are drawn from $\mathcal{N}(\mathbf{0}, \mathbf{U}\text{diag}([100, 9])\mathbf{U}^\top)$ for some $\mathbf{U} \in \mathbb{S}_{6,2}$; (ii) each column of outlier matrix $\mathbf{O}$ is equal to $\mathbf{0}_D$ with probability 85% and otherwise drawn from $\mathcal{N}(\mathbf{0}, \mathbf{U}_\text{o}\,\text{diag}([20, 5\mathbf{1}_5^\top])^2\mathbf{U}_\text{o}^\top)$ for some $\mathbf{U}_\text{o} \in \mathbb{S}_{6,6}$ such that $[\mathbf{U}_\text{o}^\top\mathbf{U}]_{1,1} = 0.1$ (the exact orientations of the inlier and outlier subspaces are not as important as their relative position); (iii) the entries of noise matrix $\mathbf{N}$ are drawn from $\mathcal{N}(0, \sigma^2 = 100)$. We conduct PCA on the corrupted data, looking for the dominant component ($K = 1$) by which we estimate the line of $[\mathbf{U}]_{:,1}$. In Fig. 1, we plot the average solution approximation ratio $\rho_d(\mathbf{X})$ versus N, over 300 independent realizations, for standard PCA, RPCA [9] (solved by means of alternating direction methods of multipliers [48]),

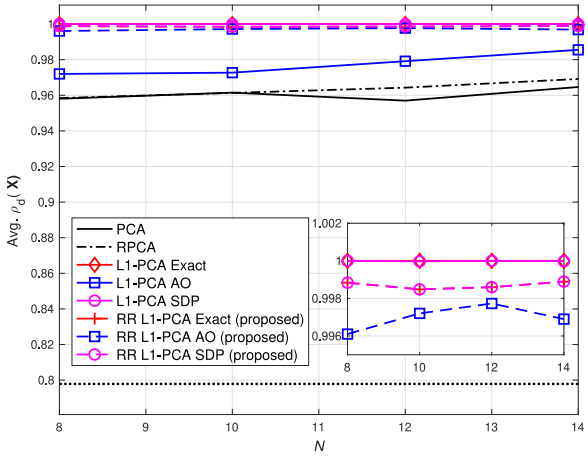

Fig. 1. Average solution approximation ratio $\rho_d(\mathbf{X})$ versus N.

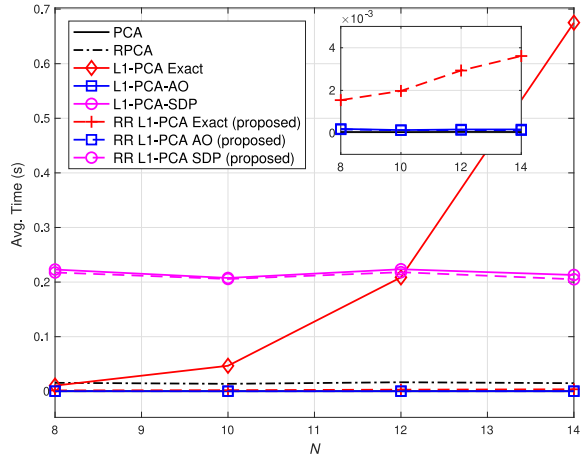

Fig. 2. Average component-analysis time versus N.

exact L1-PCA (solved by the algorithm of [1]), L1-PCA by means of alternating-optimization (L1-PCA AO) [28] (which coincides with the fixed-point method (L1-PCA FP) of [25] for $K = 1$), L1-PCA by means of semi-definite programming (L1-PCA SDP) [27], and the RR ($d = 3$) counterparts of all L1-PCA methods. Expectedly, for exact L1-PCA, the performance is a flat line at 1. Moreover, all methods attain performance above 95%. L1-PCA SDP (full-rank and RR), RR L1-PCA Exact, and RR L1-PCA AO attain approximation ratio close to 1. L1-PCA AO closely follows. PCA and RPCA (which are not designed to solve L1-PCA) follow, with performance from 95% to 97%. In the same figure, we also present the L1-PCA SDP lower-bound benchmark $\sqrt{2/\pi}$ (dotted line), as presented in [27]. In Fig. 3,

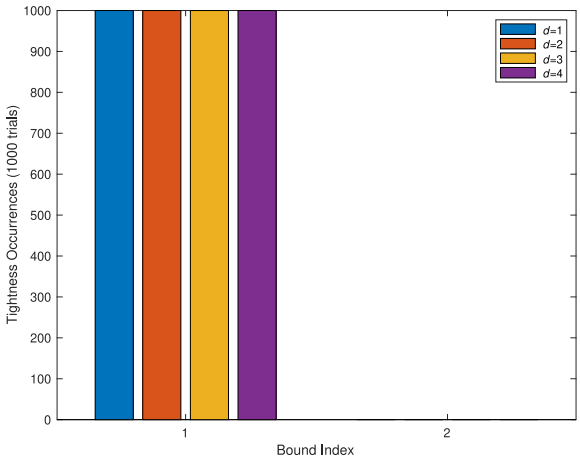

Fig. 3. Bound tightness occurrences for $\gamma_d(\mathbf{Q};\mathbf{X})$.

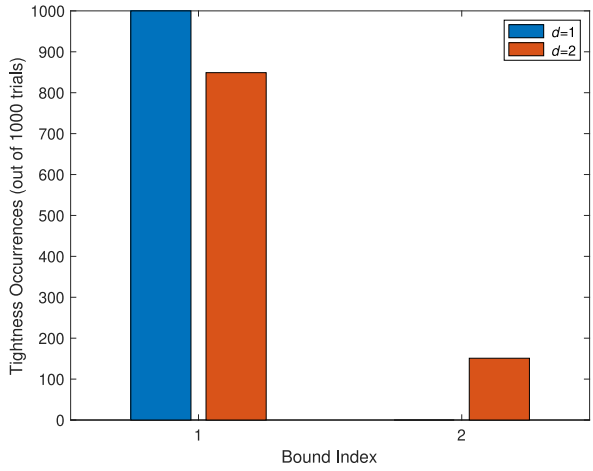

Fig. 4. Bound tightness occurrences for $\gamma_d(\mathbf{Q};\mathbf{X})$.

we plot the average computation time for the same methods.[6] We notice that rank-reduction shortens execution time, most emphatically for exact L1-PCA. Interestingly, for $d = 3$, exact RR L1-PCA attains near-optimal performance at markedly reduced computational cost.

B. Bound Tightness

We continue with an empirical study on the tightness of the bounds presented in the previous section (i.e., (16), (18), and (19)). By tightness, we refer to the proximity of a bound to the actual value of the metric and it is not to be confused with the bound tangibility.

We commence our study with the bounds for $\gamma_d(\mathbf{Q};\mathbf{X})$ in (16) and denote by γ_i the i-th bound in (16) (in order of appearance). We consider 1000 independent realizations of matrix $\mathbf{X} \in \mathbb{R}^{D\times N}$, for $D = N = \text{rank}(\mathbf{X}) = 5$. The entries of $\mathbf{X}$ are independently drawn from $\mathcal{N}(0,1)$. In Fig. 3, we plot the number of times that each bound in (16) is the tightest, for $d = 1,2,3,4$ and $K = 1$. We notice that γ_1 is the tightest bound in all 1000 realizations. Next, we repeat this study for $D = N = \text{rank}(\mathbf{X}) = 2$ and $d = 1,2$ and we plot the results in Fig. 4. We now notice that, for $d = 1$, γ_2 becomes the tightest bound 15% of the time.

Next, we study the bounds for $\rho_d(\mathbf{X})$ in (18). Once again, we consider $\mathbf{X} \in \mathbb{R}^{D\times N}$, with $D = N = \text{rank}(\mathbf{X}) = 5$ and entries from $\mathcal{N}(0,1)$. In Fig. 5 we plot the tightness occurrences of the bounds in (18) in 1000 independent realizations. We notice that ρ_1 is tight most frequently for $d = 1,2$. Then, for $d = 3$, ρ_1 is tight with frequency 80% and ρ_3 is tight with frequency 20%. For $d = 4$, ρ_3 is the tightest with frequency 75% and ρ_1 is the tightest with frequency 25%. Interestingly, for every tested value of d, ρ_2 and ρ_4 are the tightest bounds quite infrequently.

Next, we examine the effect of the SVD profile of $\mathbf{X}$ in the bounds of (18). To that end, for any given realization of $\mathbf{X}$, we define matrix $\mathbf{Y}$ by substituting $\sigma_1(\mathbf{X})$ with $10\sigma_2(\mathbf{X})$. Specifically, after computing the SVD $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ such that $\mathbf{D} = \text{diag}([\sigma_1(\mathbf{X}), \sigma_2(\mathbf{X}), \ldots, \sigma_r(\mathbf{X})]^\top)$, we define $\mathbf{Y} = \mathbf{U}\mathbf{Z}\mathbf{V}^\top$ with $\mathbf{Z} = \text{diag}([10\sigma_2(\mathbf{X}), \sigma_2(\mathbf{X}), \ldots, \sigma_r(\mathbf{X})]^\top)$. In Fig. 6, we present the number of times that each bound in (18) is the tightest, for $d = 1,2,3,4$, and $K = 1$. We notice that this plot is quite different than that of Fig. 5 since now ρ_3 is the tightest bound with much higher frequency.

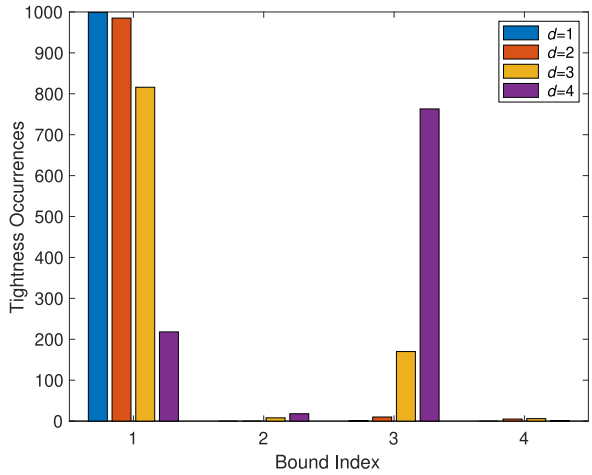

Fig. 5. Bound tightness occurrences for $\rho_d(\mathbf{X})$.

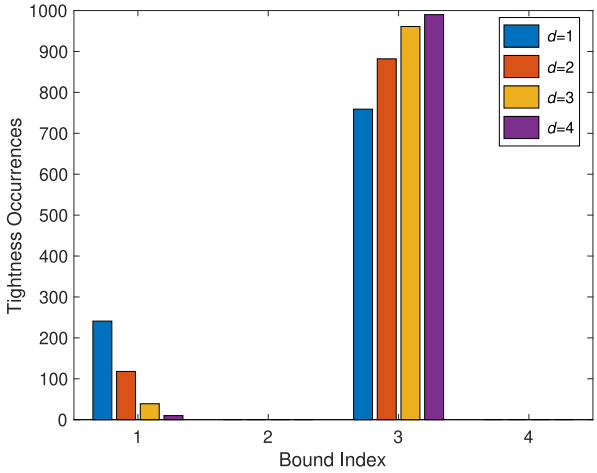

Fig. 6. Bound tightness occurrences for $\rho_d(\mathbf{Y})$.

To obtain a better insight on how the nature of the measurements (nominal data, noise, outliers) affects the tightness of the bounds, we also conduct the following studies. We consider

[6]Timed studies were conducted using MATLAB 2019a, on an Intel(R) Core(TM) i7-4790 CPU 3.60GHz, with RAM 32 GB.

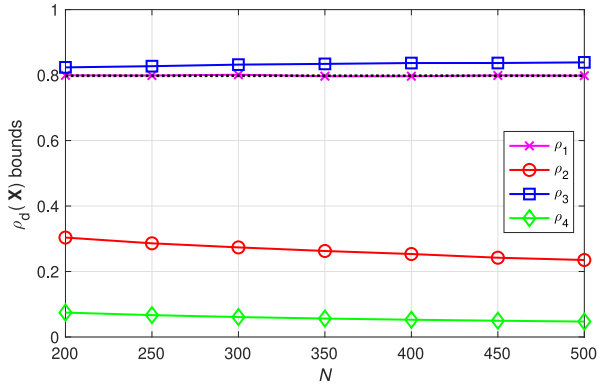

Fig. 7. Bounds $\{\rho_i\}_{i=1}^4$ versus number of measurements N (outlier-free data).

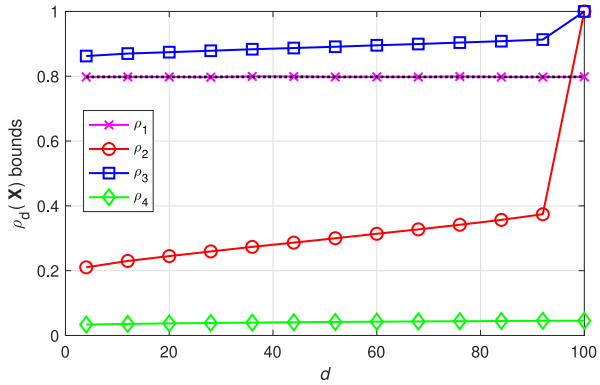

Fig. 8. Bounds $\{\rho_i\}_{i=1}^4$ versus reduced-rank d (outlier-free data).

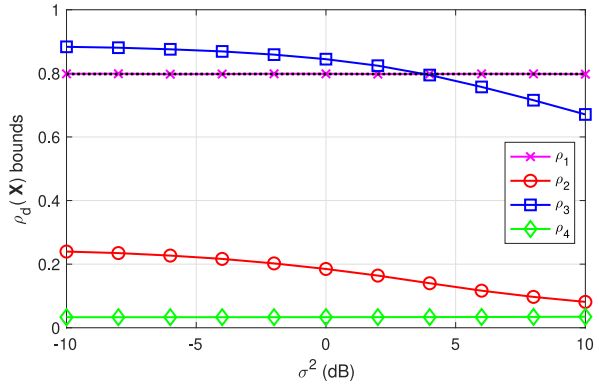

Fig. 9. Bounds $\{\rho_i\}_{i=1}^4$ versus noise variance σ^2 (outlier-free data).

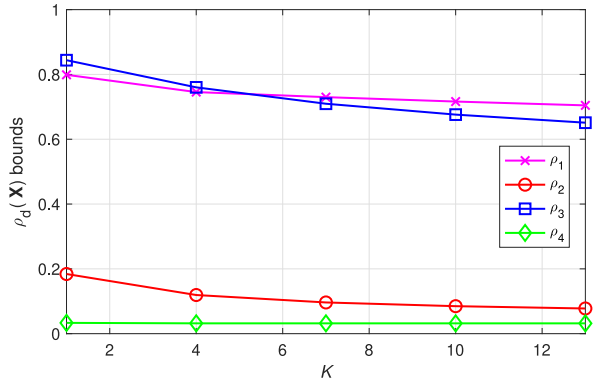

Fig. 10. Bounds $\{\rho_i\}_{i=1}^4$ versus number of components K and $d = K + 1$ (outlier-free data).

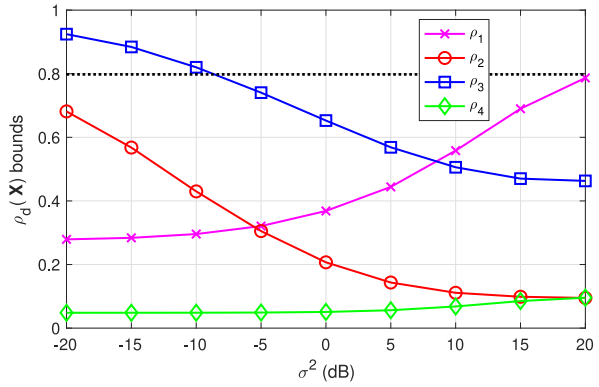

Fig. 11. Bounds $\{\rho_i\}_{i=1}^4$ versus noise variance σ^2 (outlier-corrupted data).

$\mathbf{X} = \mathbf{X}_{\text{nom}} + \mathbf{N} \in \mathbb{R}^{(D=100)\times N}$ such that: (i) the columns of $\mathbf{X}_{\text{nom}}$ are drawn from $\mathcal{N}(\mathbf{0}, \mathbf{U}\text{diag}([100, 10, \ldots, 10^{3-D}])\mathbf{U}^\top)$ for some $\mathbf{U} \in \mathbb{S}_{D,D}$ and (ii) the entries $\mathbf{N}$ are independently drawn from $\mathcal{N}(0, \sigma^2 = 1)$. We compute the top $(K = 1)$ PC of $\mathbf{X}$ by means of RR L1-PCA.

In Fig. 7, we set $d = 3$ and plot the average values of $\{\rho_i\}_{i=1}^4$, calculated over 300 realizations, versus $N = 200, 250, \ldots, 500$. We notice that ρ_3 is marginally above ρ_1 which is almost equal to 0.8 and very close to the benchmark $\sqrt{2/\pi}$.

In Fig. 8, we plot the same bounds, this time for fixed $N = 1000$ and varying $d = 2, 4, \ldots, D$. We notice that even for d as low as 2, $\rho_d(\mathbf{X})$ is lower bounded by 0.85. Moreover, this lower bound does not increase significantly as d increases from 2 to $D - 1 = 99$.

Next, in Fig. 9, we plot the bounds for $N = 1000$, $d = 2$, and varying noise variance $\sigma^2 = -10, -8, \ldots, 10dB$. We notice that for low noise variance $\rho_3 > \rho_1$. For σ^2 greater than 4dB, ρ_3 drops below ρ_1 which remains close to 0.8 across the board.

In Fig. 10, we plot the bounds for $N = 1000$, $\sigma^2 = 2$, varying $K = 1, 4, \ldots, 13$, and $d = K + 1$. We notice that for low K, $\rho_{K+1}(\mathbf{X})$ remains above 0.8 and for any K it does not fall below 0.7.

Next, we study the bounds for outlier corrupted data $\mathbf{X} = \mathbf{X}_{\text{nom}} + \mathbf{O} + \mathbf{N} \in \mathbb{R}^{(D=50)\times(N=500)}$ such that: (i) the columns of $\mathbf{X}_{\text{nom}}$ are drawn from $\mathcal{N}(\mathbf{0}, \mathbf{U}\text{diag}([5, 1, \ldots, 5^{2-D}])\mathbf{U}^\top)$ for some $\mathbf{U} \in \mathbb{S}_{D,D}$; (ii) each entry of $\mathbf{N}$ is drawn from $\mathcal{N}(0, \sigma^2)$; (iii) each column of $\mathbf{O}$ is non-zero with probability 5% and drawn from $\mathcal{N}(\mathbf{0}, 400\mathbf{u}_o\mathbf{u}_o^\top)$, for some $\mathbf{u}_o \in \mathbb{S}_{D,1}$ such that $\mathbf{u}_o^\top[\mathbf{U}]_{:,1} = 0.1$. In Fig. 11, we plot the bounds for varying noise variance $\sigma^2 = -20, -15, \ldots, 20dB$. We notice the order of the bounds changes as σ^2 increases and ρ_2 drops below ρ_1 which ascends. The results are in accordance with our discussion in Section IV.B.

Next, we study the bounds for $\xi_d(\mathbf{X})$, presented in (19). We consider the same data model as in the study of Fig. 3 and present the tightness occurrences in Fig. 12. We notice that for $d = 1, 2, 3$, ξ_2 is the tightest bound most of the times. For $d = 4$, ξ_1 is the tightest bound most often. At the same time, we notice that, though much more rarely, ξ_3 can also be the tightest bound in some realizations.

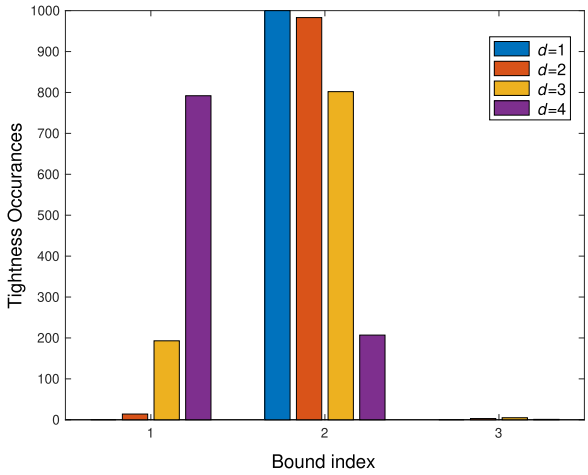

Fig. 12. Bound tightness occurrences for $\xi_d(\mathbf{X})$.

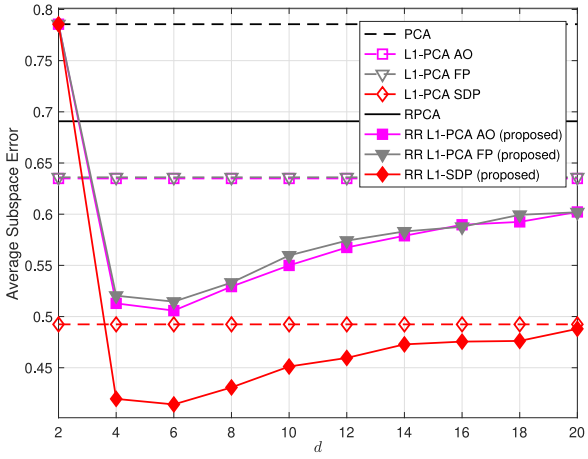

Fig. 13. Average subspace error versus reduced rank d.

C. Subspace Estimation

In this study, we evaluate the subspace estimation accuracy of the proposed framework. We consider $N = 200$ points of dimension $D = 50$ forming measurement matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$. We assume that $\mathbf{X}$ is noisy and outlier corrupted as $\mathbf{X} = \mathbf{X}_{\text{nom}} + \mathbf{O} + \mathbf{N}$ such that: (i) $\mathbf{X}_{\text{nom}} = \mathbf{Q}_{\text{nom}}\mathbf{G}$ is the rank-l matrix, where $\mathbf{Q}_{\text{nom}} \in \mathbb{S}_{D,l}$ is an orthonormal basis that determines the nominal data subspace and $\mathbf{G} \in \mathbb{R}^{l \times N}$ satisfies $\mathbb{E}\{\|\mathbf{G}\|_F^2\} = 4 \times 10^4$; (ii) $\mathbf{O}$ is the outlier matrix that has rank at most 3, has exactly 10 non-zero columns that draw values from a subspace near-orthogonal to the span of $\mathbf{Q}_{\text{nom}}$ and satisfies $\mathbb{E}\{\|\mathbf{O}\|_F^2\} = 75 \times 10^3$; (iii) $\mathbf{N}$ draws entries from $\mathcal{N}(0, \sigma^2 = 10^2)$. We consider $l = 2$ and compute the $K = l = 2$ PCs of $\mathbf{X}$, $\mathbf{Q} \in \mathbb{S}_{D,K}$. Then, we measure the subspace-estimation error (SE) $\frac{1}{2K}\|\mathbf{Q}_{\text{nom}}\mathbf{Q}_{\text{nom}}^\top - \mathbf{Q}\mathbf{Q}^\top\|_F^2$. In Fig. 13, we illustrate the average SE over 100 data/outlier/noise realizations, for standard PCA, RPCA, L1-PCA AO, L1-PCA FP, L1-PCA SDP, and the proposed reduced-rank (RR) counterparts of all the L1-PCA methods. For the RR methods, we show how the performance varies across d.

We observe that standard PCA is significantly affected by the corruption, achieving SE near 0.8. RPCA exhibits robustness, attaining SE below 0.7. The L1-PCA methods outperform RPCA. L1-PCA FP and AO attain SE 0.64, while L1-PCA SDP attains SE 0.49. The proposed reduced-rank implementations of all L1-PCA methods allow for significant SE reduction compared to their full-rank counterparts for all tested values of $d > 2$ and with a minimum attained, across all methods, for $d = 6$ (i.e., for d just higher than the sum of the ranks of the nominal data and the outliers). Specifically, for $d = 6$, RR L1-PCA FP and RR L1-PCA AO attain SE just below 0.51, while RR L1-PCA SDP attains SE 0.41. The above results corroborate that the rank reduction prior to L1-PCA processing can have beneficial denoising impact that improves the subspace estimation accuracy, while still allowing for L1-PCA to resist outliers.

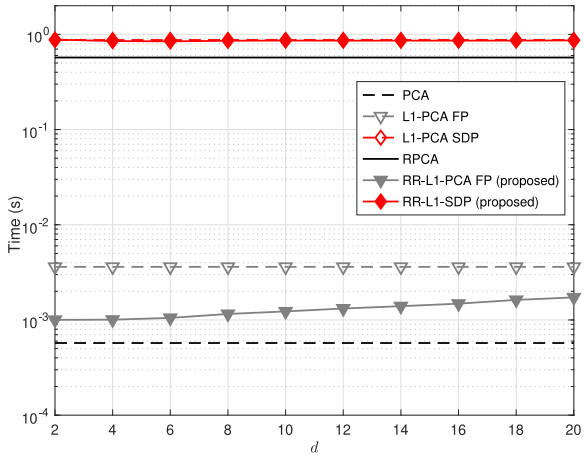

Fig. 14. Average computation time versus reduced rank d for $K = 2$.

Next, we wish to study the computational benefits of the proposed rank reduction. As discussed in Section I, for the exact L1-PCA algorithm, rank reduction can drop the cost exponentially, from $\mathcal{O}(N^{(r-1)K+1})$ to $\mathcal{O}(N^{(d-1)K+1})$. In Fig. 14, we conduct run-time (in seconds) studies for PCA, RPCA, and approximate L1-PCA methods. We notice that the computation time of the proposed RR methods is upper bounded by that of their full-rank counterparts. Certainly, for these approximate/low-cost algorithms, the computational savings of rank reduction are not as pronounced as for the exact solver. Still, for the low-cost L1-PCA FP [25], the rank reduction can offer a $5\times$ speedup. L1-PCA SDP (full rank and RR version) and RPCA, are more than $100\times$ slower than all L1-PCA counterparts. For the same setup, we decrease the nominal data rank to $l = 1$ and set $K = 1$. In Fig. 15, we plot the average run-time. Again, we observe the computational benefit of rank-reduction.

Next, for $K = 1$, we set $\mathbb{E}\{\|\mathbf{G}\|_F^2\} = 2 \times 10^4$ and compute the PC of the corrupted data matrix, for various values of noise deviation σ. In Fig. 16 we plot the average SE performance of PCA, RPCA, L1-PCA AO (which algorithmically coincides with L1-PCA FP for $K = 1$), L1-PCA SDP, as well as the proposed RR counterparts of the L1-PCA methods, for $d = 5$.

We notice that PCA is significantly affected by the outliers (drawn from a subspace almost orthogonal to that of the nominal data), attaining SE close to 0.94, for all values of σ. RPCA is certainly more robust than PCA against outliers, especially for low σ. As σ increases, the performance of RPCA tends to that of PCA (as the low-rank-plus-sparse assumption that it is based on seizes to hold). On the other hand, all L1-PCA methods (full and reduced rank) exhibit sturdy robustness, close to 0.3 for σ close to 0. As σ increases, the proposed RR methods exhibit higher noise immunity and clearly outperform their full-rank

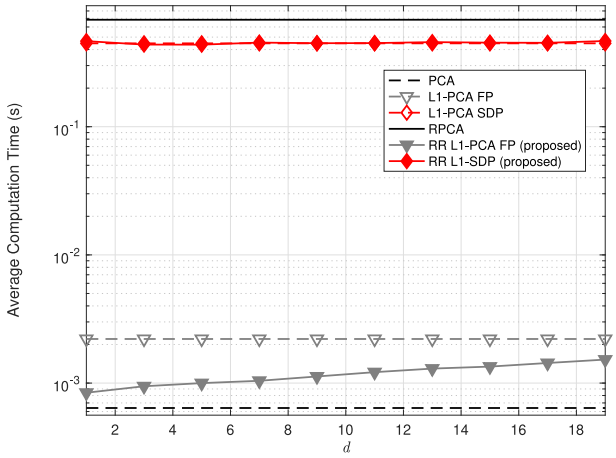

Fig. 15. Average computation time versus reduced rank d for $K = 1$.

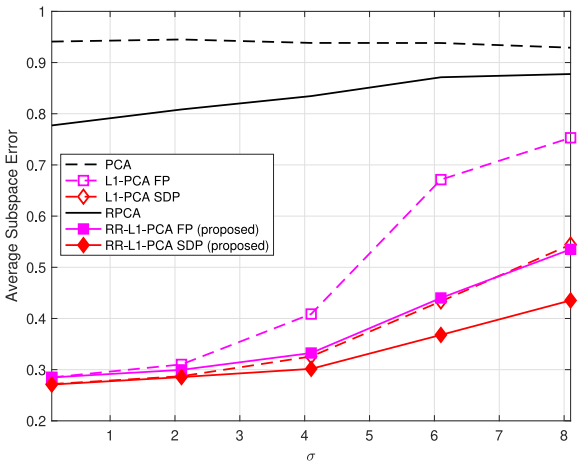

Fig. 16. Average subspace error versus noise deviation σ.

counterparts. Interestingly, RR L1-PCA FP attains quite similar performance to full-rank L1-PCA SDP, for significantly lower computational cost.

Next, we conduct a subspace estimation study for larger data matrices. Specifically, we consider data matrix $\mathbf{X} = \mathbf{X}_{\text{nom}} + \mathbf{O} + \mathbf{N} \in \mathbb{R}^{(D=50)\times(N=1000)}$ such that: (i) the columns of $\mathbf{X}_{\text{nom}}$ are drawn from $\mathcal{N}(\mathbf{0}, \mathbf{U}\text{diag}(\lambda^2, \lambda, \ldots, \lambda^{3-D})\mathbf{U}^\top)$ for some $\mathbf{U} \in \mathbb{S}_{D,D}$ and $\lambda = 12$; (ii) each entry of noise matrix $\mathbf{N}$ is drawn from $\mathcal{N}(0, \sigma^2 = 10)$; (iii) each column of outlier matrix $\mathbf{O}$ is non-zero with probability p and drawn from $\mathcal{N}(\mathbf{0}, \mathbf{U}_\text{o}\text{diag}(\lambda_\text{o}^2, \lambda_\text{o}, \ldots, \lambda_\text{o}^{3-D})\mathbf{U}_\text{o}^\top)$, with $\lambda_\text{o} = 25$ some $\mathbf{U}_\text{o} \in \mathbb{S}_{D,D}$ such that $[\mathbf{U}_\text{o}^\top \mathbf{U}]_{1,1} = 0.1$.

We estimate $\mathbf{U}_{:,1}$ by the $K = 1$ PC of $\mathbf{X}$, computed by means of PCA, RPCA [9], RPCA with Outlier Pursuit (RPCA OP) [49], Coherent Pursuit PCA [51], and L1-PCA AO (full and reduced rank). In Fig. 17, we plot the average subspace error versus corruption probability p. We observe that Coherent PCA exhibits sturdy resistance against the corruption. Remarkable robustness is also attained by L1-PCA AO and RR L1-PCA AO. RPCA and RPCA OP are somewhat more robust than PCA, which is critically affected by the corruption. In Fig. 18, we plot the computation time for the same methods. Quite interestingly, RR L1-PCA is the fastest robust PC calculator, with computation time slightly higher than standard PCA.

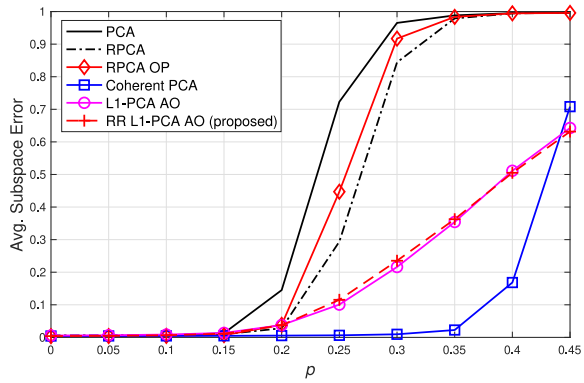

Fig. 17. Average subspace error versus corruption probability p ($\sigma^2 = 10$).

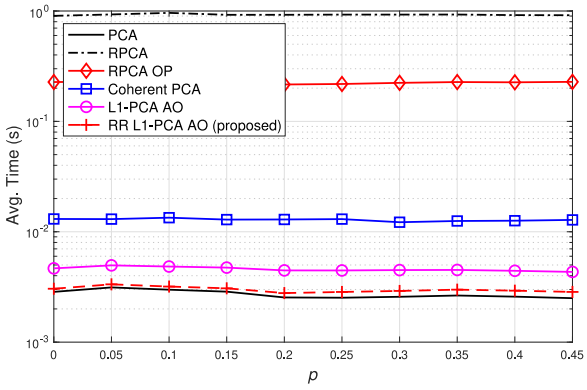

Fig. 18. Average computation time versus corruption probability p ($\sigma^2 = 10$).

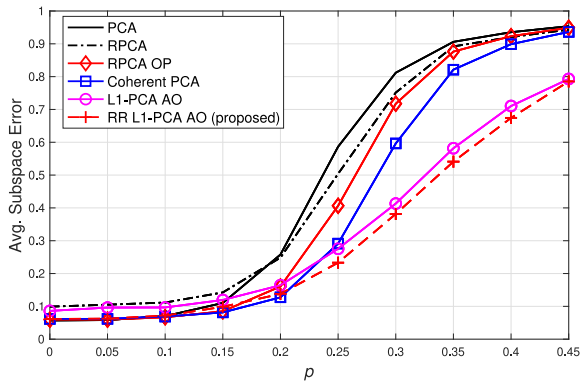

Fig. 19. Average subspace error versus corruption probability p ($\sigma^2 = 100$).

In Figures 19 and 20 we repeat the studies of Figures 17 and 18, respectively, this time for $\sigma^2 = 100$. We notice that the proposed method attains the lowest estimation error (in particular for $p > 0.25$) at the lowest computation time (similar to standard PCA).

D. Image Reconstruction

In this experiment, we consider original image $\mathbf{X}_0 \in \{0, 1, 2, \ldots, 255\}^{H\times W}$ (cameraman image), with $H = W = 256$, as shown in Fig. 21(a). We collect $N = 10$ copies of $\mathbf{X}_0$, $\{\mathbf{X}_n\}_{n=1}^N$, such that each copy is corrupted by zero-mean Additive White Gaussian Noise (AWGN) with variance $\sigma^2 = 100$ (e.g., see an instance of a noisy copy in Fig. 21(b)). Next, in six

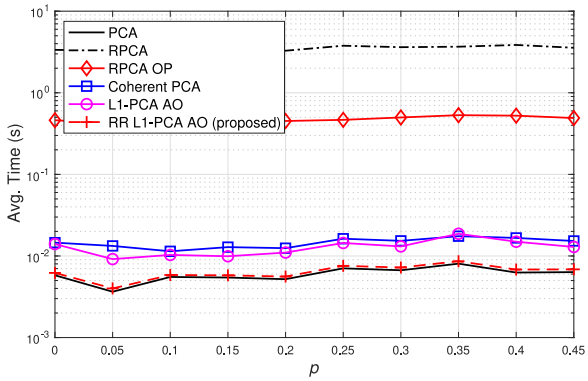

Fig. 20. Average computation time versus corruption probability p ($\sigma^2 = 100$).

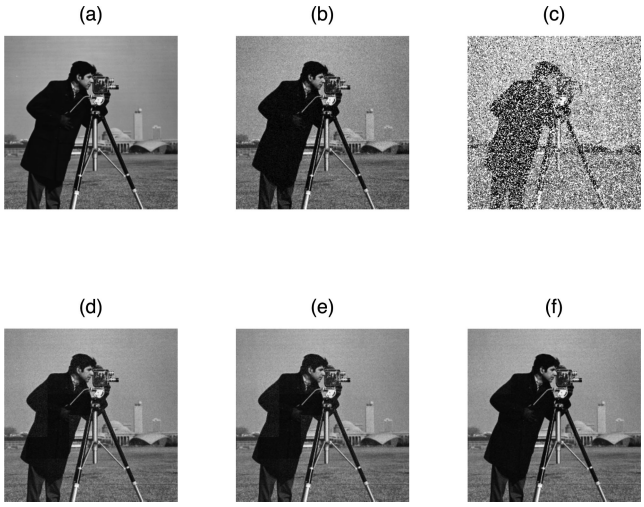

Fig. 21. (a) Original image; (b) AWGN corrupted image; (c) AWGN and S&P corrupted image; (d) Reconstructed image by RR L1-PCA ($d = 3$; RR with SVD); (e) Reconstructed image by RR L1-PCA ($d = 3$; RR with NNMF); (f) Reconstructed image by RR L1-PCA ($d = 3$; RR with SVD after centralization).

arbitrary copies, 40% of the pixels are overwritten by "salt and pepper" (S&P) noise; that is, each corrupted pixel is set white, with probability $p = 0.9$, or black, with probability $1 - p$. An instance of a S&P corrupted copy is shown in Fig. 21(c)). We consider that $\mathbf{X}_0$ is unavailable and we wish to reconstruct it as $\hat{\mathbf{X}}_0$, using $\{\mathbf{X}_n\}_{n=1}^N$. To that end, we first divide each corrupted copy $\mathbf{X}_n$ into $Z = \frac{HW}{hw}$ patches, $\{\mathbf{X}_n^{(i)}\}_{i=1}^Z$, such that the size of each patch is $h \times w$ (for $h = w = 32$). Then, we vectorize each patch as $\mathbf{a}_n^{(i)} := \text{vec}(\mathbf{X}_n^{(i)})$ by vertical concatenation of its columns. Next, for the i-th patch, $i \in [Z]$, we define

$$\mathbf{A}^{(i)} := [\mathbf{a}_1^{(i)}, \mathbf{a}_2^{(i)}, \ldots, \mathbf{a}_N^{(i)}] \tag{21}$$

and, similar to (7), its L1-PC

$$\mathbf{q}^{(i)^*} = \mathbf{A}^{(i)}\mathbf{b}^{(i)^*}\|\mathbf{A}^{(i)}\mathbf{b}^{(i)^*}\|_2^{-1}, \tag{22}$$

where $\mathbf{b}^{(i)^*} = \text{argmax}_{\mathbf{b}\in\{\pm 1\}^N} \|\mathbf{A}^{(i)}\mathbf{b}\|_2$. In practice, adding AWGN to $\mathbf{A}^{(i)}$ may result to some very small negative entries, which we truncate setting them to 0 so that $\mathbf{A}^{(i)} \in \{0, 1, 2, \ldots, 255\}^{hw\times N}$. Since the entries of $\mathbf{A}^{(i)}$ are non-negative, $\mathbf{b}^{(i)^*} = \mathbf{1}_N$ and (22) equals[7]

$$\mathbf{q}^{(i)} = \mathbf{A}^{(i)}\mathbf{1}_N\|\mathbf{A}^{(i)}\mathbf{1}_N\|_2^{-1}. \tag{23}$$

Using the L1-PC, we form the L1-reliability of the n-th copy of the i-th patch as [36], [54]:

$$r_n^{(i)} = \|\mathbf{a}_n^{(i)} - \mathbf{q}^{(i)}\mathbf{q}^{i^\top}\mathbf{a}_n^{(i)}\|_1^{-2}. \tag{24}$$

Subsequently, we define the normalized reliability weights

$$w_n^{(i)} := r_n^{(i)}\|\mathbf{r}^{(i)}\|_1^{-1}, \tag{25}$$

where $\mathbf{r}^{(i)} := [r_1^{(i)}, r_2^{(i)}, \ldots, r_N^{(i)}]^\top$. Then, we reconstruct the i-th vectorized patch as

$$\hat{\mathbf{a}}^{(i)} = \sum_{n=1}^N w_n^{(i)}\mathbf{a}_n^{(i)} = \mathbf{A}^{(i)}\mathbf{w}^{(i)}, \tag{26}$$

where $\mathbf{w}^{(i)} := [w_1^{(i)}, w_2^{(i)}, \ldots, w_N^{(i)}]^\top$. Finally, we reshape $\hat{\mathbf{a}}^{(i)}$ into patch $\hat{\mathbf{A}}^{(i)} \in \mathbb{R}^{h\times w}$ and appropriately arrange/assemble the reconstructed patches to form $\hat{\mathbf{X}}_0 \in \mathbb{R}^{H\times W}$.

Next, we repeat the above study, using instead low-rank approximations of $\mathbf{A}^{(i)}$ and the corresponding RR L1-PCs.

a) SVD Approximation of Patches: According to the proposed method, the rank-d approximation of $\mathbf{A}^{(i)}$ is defined as $\mathbf{A}_d^{(i)} := \mathbf{U}_{1\to d}\mathbf{U}_{1\to d}^\top\mathbf{A}^{(i)} = \mathbf{U}\mathbf{D}_d\mathbf{V}^\top$. Then, we use the L1-PC of $\mathbf{A}_d^{(i)}$ instead of $\mathbf{q}^{(i)}$ to reconstruct the original image as presented above.

b) NNMF Approximation of Patches: Next, we use Non-Negative Matrix Factorization (NNMF) [55] to obtain the rank-d approximation matrix $\mathbf{A}_d^{(i)}$. Specifically, we used the multiplicative method of [55] to find non-negative matrices $\mathbf{R} \in \mathbb{R}_+^{hw\times d}$ and $\mathbf{S} \in \mathbb{R}_+^{d\times N}$ such as $\|\mathbf{A}^{(i)} - \mathbf{R}\mathbf{S}\|_F^2$ is minimized. Then we approximated $\mathbf{A}^{(i)}$ by $\mathbf{A}_d^{(i)} = \mathbf{R}\mathbf{S}$ and used the L1-PC of $\mathbf{A}_d^{(i)}$ instead of $\mathbf{q}^{(i)}$ to reconstruct $\mathbf{X}_0$.

c) SVD approximation of centralized patches: Next, we centralize the patches in $\mathbf{A}^{(i)}$, as

$$\mathbf{T}^{(i)} := [\mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \ldots, \mathbf{t}_N^{(i)}]_{hw\times N} = \mathbf{A}^{(i)}\mathbf{C}_N, \tag{27}$$

where $\mathbf{C}_N = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$. Then, we rank-$d$ approximate $\mathbf{T}^{(i)}$ by SVD, obtaining $\mathbf{T}_d^{(i)}$ and we find the L1-PC of $\mathbf{T}_d^{(i)}$, $\mathbf{q}_c^{(i)}$. For the n-th copy of the i-th patch we compute its L1-reliability as $\|\mathbf{t}_n^{(i)} - \mathbf{q}_c^{(i)}\mathbf{q}_c^{i^\top}\mathbf{t}_n^{(i)}\|_1^{-2}$ instead of $r_n^{(i)}$ in (24).

In Fig. 21 (d), (e), and (f) we present the reconstructed image, by means of the three RR methods above, for $d = 3$. We notice that centralization before L1-PC-based reconstruction offers a better result. For a more detailed comparison between the three approaches, we plot in Fig. 22 the Mean Squared Estimation Error (MSE), evaluated as the average of $\frac{1}{HW}\|\mathbf{X}_0 - \hat{\mathbf{X}}_0\|_2^2$ over 1000 independent noise/corruption realizations, versus the reduced rank $d = 1, 3, \ldots, 9$. We notice that for $d = 1$, RR L1-PCA coincides with standard PCA. Also, we notice that the MSE decreases as d increases, for all approaches. Finally, we observe that the third approach (RR by SVD on centralized data) clearly outperforms the other two, for every value of d.

[7]For any matrix $\mathbf{X} \in \mathbb{R}_+^{D\times N}$ with non-negative entries, it holds that $\mathbf{1}_N \in \text{argmax}_{\mathbf{b}\in\{\pm 1\}^N} \|\mathbf{X}\mathbf{b}\|_2$ and, accordingly, $\mathbf{X}\mathbf{1}_N\|\mathbf{X}\mathbf{1}_N\|_2^{-1} \in \text{argmax}_{\mathbf{q}\in\mathbb{S}_{D,1}} \|\mathbf{X}^\top\mathbf{q}\|_1$ [36].

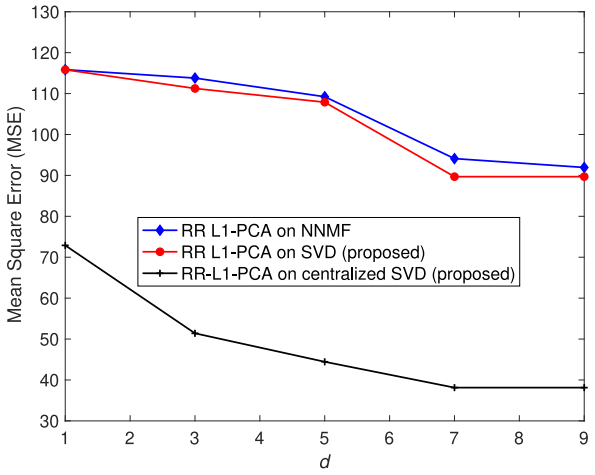

Fig. 22. MSE versus d attained by RR L1-PCA based on SVD, NNMF, and centralized SVD.

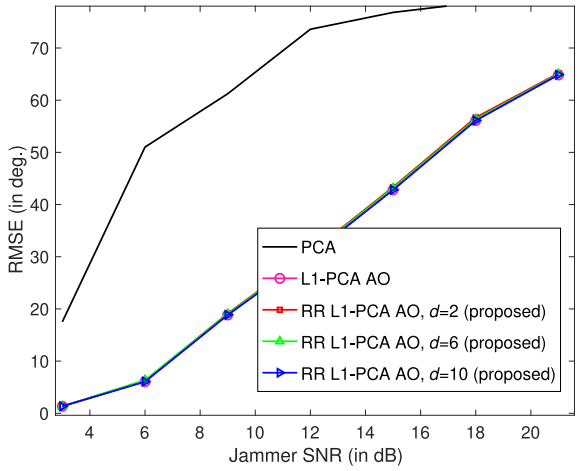

Fig. 24. RMSE versus jammer SNR based on PCA, L1-PCA, and RR L1-PCA ($d = 2, 6, 10$) when $D = 7$ and $N = 17$.

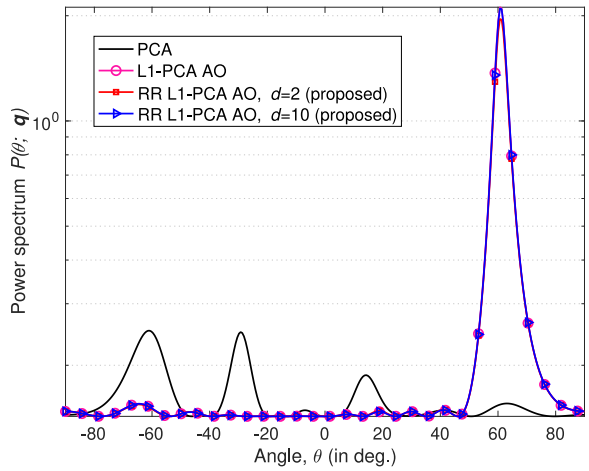

Fig. 23. MUSIC power spectrum based on PCA, L1-PCA, and RR L1-PCA for $d = 2$ and 10.

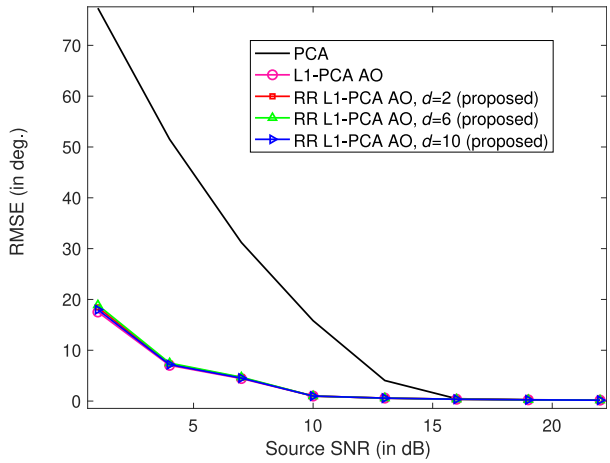

Fig. 25. RMSE versus source SNR based on PCA, L1-PCA, and RR L1-PCA ($d = 2, 6, 10$) when $\text{SNR}_\text{j} = 7$dB, $D = 7$, and $N = 17$.

E. Direction-of-Arrival Estimation

In this section we consider a uniform linear antenna array of $D = 7$ elements that collects $N = 17$ snapshots of Binary Phase Shift Keying (BPSK) signals from angle $\phi = 60°$,

$$\mathbf{x}_n = A b_n \mathbf{s}_\phi + \mathbf{n}_n, \tag{28}$$

$n = 1, 2, \ldots, 17$. In (28), A denotes the received signal amplitude, $\mathbf{s}_\phi$ is the array response vector for angle ϕ, $b_n \in \{\pm 1\}$ is the BPSK symbol, and $\mathbf{n}_n \sim \mathcal{CN}(\mathbf{0}_7, \sigma^2 \mathbf{I}_7)$ is AWGN. We assume that the Signal-to-Noise-Ratio (SNR) is $\text{SNR} = 10 \log_{10} \frac{A^2}{\sigma^2} = 3$dB. Next we consider that three arbitrarily selected observations are corrupted by three jammers with $\text{SNR}_{\text{j}1} = \text{SNR}_{\text{j}2} = \text{SNR}_{\text{j}3} = 7$dB and angles of arrival $\phi_{j1} = -60°$, $\phi_{j2} = -30°$, and $\phi_{j3} = 15°$, respectively. The resulting corrupted collection of snapshots is denoted by $\mathbf{X}^{\text{CRPT}} \in \mathbb{C}^{7 \times 17}$. Next, we transform $\mathbf{X}^{\text{CRPT}}$ by $\text{Re}\{\cdot\}, \text{Im}\{\cdot\}$ part concatenation to its real-domain version $\tilde{\mathbf{X}}^{\text{CRPT}} = [\text{Re}(\mathbf{X}^{\text{CRPT}})^\top, \text{Im}(\mathbf{X}^{\text{CRPT}})^\top]^\top \in \mathbb{R}^{14 \times 17}$ similar to [38]. Then, we calculate and plot in Fig. 23 , the MUSIC spectrum

$$P(\theta; \mathbf{q}) = \frac{1}{D - (\mathbf{q}^{*\top} \tilde{\mathbf{s}}_\theta)^2}, \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right), \tag{29}$$

where $\tilde{\mathbf{s}}_\theta = [\text{Re}(\mathbf{s}_\theta)^\top, \text{Im}(\mathbf{s}_\theta)^\top]^\top$ and $\mathbf{q}$ is set to $\mathbf{q}^* \in \mathbb{S}_{14,1}$, the top principal component of $\tilde{\mathbf{X}}^{\text{CRPT}}$ produced by L1-PCA or PCA. Also, we consider the rank-d approximation matrix $\tilde{\mathbf{X}}_d^{\text{CRPT}} = \mathbf{U}\mathbf{D}_d\mathbf{V}^\top$ and find the DoA by means the peak of $P(\theta; \mathbf{q}_d^*)$, where $\mathbf{q}_d^* \in \mathbb{S}_{14,1}$ is the L1-PC of $\tilde{\mathbf{X}}_d^{\text{CRPT}}$ (proposed method). The true angle of arrival ϕ is estimated by the peak of $P(\theta; \mathbf{q})$. In Fig. 23, we notice that the spectrum based on PCA is more affected by jammers rather than that obtained by L1-PCA and RR L1-PCA. Also, we notice that the performances of L1-PCA and the proposed RR L1-PCA are very similar. Next, we repeat the study $M = 10^3$ times, for M independent realizations. We denote by $\hat{\phi}(m)$ the estimated angle obtained by finding the peak of $P(\theta; \mathbf{q})$ in the mth experiment. In Fig. 24, we plot the Root-Mean-Squared-Error $\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\phi - \hat{\phi}(m))^2}$ as a function of the jammer SNR for different values of d. Interestingly, we notice that the performance of L1-PCA and RR L1-PCA is almost identical (for every value of d) and significantly superior to that of standard PCA by means of SVD. In Fig. 25, we plot the RMSE versus source SNR when jammer $\text{SNR} = 7$ dB, $D = 7$, and $N = 17$. According to this figure,

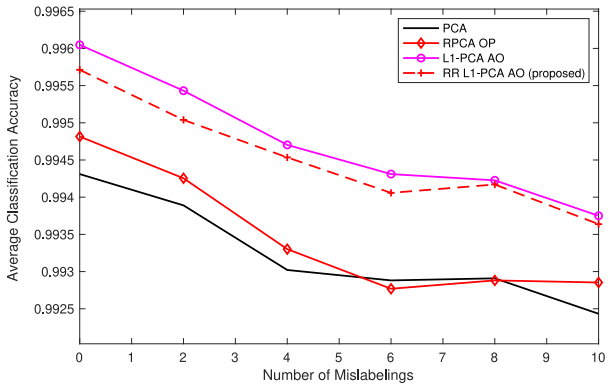

Fig. 26. Average classification accuracy versus number of mislabeled points in the training batch of each class.

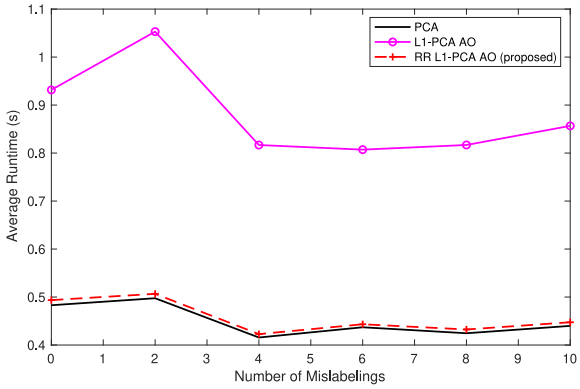

Fig. 27. Average runtime versus number of mislabeled points in the training batch of each class.

by increasing the source SNR, the RMSE will decrease. Also, similar to Fig. 24, increasing d does not affect RMSE in Fig. 25.

F. MNIST Data Classification

In this experiment, we consider the MNIST dataset [56] and particularly the classes of digits '0' and '1'. From each class, we consider 1000 training points and 892 testing points. Each point is a vectorized image of length 784. The training data are organized in matrix $\mathbf{X} \in \mathbb{R}^{784\times 2000}$. Each training point is additively corrupted with noise from $\mathcal{N}(0, 25)$. Also, we consider that N_{mis} of the training points (equal number from each class) are in fact a mislabeled image of digit '4'.

In our classification experiment, we work as follows. First, we compute a rank-2 principal basis $\mathbf{Q}$ by decomposing the corrupted training dataset $\mathbf{X}$. Then, we project all training/testing data on $\mathbf{Q}$. We use the projected data to train and test a standard nearest neighbor classifier (1-NN) and measure its accuracy. We repeat this study over 100 realizations of training/testing data. In Fig. 26 we plot average classification accuracy vs. N_{mis}, when $\mathbf{Q}$ is computed by means of PCA, RPCA OP, L1-PCA AO, and the proposed RR L1-PCA AO ($d = 5$). L1-PCA and RR L1-PCA attain the highest performance.

In Fig. 27 we also plot the average subspace computation time vs. the number of mislabeled points in each training class. The runtime of RPCA OP was above 40 sec. across the board and therefore it is not even plotted. We notice that the proposed RR L1-PCA AO is almost two times faster than full-rank L1-PCA AO.

VI. Conclusion

We presented RR L1-PCA: a new framework for reduced-rank L1-PCA. The presented method approximates the K L1-PCs of $\mathbf{X}$ by the L1-PCs of its SVD-obtained rank-d approximation ($d \leq r$) $\mathbf{X}_d$. Importantly, we also derived formal performance bounds for RR L1-PCA. Our experimental studies corroborate that the proposed framework effectively combines the denoising capabilities and low computation cost of standard PCA with the outlier-resistance of L1-PCA.

A. Proof of Lemma 1

It has been shown that the solution to (2) is given by (5). The argument of (3) can be rewritten as

$$\|\mathbf{XB}\|_* = \mathrm{Tr}(\sqrt{\mathbf{B}^\top\mathbf{X}^\top\mathbf{XB}}) = \mathrm{Tr}(\sqrt{\mathbf{B}^\top\mathbf{V}\mathbf{D}^\top\mathbf{D}\mathbf{V}^\top\mathbf{B}}) = \|\mathbf{D}\mathbf{V}^\top\mathbf{B}\|_*. \tag{30}$$

By (30), $\mathbf{B}^*$ maximizes both $\|\mathbf{XB}\|_*$ and $\|\mathbf{D}\mathbf{V}^\top\mathbf{B}\|_*$. Thus,

$$\bar{\mathbf{Q}}^* = \mathbf{\Phi}(\mathbf{D}\mathbf{V}^\top\mathbf{B}^*) := \underset{\mathbf{Q}\in\mathbb{S}_{D,K}}{\mathrm{argmax}} \ \|\mathbf{V}\mathbf{D}^\top\mathbf{Q}\|_1 = \mathbf{\Omega}\mathbf{W}^\top, \tag{31}$$

where $\mathbf{D}\mathbf{V}^\top\mathbf{B}^* \overset{\text{svd}}{=} \mathbf{\Omega}\mathbf{\Lambda}\mathbf{W}^\top$. Moreover, if we consider $\mathbf{X}\mathbf{B}^* = \mathbf{U}\mathbf{D}\mathbf{V}^\top\mathbf{B}^* \overset{\text{svd}}{=} (\mathbf{U}\mathbf{\Omega})\mathbf{\Lambda}\mathbf{W}^\top$, then

$$\mathbf{Q}^* = \mathbf{\Phi}(\mathbf{X}\mathbf{B}^*) = \mathbf{\Phi}(\mathbf{U}\mathbf{D}\mathbf{V}^\top\mathbf{B}^*) = \mathbf{U}\mathbf{\Omega}\mathbf{W}^\top = \mathbf{U}\mathbf{\Phi}(\mathbf{D}\mathbf{V}^\top\mathbf{B}^*) = \mathbf{U}\bar{\mathbf{Q}}^*. \tag{32}$$

By (32), we conclude the proof of Lemma 1.

B. Proof of Theorem 2

Let $\tilde{\mathbf{Q}} \in \mathbb{S}_{D,D-K}$ be orthonormal matrix spanning the orthogonal complement of $\mathrm{span}(\mathbf{Q})$.[8] Then, by the Pythagorean theorem, it holds that $\|\mathbf{X}^\top\|_F^2 = \|\mathbf{X}^\top\mathbf{Q}\|_F^2 + \|\mathbf{X}^\top\tilde{\mathbf{Q}}\|_F^2$. Next, we consider the following Lemmas 2, 3, and 4, presented in [52] and [53].

Lemma 2: ([52]) Consider $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{B} \in \mathbb{R}^{p\times m}$. For any $z > 0$, it holds that

$$\sum_{i=1}^{\min\{n,m,p\}} \sigma_i^z(\mathbf{AB}) \leq \sum_{i=1}^{\min\{n,m,p\}} \sigma_i^z(\mathbf{A})\sigma_i^z(\mathbf{B}). \tag{33}$$

Lemma 3: ([53]) For every conformable matrix pair $\mathbf{A}$, $\mathbf{B}$, and any matrix norm $\|\cdot\|$, it holds that $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ and $\|\mathbf{A}\| - \|\mathbf{B}\| \leq \|\mathbf{A}-\mathbf{B}\|$.

Lemma 4: ([53]) For every matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank $r \leq \min(m,n)$ and every vector $\mathbf{q} \in \mathbb{R}^N$, we have $\|\mathbf{A}\|_F \leq \|\mathbf{A}\|_* \leq \|\mathbf{A}\|_1 \leq \sqrt{mn}\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_* \leq \sqrt{r}\|\mathbf{A}\|_F$, and $\|\mathbf{q}\|_2 \leq \|\mathbf{q}\|_1 \leq \sqrt{N}\|\mathbf{q}\|_2$.

In addition, we derive the following Lemma 5, the proof of which is provided below.

[8]For any $\mathbf{A} \in \mathbb{R}^{n\times p}$, $\mathrm{span}(\mathbf{A}) := \{\bar{\mathbf{x}} \in \mathbb{R}^n : \bar{\mathbf{x}} = \mathbf{A}\bar{\mathbf{y}} \text{ for some } \bar{\mathbf{y}} \in \mathbb{R}^p\}$. Subspace $\mathcal{X} \subseteq \mathbb{R}^n$ is the "orthogonal complement" of subspace $\mathcal{Y} \subseteq \mathbb{R}^n$, if and only if (i) $\bar{\mathbf{x}}^\top\bar{\mathbf{y}} = 0\ \forall \bar{\mathbf{x}} \in \mathcal{X}$ and $\bar{\mathbf{y}} \in \mathcal{Y}$ and (ii) $\mathcal{X} \cup \mathcal{Y} = \mathbb{R}^n$.

Lemma 5: Consider any $\mathbf{Q} \in \mathbb{S}_{D,K}$ for $K \le d \le r \le \min\{D, N\}$. It holds that

$$\lambda_{h+1\to r}(\mathbf{X}) \le \|\mathbf{X}^\top \mathbf{Q}\|_F \le \lambda_{1\to K}(\mathbf{X}), \tag{34}$$

where $h = \min\{r, D-K\}$.

Similar to (34), for the reduced-rank approximation matrix $\mathbf{X}_d$, it holds that

$$\lambda_{h+1\to d}(\mathbf{X}) \le \|\mathbf{X}_d^\top \mathbf{Q}\|_F \le \lambda_{1\to K}(\mathbf{X}). \tag{35}$$

In the above, if $h+1 > d$, then $\lambda_{h+1\to d}(\mathbf{X}) = 0$. By Lemma 3, Lemma 4, and (35), for every matrix $\mathbf{Q}$ with orthonormal columns, it holds

$$\begin{aligned}
&\|\mathbf{X}^\top \mathbf{Q}\|_1 - \|\mathbf{X}_d^\top \mathbf{Q}\|_1 = \|(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top \mathbf{Q}\|_1 - \|(\mathbf{U}\mathbf{D}_d\mathbf{V}^\top)^\top \mathbf{Q}\|_1 \\
&\le \|(\mathbf{U}(\mathbf{D}-\mathbf{D}_d)\mathbf{V}^\top)^\top \mathbf{Q}\|_1 \\
&\le \sqrt{NK}\|(\mathbf{U}(\mathbf{D}-\mathbf{D}_d)\mathbf{V}^\top)^\top \mathbf{Q}\|_F \\
&\overset{(35)}{\le} \sqrt{NK}\lambda_{\min(d+1,r)\to\min(K+d,r)}(\mathbf{X}).
\end{aligned} \tag{36}$$

Also, by Lemma 4 and (35), for every matrix $\mathbf{Q}$ with orthonormal columns, we can write

$$\|\mathbf{X}^\top \mathbf{Q}\|_1 \le \sqrt{NK}\|\mathbf{X}^\top \mathbf{Q}\|_F \le \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) \tag{37}$$

and

$$\|\mathbf{X}_d^\top \mathbf{Q}\|_1 \ge \|\mathbf{X}_d^\top \mathbf{Q}\|_F \ge \lambda_{h+1\to d}(\mathbf{X}). \tag{38}$$

Then, by (37) and (38), it holds

$$\|\mathbf{X}^\top \mathbf{Q}\|_1 - \|\mathbf{X}_d^\top \mathbf{Q}\|_1 \le \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) - \lambda_{h+1\to d}(\mathbf{X}). \tag{39}$$

As a result, by (36) and (39), we have

$$\begin{aligned}
&\|\mathbf{X}^\top \mathbf{Q}\|_1 - \|\mathbf{X}_d^\top \mathbf{Q}\|_1 \qquad (40)\\
&\le \min\Big\{\sqrt{NK}\lambda_{\min(d+1,r)\to\min(K+d,r)}(\mathbf{X}), \\
&\qquad \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) - \lambda_{h+1\to d}(\mathbf{X})\Big\}. \qquad (41)
\end{aligned}$$

On the other hand, by Lemma 4 and Lemma 5, it holds

$$\|\mathbf{X}^\top \mathbf{Q}\|_1 \ge \|\mathbf{X}^\top \mathbf{Q}\|_F \ge \lambda_{h+1\to r}(\mathbf{X}). \tag{42}$$

Also, for the reduced-rank matrix $\mathbf{X}_d$, we have

$$\|\mathbf{X}_d^\top \mathbf{Q}\|_1 \le \sqrt{NK}\|\mathbf{X}_d^\top \mathbf{Q}\|_F \le \sqrt{NK}\lambda_{1\to K}(\mathbf{X}). \tag{43}$$

Therefore, by (42) and (43), it holds that

$$\lambda_{h+1\to r}(\mathbf{X}) - \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) \le \|\mathbf{X}^\top \mathbf{Q}\|_1 - \|\mathbf{X}_d^\top \mathbf{Q}\|_1. \tag{44}$$

Finally, by (41) and (44), we conclude Theorem 2.

C. Proof of Lemma 5

By (33), it holds

$$\begin{aligned}
\|\mathbf{X}^\top \mathbf{Q}\|_F &= \lambda_{1\to K}(\mathbf{X}^\top \mathbf{Q}) \le \sqrt{\sum_{i=1}^{K} \sigma_i^2(\mathbf{X}^\top)\sigma_i^2(\mathbf{Q})} \\
&= \lambda_{1\to K}(\mathbf{X}^\top) = \lambda_{1\to K}(\mathbf{X}).
\end{aligned} \tag{45}$$

Then, similar to (45), we have

$$\begin{aligned}
\|\mathbf{X}^\top \tilde{\mathbf{Q}}\|_F &= \lambda_{1\to h}(\mathbf{X}^\top \tilde{\mathbf{Q}}) \le \sqrt{\sum_{i=1}^{h} \sigma_i^2(\mathbf{X}^\top)\sigma_i^2(\tilde{\mathbf{Q}})} \\
&= \lambda_{1\to h}(\mathbf{X}^\top) = \lambda_{1\to h}(\mathbf{X}).
\end{aligned} \tag{46}$$

Therefore, by the Pythagorean theorem and (46), it holds that

$$\|\mathbf{X}^\top \mathbf{Q}\|_F \ge \lambda_{h+1\to r}(\mathbf{X}). \tag{47}$$

In the above, we note that if $h+1 > r$, then $\lambda_{h+1\to r}(\mathbf{X}) = 0$. Finally, combining (45) with (47) yields (34), which concludes the proof of Lemma 5.

D. Proof of Theorem 3

For any $i, j \in [r]$, with $i < j$, we define $\mathbf{X}_{i\to j} := \mathbf{U}_{i\to j}\mathbf{U}_{i\to j}^\top \mathbf{X}$. Accordingly, $\mathbf{X} = \mathbf{X}_{1\to r}$ and $\mathbf{X}_d = \mathbf{X}_{1\to d}$. In addition, we define $\mathbf{B}_{i\to j}^* := \operatorname{argmax}_{\mathbf{B}\in\{\pm 1\}^{N\times K}} \|\mathbf{X}_{i\to j}\mathbf{B}\|_*$, $\mathbf{Q}_{i\to j}^* = \mathbf{\Phi}(\mathbf{X}_{i\to j}\mathbf{B}_{i\to j}^*)$, and $f_{i\to j}^* := \|\mathbf{X}_{i\to j}^\top \mathbf{Q}_{i\to j}^*\|_1 = \|\mathbf{X}_{i\to j}\mathbf{B}_{i\to j}^*\|_*$. We commence our proof by noting that, for all $i, j \in [r]$ with $i < j$, it holds

$$\begin{aligned}
\operatorname{span}(\mathbf{Q}_{i\to j}^*) &= \operatorname{span}(\mathbf{X}_{i\to j}\mathbf{B}_{i\to j}^*) \subseteq \operatorname{span}(\mathbf{X}_{i\to j}) \\
&\subseteq \operatorname{span}(\mathbf{U}_{i\to j}).
\end{aligned} \tag{48}$$

Since $\mathbf{U}_{1\to j}^\top \mathbf{U}_{(i+1)\to j} = \mathbf{0}_{j\times(j-i)}$, (48) implies that

$$\mathbf{X}_{(i+1)\to j}^\top \mathbf{Q}_{1\to i}^* = \mathbf{0}_{N\times K}. \tag{49}$$

Therefore, $\mathbf{X}_{1\to j}^\top \mathbf{Q}_{1\to i}^* = (\mathbf{X}_{1\to i} + \mathbf{X}_{(i+1)\to j})^\top \mathbf{Q}_{1\to i}^* = \mathbf{X}_{1\to i}^\top \mathbf{Q}_{1\to i}^* + \mathbf{X}_{(i+1)\to j}^\top \mathbf{Q}_{1\to i}^* = \mathbf{X}_{1\to i}^\top \mathbf{Q}_{1\to i}^*$ and

$$\begin{aligned}
f_{1\to i}^* &= \|\mathbf{X}_{1\to i}^\top \mathbf{Q}_{1\to i}^*\|_1 = \|\mathbf{X}_{1\to j}^\top \mathbf{Q}_{1\to i}^*\|_1 \\
&\le \|\mathbf{X}_{1\to j}^\top \mathbf{Q}_{1\to j}^*\|_1 = f_{1\to j}^*.
\end{aligned} \tag{50}$$

By (50), the accuracy ratio of (14) can be re-written as

$$\rho_d(\mathbf{X}) := \frac{f_{1\to d}^*}{f_{1\to r}^*} \le 1. \tag{51}$$

Accordingly, by Lemma 3 and Lemma 4, we find that

$$\begin{aligned}
&f_{1\to j}^* \\
&= \|\mathbf{X}_{1\to j}\mathbf{B}_{1\to j}^*\|_* \\
&\le \|\mathbf{X}_{1\to i}\mathbf{B}_{1\to j}^*\|_* + \|\mathbf{X}_{(i+1)\to j}\mathbf{B}_{1\to j}^*\|_* \qquad (52)\\
&\le f_{1\to i}^* + \sqrt{\operatorname{rank}(\mathbf{X}_{(i+1)\to j}\mathbf{B}_{1\to j}^*)}\,\|\mathbf{X}_{(i+1)\to j}\mathbf{B}_{1\to j}^*\|_F \\
&\le f_{1\to i}^* + \sqrt{\min\{(j-i), K\}}\,\|\mathbf{X}_{(i+1)\to j}\mathbf{B}_{1\to j}^*\|_F \\
&\le f_{1\to i}^* + \sqrt{\min\{(j-i), K\}NK}\max_{\mathbf{b}\in\{\pm\frac{1}{\sqrt{N}}\}^N}\|\mathbf{X}_{(i+1)\to j}\mathbf{b}\|_2 \\
&\le f_{1\to i}^* + \sqrt{\min\{(j-i), K\}NK}\max_{\mathbf{b}\in\mathbb{S}_{N,1}}\|\mathbf{X}_{(i+1)\to j}\mathbf{b}\|_2 \\
&= f_{1\to i}^* + \sqrt{\min\{(j-i), K\}NK}\,\sigma_{i+1}(\mathbf{X}). \qquad (53)
\end{aligned}$$

Inequalities (50) and (53) are summarized in the following Lemma 6.

Lemma 6: For any $i, j \in [r]$ with $i \le j$, it holds

$$\begin{aligned}
f_{1\to i}^* &\le f_{1\to j}^* \\
&\le f_{1\to i}^* + \sqrt{\min\{(j-i), K\}NK}\,\sigma_{i+1}(\mathbf{X}).
\end{aligned} \tag{54}$$

Substituting i and j in the right-hand inequality of (54) by d and r, respectively, we obtain

$$f_{1\to r}^* \le f_{1\to d}^* + \sqrt{\min\{(r-d), K\}NK}\,\sigma_{d+1}(\mathbf{X}). \tag{55}$$

Dividing both sides of (55) by $f^*_{1\to d}$, we find

$$\frac{f^*_{1\to r}}{f^*_{1\to d}} \leq 1 + \frac{\sqrt{\min\{(r-d),K\}NK}\sigma_{d+1}(\mathbf{X})}{f^*_{1\to d}}. \quad (56)$$

Therefore,

$$\rho_d(\mathbf{X}) = \frac{f^*_{1\to d}}{f^*_{1\to r}} \geq \frac{1}{1+\frac{\sqrt{\min\{(r-d),K\}NK}\sigma_{d+1}(\mathbf{X})}{f^*_{1\to d}}}. \quad (57)$$

For obtaining lower and upper bounds for (18) and (19), respectively, we present the following Lemma 7, a proof for which is offered below.

Lemma 7: For every $j \in [r]$, it holds

$$\sqrt{K}\|\mathbf{X}_{1\to j}\|_F \leq f^*_{1\to j} \leq \sqrt{\min\{j,K\}NK}\,\sigma_1(\mathbf{X}). \quad (58)$$

By (37), we find

$$f^*_{1\to r} \leq \sqrt{NK}\lambda_{1\to K}(\mathbf{X}). \quad (59)$$

By (58) and $r > K$, it holds

$$f^*_{1\to r} \leq K\sqrt{N}\sigma_1(\mathbf{X}). \quad (60)$$

It is clear that $\sqrt{NK}\lambda_{1\to K}(\mathbf{X}) = \sqrt{NK}\sqrt{\sum_{i=1}^K \sigma_i^2(\mathbf{X})} \leq \sqrt{NK}\sqrt{K\sigma_1^2(\mathbf{X})} = K\sqrt{N}\sigma_1(\mathbf{X})$. Interestingly, by (54), it holds

$$\begin{aligned} f^*_{1\to d} \geq f^*_{1\to 1} &= \max_{\mathbf{B}\in\{\pm 1\}^{N\times K}} \sigma_1(\mathbf{X})\|\mathbf{u}_1\mathbf{v}_1^\top\mathbf{B}\|_* \\ &= \sigma_1(\mathbf{X}) \max_{\mathbf{B}\in\{\pm 1\}^{N\times K}} \|\mathbf{v}_1^\top\mathbf{B}\|_2 \\ &= \sqrt{K}\sigma_1(\mathbf{X}) \max_{\mathbf{b}\in\{\pm 1\}^{N}} |\mathbf{v}_1^\top\mathbf{b}| \\ &= \sqrt{K}\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1. \end{aligned} \quad (61)$$

Then, by (59) and (61), we find

$$\rho_d(\mathbf{X}) = \frac{f^*_{1\to d}}{f^*_{1\to r}} \geq \frac{\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1}{\sqrt{N}\lambda_{1\to K}(\mathbf{X})}. \quad (62)$$

By Lemma 7, it holds

$$f^*_{1\to d} \geq \sqrt{K}\|\mathbf{X}_{1\to d}\|_F. \quad (63)$$

As a result,

$$\begin{aligned} \rho_d(\mathbf{X}) = \frac{f^*_{1\to d}}{f^*_{1\to r}} &\geq \frac{1}{1+\frac{\sqrt{\min\{(r-d),K\}NK}\sigma_{d+1}(\mathbf{X})}{\sqrt{K}\|\mathbf{X}_{1\to d}\|_F}} \\ &= \frac{1}{1+\frac{\sqrt{\min\{(r-d),K\}N}\sigma_{d+1}(\mathbf{X})}{\lambda_{1\to d}(\mathbf{X})}} \end{aligned} \quad (64)$$

and

$$\begin{aligned} \rho_d(\mathbf{X}) = \frac{f^*_{1\to d}}{f^*_{1\to r}} &\geq \frac{1}{1+\frac{\sqrt{\min\{(r-d),K\}NK}\sigma_{d+1}(\mathbf{X})}{\sqrt{K}\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1}} \\ &= \frac{1}{1+\frac{\sqrt{\min\{(r-d),K\}N}\sigma_{d+1}(\mathbf{X})}{\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1}}. \end{aligned} \quad (65)$$

Finally, by (59) and (63), it holds

$$\begin{aligned} \rho_d(\mathbf{X}) = \frac{f^*_{1\to d}}{f^*_{1\to r}} &\geq \frac{f^*_{1\to d}}{\sqrt{NK}\lambda_{1\to K}(\mathbf{X})} \\ &\geq \frac{\sqrt{K}\|\mathbf{X}_{1\to d}\|_F}{\sqrt{NK}\lambda_{1\to K}(\mathbf{X})} = \frac{\lambda_{1\to d}(\mathbf{X})}{\sqrt{N}\lambda_{1\to K}(\mathbf{X})}, \end{aligned} \quad (66)$$

which concludes our proof of (18). That is, so far we showed that, expectedly, $\rho_d(\mathbf{X})$ is upper bounded by 1. Also, we presented a lower bound for $\rho_d(\mathbf{X})$ that can be calculated by means of SVD of $\mathbf{X}$ with cost $\mathcal{O}(ND\min\{N,D\})$.

To prove (19), we first note that

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = \mathbf{U}\mathbf{D}_d\mathbf{V}^\top + \mathbf{U}(\mathbf{D}-\mathbf{D}_d)\mathbf{V}^\top. \quad (67)$$

By (50),

$$\|\mathbf{X}_d^\top\mathbf{Q}_d^*\|_1 = \|\mathbf{X}^\top\mathbf{Q}_d^*\|_1 \leq \|\mathbf{X}^\top\mathbf{Q}^*\|_1. \quad (68)$$

Therefore, we find

$$0 \leq \|\mathbf{X}^\top\mathbf{Q}^*\|_1 - \|\mathbf{X}^\top\mathbf{Q}_d^*\|_1. \quad (69)$$

By (54), we find

$$\begin{aligned} &\|\mathbf{X}^\top\mathbf{Q}^*\|_1 - \|\mathbf{X}^\top\mathbf{Q}_d^*\|_1 \\ &\leq \sqrt{\min\{(r-d),K\}NK}\,\sigma_{d+1}(\mathbf{X}). \end{aligned} \quad (70)$$

Also, by (59),

$$\|\mathbf{X}^\top\mathbf{Q}^*\|_1 \leq \sqrt{NK}\lambda_{1\to K}(\mathbf{X}). \quad (71)$$

Therefore, by (71) and (61), we find

$$\begin{aligned} &\|\mathbf{X}^\top\mathbf{Q}^*\|_1 - \|\mathbf{X}^\top\mathbf{Q}_d^*\|_1 \\ &\leq \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) - \sqrt{K}\,\sigma_1(\mathbf{X})\|\mathbf{v}_1\|_1. \end{aligned} \quad (72)$$

In addition, by (50) and (59), we have

$$\begin{aligned} \sqrt{K}\|\mathbf{U}\mathbf{D}_d\mathbf{V}^\top\|_F &\leq \|(\mathbf{U}\mathbf{D}_d\mathbf{V}^\top)^\top\mathbf{Q}_d^*\|_1 \\ &= \|(\mathbf{U}\mathbf{D}\mathbf{V}^\top)^\top\mathbf{Q}_d^*\|_1 \\ &\leq \sqrt{NK}\lambda_{1\to K}(\mathbf{X}). \end{aligned} \quad (73)$$

As a result, by (71), (68), and (73),

$$\begin{aligned} &\|\mathbf{X}^\top\mathbf{Q}^*\|_1 - \|\mathbf{X}^\top\mathbf{Q}_d^*\|_1 \\ &\leq \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) - \sqrt{K}\,\|\mathbf{U}\mathbf{D}_d\mathbf{V}^\top\|_F \\ &= \sqrt{NK}\lambda_{1\to K}(\mathbf{X}) - \sqrt{K}\lambda_{1\to d}(\mathbf{X}). \end{aligned} \quad (74)$$

Finally, by (69), (70), (72), and (74), we derive (19).

E. Proof of Lemma 7

For the right side of inequality (58), we find

$$\begin{aligned} f^*_{1\to j} &\leq \sqrt{\operatorname{rank}(\mathbf{X}_{1\to j}\mathbf{B}^*_{1\to j})}\|\mathbf{X}_{1\to j}\mathbf{B}^*_{1\to j}\|_F \\ &\leq \sqrt{\min\{j,K\}K}\max_{\mathbf{b}\in\{\pm 1\}^N}\|\mathbf{X}_{1\to j}\mathbf{b}\|_2 \\ &\leq \sqrt{\min\{j,K\}KN}\max_{\mathbf{b}\in\mathbb{S}_{N,1}}\|\mathbf{X}_{1\to j}\mathbf{b}\|_2 \\ &= \sqrt{\min\{j,K\}KN}\,\sigma_1(\mathbf{X}). \end{aligned} \quad (75)$$

By Lemma 4 and the definition of $f^*_{1\to j}$,

$$\begin{aligned} f^*_{1\to j} &\geq \|\mathbf{X}_{1\to j}\mathbf{b}^*_{1\to j}\mathbf{1}_K^\top\|_* \\ &\geq \|\mathbf{X}_{1\to j}\mathbf{b}^*_{1\to j}\mathbf{1}_K^\top\|_F = \sqrt{K}\|\mathbf{X}_{1\to j}\mathbf{b}^*_{1\to j}\|_2, \end{aligned} \quad (76)$$

where $\mathbf{b}^*_{1\to j} = \operatorname{argmax}_{\mathbf{b}\in\{\pm 1\}^N}\|\mathbf{X}_{1\to j}\mathbf{b}\|_2$. Then, we observe that, as it was originally shown in [30],

$$[\mathbf{b}^*_{1\to j}]_n[\mathbf{X}_{1\to j}]_{:,n}^\top\mathbf{X}_{1\to j}\mathbf{b}^*_{1\to j} \geq \|[\mathbf{X}_{1\to j}]_{:,n}\|_2^2 \quad \forall n, \quad (77)$$

which implies that

$$\|\mathbf{X}_{1\to j}\mathbf{b}^*_{1\to j}\|_2 \geq \|\mathbf{X}_{1\to j}\|_F. \quad (78)$$

Next, we prove (77) below for completeness purposes. We start the to-be-contradicted assumption

$$[\mathbf{b}^*_{1\to j}]_n[\mathbf{X}_{1\to j}]^\top_{:,n}\,\mathbf{X}_{1\to j}\mathbf{b}^*_{1\to j} < \|[\mathbf{X}_{1\to j}]_{:,n}\|_2^2, \tag{79}$$

for some $n=\{1,2,\ldots,N\}$. Then, we define $\mathbf{b}' := \mathbf{b}^*_{1\to j} - 2[\mathbf{b}^*_{1\to j}]_n\mathbf{e}_{n,N} \in \{\pm 1\}^N$, where $\mathbf{e}_{n,N} := [\mathbf{I}_N]_{:,n}$; that is, we set $b'_m = [\mathbf{b}^*_{1\to j}]_m$ for every $m \neq n$ and $b'_n = -[\mathbf{b}^*_{1\to j}]_n$. Then, we find

$$\begin{aligned}
&\|\mathbf{X}_{1\to j}\mathbf{b}'\|_2^2 - \|\mathbf{X}_{1\to j}\mathbf{b}^*\|_2^2\\
&= b'_n[\mathbf{X}_{1\to j}]^\top_{:,n}\,\mathbf{X}_{1\to j}\mathbf{b}' - [\mathbf{b}^*_{1\to j}]_n[\mathbf{X}_{1\to j}]^\top_{:,n}\,\mathbf{X}_{1\to j}\mathbf{b}^*_{1\to j}\\
&= b'_n[\mathbf{X}_{1\to j}]^\top_{:,n}\sum_{m\neq n}[\mathbf{X}_{1\to j}]_{:,m}b'_m + \|[\mathbf{X}_{1\to j}]_{:,n}\|_2^2\\
&- \left([\mathbf{b}^*_{1\to j}]_n[\mathbf{X}_{1\to j}]^\top_{:,n}\sum_{m\neq n}[\mathbf{X}_{1\to j}]_{:,m}[\mathbf{b}^*_{1\to j}]_m + \|[\mathbf{X}_{1\to j}]_{:,n}\|_2^2\right)\\
&= -2[\mathbf{b}^*_{1\to j}]_n[\mathbf{X}_{1\to j}]^\top_{:,n}\sum_{m\neq n}[\mathbf{X}_{1\to j}]_{:,m}[\mathbf{b}^*_{1\to j}]_m\\
&= -2([\mathbf{b}^*_{1\to j}]_n[\mathbf{X}_{1\to j}]^\top_{:,n}\,\mathbf{X}_{1\to j}\mathbf{b}^*_{1\to j} - \|[\mathbf{X}_{1\to j}]_{:,n}\|_2^2)\\
&\overset{(79)}{>} 0,
\end{aligned} \tag{80}$$

which cannot hold true since by the definition of $\mathbf{b}^*_{1\to j}$ and $\|\mathbf{X}_{1\to j}\mathbf{b}^*_{1\to j}\|_2^2 \geq \|\mathbf{X}_{1\to j}\mathbf{b}\|_2^2\ \forall \mathbf{b}\in\{\pm 1\}^N$. Therefore, (79) is contradicted and (77) holds true. By (76) and (78), the left side of inequality of Lemma 7 holds true as well. Finally, by (75), (76), and (78), we conclude the proof of Lemma 7.

ACKNOWLEDGMENT

The authors would like to thank the Department of Electrical Engineering at the University at Buffalo for hosting H. Kamrani as a visiting scholar.

REFERENCES

[1] P. P. Markopoulos, G. N. Karystinos, and D. A. Pados, "Optimal algorithms for L1-subspace signal processing," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5046–5058, Oct. 2014.

[2] P. P. Markopoulos, G. N. Karystinos, and D. A. Pados, "Some options for L1-subspace signal processing," in *Proc. 10th Int. Symp. Wireless Commun. Syst.*, Ilmenau, Germany, Aug. 2013, pp. 622–626.

[3] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag.*, vol. 2, no. 11, pp. 559–572, 1901.

[4] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 1986.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., New York, NY, USA: Wiley, 2001.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[7] L. L. Kasun, Y. Yang, G.-B. Huang, and Z. Zhang, "Dimension reduction with extreme learning machine," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3906–3918, Aug. 2016.

[8] Z. Allen-Zhu and Y. Li, "LazySVD: Even faster SVD decomposition yet without agonizing pain," in *Proc. Adv. Neural Info. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 974–982.

[9] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, May 2011, Art. no. 11.

[10] V. Barnett and T. Lewis, *Outliers in Statistical Data*. New York, NY, USA: Wiley, 1994.

[11] J. W. Tukey, "The future of data analysis," *Ann. Math. Stat.*, vol. 33, no. 1, pp. 1–67, 1962.

[12] G. Lerman and T. Maunu, "An overview of robust subspace recovery," *Proc. IEEE*, vol. 106, no. 8, pp. 1380–1410, Aug. 2018.

[13] Q. Ke and T. Kanade, "Robust subspace computation using L1 norm," Computer Science Dept., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-03-172, Aug. 2003.

[14] Q. Ke and T. Kanade, "Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 739–746.

[15] N. Tsagkarakis, P. P. Markopoulos, and D. A. Pados, "On the L1-norm approximation of a matrix by another of lower rank," in *Proc. IEEE Int. Conf. Machine Learn. App.*, Anaheim CA, USA, Dec. 2016, pp. 768–773.

[16] A. Eriksson and A. v. d. Hengel, "Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L1 norm," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 771–778.

[17] L. Yu, M. Zhang, and C. Ding, "An efficient algorithm for L1-norm principal component analysis," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 1377–1380.

[18] I. Barrodale, "L1 approximation and the analysis of data," *J. Royal Stat. Soc. Appl. Stat.*, vol. 17, pp. 51–57, 1968.

[19] I. Barrodale and F. D. K. Roberts, "An improved algorithm for discrete L1 linear approximation," *SIAM J. Numer. Anal.*, vol. 10, no. 5, pp. 839–848, Oct. 1973.

[20] J. P. Brooks and J. H. Dulá, "The L1-norm best-fit hyperplane problem," *Appl. Math. Lett.*, vol. 26, no. 1, pp. 51–55, Jan. 2013.

[21] J. P. Brooks, J. H. Dulá, and E. L. Boone, "A pure L1-norm principal component analysis," *J. Comput. Stat. Data Anal.*, vol. 61, pp. 83–98, May 2013.

[22] Y. W. Park and D. Klabjan, "Three iteratively reweighted least squares algorithms for L1-norm principal component analysis," *Knowledge Info. Syst.*, vol. 54, no. 3, pp. 541–565, 2018.

[23] Y. W. Park, "Optimization for l1-norm error fitting via data aggregation," *INFORMS J. Comput.* 2020. Accessed: Jun. 15, 2020. [Online]. Available: https://pubsonline.informs.org/doi/pdf/10.1287/ijoc.2019.0908

[24] Y. W. Park and D. Klabjan, "Iteratively reweighted least squares algorithms for L1-norm principal component analysis," in *Proc. IEEE Int. Conf. Data Mining*, Barcelona, Spain, Dec. 2016, pp. 430–438.

[25] N. Kwak and J. Oh, "Feature extraction for one-class classification problems: Enhancements to biased discriminant analysis," *Pattern Recognit.*, vol. 42, pp. 17–26, Jan. 2009.

[26] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[27] M. McCoy and J. A. Tropp, "Two proposals for robust PCA using semidefinite programming," *Electron. J. Stat.*, vol. 5, pp. 1123–1160, Jun. 2011.

[28] F. Nie, H. Huang, C. Ding, D. Luo, and H. Wang, "Robust principal component analysis with non-greedy L1-norm maximization," in *Proc. Int. Joint Conf. Artif. Intell.*, Barcelona, Spain, Jul. 2011, pp. 1433–1438.

[29] S. Kundu, P. P. Markopoulos, and D. A. Pados, "Fast computation of the L1-principal component of real-valued data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 8028–8032.

[30] P. P. Markopoulos, S. Kundu, S. Chamadia, and D. A. Pados, "Efficient L1-norm Principal-Component Analysis via bit flipping," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4252–4264, Aug. 2017.

[31] N. Tsagkrakis, P. P. Markopoulos, G. Sklivanitis, and D. A. Pados, "L1-norm principal-component analysis of complex data," *IEEE Trans. Signal Process.*, vol. 66, no. 12, pp. 3256–3267, Jun. 2018

[32] C. Ding, D. Zhou, X. He, and H. Zha, "R_1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization," in *Proc. Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, 2006, pp. 281–288.

[33] D. Meng, Q. Zhao, and Z. Xu, "Improve robustness of sparse PCA by L1-norm maximization," *Pattern Recognit.*, vol. 45, no. 1, pp. 487–497, Jan. 2012.

[34] H. Wang, "Block principal component analysis with L1-norm for image analysis," *Pattern Recognit. Lett.*, vol. 33, pp. 537–542, Apr. 2012.

[35] P. P. Markopoulos, "Reduced-rank filtering on L1-norm subspaces," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop.*, Rio de Janeiro, Brazil, Jul. 2016, pp. 1–5.

[36] P. P. Markopoulos, S. Kundu, and D. A. Pados, "L1-fusion: Robust linear-time image recovery from few severely corrupted copies," in *Proc. IEEE Int. Conf. Image Process.*, Quebec City, Canada, Sep. 2015, pp. 1225–1229.

[37] M. Johnson and A. Savakis, "Fast L1-eigenfaces for robust face recognition," in *Proc. IEEE Western New York Image Signal Process. Workshop. (WNYISPW)*, Rochester, NY, USA, Nov. 2014, pp. 1–5.

[38] P. P. Markopoulos, N. Tsagkarakis, D. A. Pados, and G. N. Karystinos, "Realified L1-PCA for direction-of-arrival estimation: Theory and algorithms," *EURASIP J. Adv. Signal Process.*, vol. 30, no. 1, pp. 1–16, Jun. 2019, doi: 10.1186/s13634-019-0625-5.

[39] Y. Liu and D. A. Pados, "Compressed-sensed-domain L1-PCA video surveillance," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 351–363, Mar. 2016.

[40] P. P. Markopoulos, M. Dhanaraj, and A. Savakis, "Adaptive L1-norm principal-component analysis with online outlier rejection," *IEEE J. Select. Topics Signal Process.*, vol. 12, no. 6, pp. 1131–1143, Dec. 2018.

[41] M. Dhanaraj, D. G. Chachlakis, and P. P. Markopoulos, "Incremental complex L1-PCA for direction-of-arrival estimation," in *Proc. IEEE Western NY Image Signal Process. Workshop*, Rochester, NY, USA, Oct. 2018, pp. 1–5.

[42] D. G. Chachlakis, A. Prater-Bennette, and P. P. Markopoulos, "L1-norm Tucker tensor decomposition," *IEEE Access*, vol. 7, pp. 178454–178465, Nov. 2019.

[43] D. G. Chachlakis, M. Dhanaraj, A. Prater-Bennette, and P. P. Markopoulos, "Options for multimodal classification based on L1-Tucker decomposition," in *Proc. SPIE DCS, Big Data: Learn., Analytics, Appl.*, Baltimore, MD, USA, May 2019, pp. 109890O:1–109890O:13.

[44] P. P. Markopoulos, D. G. Chachlakis, and A. Prater-Bennette, "L1-norm higher-order singular-value decomposition," in *Proc. IEEE Global Conf. Signal Info. Process.*, Anaheim CA, Nov. 2018, pp. 1353–1357.

[45] D. G. Chachlakis, M. Dhanaraj, A. Prater-Bennette, and P. P. Markopoulos, "Dynamic L1-norm Tucker tensor decomposition," *IEEE J. Selected Topics Signal Process.*, Aug. 2020, doi: 10.36227/techrxiv.12762392.v1.

[46] K. Tountas, D. A. Pados, and M. J. Medley, "Conformity evaluation and L1-norm principal-component analysis of tensor data," in *Proc. SPIE Def. Commerce Sens.*, Baltimore, MD, USA, Apr. 2019, vol. 109890P, pp. 1–11.

[47] P. P. Markopoulos, D. A. Pados, G. N. Karystinos, and M. Langberg, "L1-norm principal-component analysis in L2-norm-reduced-rank data subspaces," in *Proc. SPIE Def. Commerce Sens.*, Anaheim, CA, USA, Apr. 2017, vol. 1021104, pp. 1–10.

[48] X. Yuan and J. Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *Pacific J. Optim.*, vol. 9, no. 1, pp. 167–180, 2013.

[49] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," *Proc. Adv. Neural Info. Process. Syst.*, Vancouver, Canada, Dec. 2010, pp. 2496–2504.

[50] H. Xu, Academic Webpage, Georgia Institute of Technology. Accessed on: May 23, 2020. [Online]. Available: https://pwp.gatech.edu/huan-xu/publications.

[51] M. Rahmani and G. Atia, "Coherence pursuit: Fast, simple, and robust principal component analysis," *IEEE Trans. Signal Process.*, vol. 65, no. 23, pp. 6260–6275, Dec. 2017.

[52] R. Horn and C. Johnson, *Topics in Matrix Analysis*, 1st ed., New York, NY, USA: Cambridge Univ. Press, 1994.

[53] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.

[54] S. Chamadia and D. A. Pados, "Optimal sparse L1-norm principal-component analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 2686–2690.

[55] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Info. Process. Syst. Conf.* (NeurIPS), Vancouver, Canada, Dec. 2001, pp. 556–562.

[56] Y. LeCun, C. Cortes, and C. J. C. Burges, "THE MNIST DATABASE of handwritten digits," 2020. Accessed: Sep. 27, 2020. [Online]. Available: https://yann.lecun.com/exdb/mnist/