Thresholding based Efficient Outlier Robust PCA

Yeshwanth Cherapanamjeri   T-YECHER@MICROSOFT.COM

Prateek Jain   PRATEEK@MICROSOFT.COM

Praneeth Netrapalli   PRANEETH@MICROSOFT.COM

Microsoft Research, India

Abstract

We consider the problem of outlier robust PCA (**OR-PCA**) where the goal is to recover principal directions despite the presence of outlier data points. That is, given a data matrix M^*, where $(1 - \alpha)$ fraction of the points are noisy samples from a low-dimensional subspace while α fraction of the points can be arbitrary outliers, the goal is to recover the subspace accurately. Existing results for **OR-PCA** have serious drawbacks: while some results are quite weak in the presence of noise, other results have runtime quadratic in dimension, rendering them impractical for large scale applications.

In this work, we provide a novel thresholding based iterative algorithm with per-iteration complexity at most linear in the data size. Moreover, the fraction of outliers, α, that our method can handle is tight up to constants while providing nearly optimal computational complexity for a general noise setting. For the special case where the inliers are obtained from a low-dimensional subspace with additive Gaussian noise, we show that a modification of our thresholding based method leads to significant improvement in recovery error (of the subspace) even in the presence of a large fraction of outliers.

1. Introduction

Principal Component Analysis (PCA) is a critical first step for any typical data exploration/analysis effort and is widely used in a variety of applications. A key reason for the success of PCA is that it can be performed efficiently using Singular Value Decomposition (SVD).

However, due to various practical reasons like measurement error, presence of anomalies etc., a large fraction of data points can be corrupted in a somewhat correlated and even adversarial manner. Unfortunately, SVD is fragile with respect to outliers and can lead to arbitrarily inaccurate principal directions in the presence of even a small number of outliers. So, designing an outlier robust PCA (**OR-PCA**) algorithm is critical for several application domains.

Formally, the setting of **OR-PCA** is as follows: given a data matrix $M^* = D^* + C^* \in \mathbb{R}^{d \times n}$ where $D^* = [x_1, \ldots, x_n]$ corresponds to n clean "inlier" data points and C^* has at most α-fraction of non-zero columns that can corrupt the corresponding clean data points arbitrarily, the goal of **OR-PCA** is to estimate the principal components of D^* accurately, i.e., recover $U^* \in \mathbb{R}^{d \times r}$, the top-$r$ left singular vectors of D^*.

Vanilla SVD does not do the job since the top singular vectors of M^* can be arbitrarily far from U^* if the operator norm of C^* ($\|C^*\|_2$) is large, as can be the case when $\alpha = \Omega(\frac{1}{n})$. Any algorithm trying to solve **OR-PCA** needs to exploit the column sparsity of C^* to obtain a better estimate of U^*. In particular, they need to find $S^* = \text{Supp}(C^*)$—$\text{Supp}(A)$ is the index of non-zero

columns of A—so that U^* can be estimated using top singular directions of $M^*_{\backslash S^*}$, i.e., columns of M^* restricted to complement set of S^*.

Existing results for **OR-PCA** fall into two categories based on: a) Nuclear norm (Xu et al., 2012a; Zhang et al., 2016), b) iterative PCA (Xu et al., 2013). Nuclear norm based approaches work with exactly same setting as ours, but require $O(nd^2)$ computational time which is prohibitive for typical applications. Iterative PCA based techniques require $O(n^2d)$ computation which in general is significantly higher than our algorithms. Moreover, these results do not recover the exact principal directions even if the inliers are restricted to a low-dimensional subspace and just a constant number of outliers are present.

Our approach is based on solving the following natural optimization problem:

$$\textbf{OR-PCA}: \min_{D,C \in \mathbb{R}^{d \times n}} \|M^* - D - C\|_F^2 \text{ s.t. } rank(D) \leq r, \ |\text{Supp}(C)| \leq \alpha n.$$

Technical Challenges: The main challenges with **OR-PCA** are its non-convexity and combinatorial structure which rules out standard tools from convex optimization. Furthermore, due to the column-sparsity constraint standard SVD based techniques also do not apply. Instead, we propose a simple iterative method that constructs an estimate of outliers C and using that, an estimate of the inliers $M^* - C$. SVD of the estimated inliers is then used to obtain an estimate of the principal directions. Now, a significant challenge is to use these principal directions to re-estimate outliers. There are two different scenarios here:

- *Length of all the outliers is smaller than the smallest singular value of the inliers and hence do not stand out*: In this case however, the principal directions are not much affected by outliers. So, outliers can be recognized by taking a projection on to the orthogonal space to the estimated principal directions. As we get better estimate of principal directions, we can be more aggressive in determining the outliers.

- *Length of at least one of the outliers is larger than the small singular values of inliers*: In this case again, length based thresholding fails since the lengths of the inliers are dominated by the larger singular values. Similarly, the above mentioned thresholding scheme also fails as some of the estimated principal directions will be heavily biased towards those outliers. A key and somewhat surprising algorithmic insight of our work is that: as some of the estimated singular vectors are heavily affected by outliers, projection of such outliers on these spurious singular vectors will be inordinately high. So, in contrast to the above thresholding operator, we can use the length of projection of points along estimated principal directions as well to detect and threshold outliers.

In both the scenarios, we can identify more outliers and repeat the procedure till convergence. A key assumption we make, in order to ensure that inliers are not thresholded, is that most of the inliers have "limited" influence on the principal directions, i.e., the data matrix is incoherent (see Assumption 1). Such an assumption holds for various typical settings, for example when inliers are noisy and uniform samples from a low-dimensional subspace.

Contributions: *The main contribution of our work is to show that, under some regularity conditions, it is indeed possible to solve **OR-PCA** near optimally, in essentially the same time as that taken by vanilla PCA*. In particular, we propose a thresholding based approach that iteratively estimates inliers using *two* different thresholding operators and then use SVD to estimate the principal

directions. We also show that our method recovers U^* nearly optimally and efficiently as long as the fraction of corrupted data points (column-support of C^*) is less than $O(1/r)$ where r is the dimensionality of principal subspace U^*. More concretely, we study the problem in three settings:

1. *Noiseless setting*: $M^* = D^* + C^*$ where the clean data matrix D^* is a rank-r matrix with μ-incoherent right singular vectors (see Assumption 1) and C^* has at most $\alpha \cdot n$ non-zero columns. For this setting, we design a novel **T**hresholding based **O**uliter **R**obust **PCA** algorithm (**TORP**) that recovers U^* up to an error of ϵ in time $O\left(ndr \log \frac{n\|M^*\|_2}{\epsilon}\right)$, if $\alpha \leq \frac{1}{128\mu^2 r}$. Vanilla PCA, on the other hand, can be computed in time $O(nd + \frac{nk^2}{\epsilon^2})$ via input sparsity time algorithms (See, for example, Clarkson and Woodruff (2013)) and in time $O(ndr)$ via orthonormal iteration (Golub and Loan (1996)). Hence, we essentially match the running time of Vanilla PCA except in the regime where r is small where the extra r factor doesn't affect the run time significantly. In contrast, existing results for the same setting require $O(\frac{nd^2}{\epsilon^2})$[1] computation to recover U^*. Note that the number of outliers our results can handle (i.e., $\alpha \leq 1/128\mu^2 r$) is optimal up to constant factors in the sense that if $\alpha > 1/\mu^2 r$, then there exists a matrix M^* which has more than one decomposition satisfying the above conditions.

2. *Arbitrary noise*: In the second setting, $M^* = D^* + C^*$ where the clean data matrix D^* can be written as $L^* + N^*$ and L^*, the rank-r projection of D^* has μ incoherent right singular vectors. C^* on the other hand, again has at most $\alpha \cdot n$ non-zero columns. If $\alpha \leq \frac{1}{128\mu^2 r}$, our proposed algorithm **TORP-N** guarantees recovery of U^* (left singular vectors of L^*) up to $O\left(\sqrt{r}\|N^*\|_F + \epsilon\right)$ error, in $O\left(ndr \log \frac{n\|M^*\|_2}{\sqrt{r}\|N^*\|_F + \epsilon}\right)$ time. Again this is essentially the same as the time taken for Vanilla PCA. In contrast, existing results for this problem get stuck at a significantly larger error of $(\sqrt{n}\|N^*\|_F)$, with a runtime of $O(\frac{nd^2}{\epsilon^2})$, which is slower than ours by a factor of d.

3. *Gaussian noise*: In this setting, we again have $M^* = D^* + C^*$ where the clean data matrix D^* is a sum of low-rank matrix L^* and a Gaussian noise matrix N^*, i.e., each element of N^* is sampled independently and identically (iid) from $\mathcal{N}(0, \sigma^2)$. **TORP-G**, which is our proposed algorithm for this special case, recovers U^* up to an error of $O\left(\sqrt{r \log d}\|N^*\|_2\right)$. This not only improves upon the result we obtained for the arbitrary noise case above, which is $O\left(\sqrt{r}\|N^*\|_F\right)$, it also improves significantly upon the existing results. However, in order to achieve this improvement in error, we require $n > d^2$ and the algorithm has a runtime of $O\left(n^2 d\right)$.

To summarize, our results obtain stronger guarantees in terms of both final error and runtime, while being able to handle a large number of outliers, for three different settings of **OR-PCA**. Moreover, in the first two settings, our run time essentially matches that of standard PCA. Please refer Tables 1 and 2 for comparison of our results with existing results.

Paper Outline: The paper is organized as follows. We will review related work in Section 1.1. We then present a formal definition of the problem in Section 2 and our main results in Section 3. We then present our algorithm for each of the three settings: a) noise-less setting, b) arbitrary data,

---

1. Dependence on ϵ is due to the standard rates of gradient descent when applied to the non-smooth non-strongly convex optimization problem given in (1); however, using more refined RSC-style analysis, ϵ dependency might be improved but we are not aware of such an existing result.

	Run time	**Error** ($\left\lVert (I - UU^\top)L^* \right\rVert_F$)
Xu et al. (2012b)	$O\left(\frac{d^2 n}{\epsilon^2}\right)$	$O\left(\sqrt{n}\left\lVert N^* \right\rVert_F + \epsilon\right)$
TORP-N	$O\left(dnr \log \frac{n\lVert M^* \rVert_2}{\epsilon}\right)$	$O\left(\sqrt{r}\left\lVert N^* \right\rVert_F + \epsilon\right)$

Table 1: Arbitrary noise: comparison of our results and existing results for the arbitrary noise setting. Here, $M^* = D^* + C^*$, where $D^* = L^* + N^*$ is the clean data matrix with L^*, the rank-r projection of D^* having μ-incoherent right singular vectors (Assumption 1) and C^* has $O(\frac{1}{\mu^2 r})$ is a column sparse matrix. The error is measured as the residual of L^* when projected on to the estimated space U. Note that we obtain better error and better runtime compared to existing results. Also, the noiseless case is a special case with $N^* = 0$.

	Run time	**Error**($\left\lVert (I - UU^\top)L^* \right\rVert_F$)
Xu et al. (2012b)	$O\left(\frac{nd^2 r}{\epsilon^2}\right)$	$O\left(\sigma n\sqrt{d} + \epsilon\right)$
TORP-N	$O\left(ndr^2 \log \frac{\lVert M^* \rVert_2}{\epsilon}\right)$	$O\left(\sigma\sqrt{nd} + \epsilon\right)$
TORP-G	$O\left(n^2 dr \log \frac{\lVert M^* \rVert_2}{\epsilon}\right)$	$O\left(\sigma\sqrt{n \log(d)} + \epsilon\right)$

Table 2: Gaussian noise: comparison of existing results with ours for the Gaussian noise setting: $M^* = D^* + C^*$, where inliers $D^* = L^* + N^*$ with L^* being a rank-r matrix with μ-incoherent right singular vectors (Assumption 1) and N^* is a Gaussian matrix with variance σ^2. Specializing **TORP-N** for this setting already gives us faster and better results than existing ones. **TORP-G** further improves the error by a factor of $\sqrt{d/\log d}$. The algorithm however requires extra $O(n)$ factor in the runtime.

c) Gaussian noise, in Section 4, 5, 6, respectively. We provide a brief overview of our proofs in Section 7. Finally, we conclude with a few open problems and promising future directions in Section 8.

1.1. Related Works

In this section, we will discuss related work and compare existing results with ours. Existing theoretical results for OR-PCA fall into two categories:

a) The first category of approaches, more in-line with our own work, are based on Outlier-Pursuit (Xu et al. (2012a)) which optimizes a convex relaxation of the OR-PCA problem where the rank and column sparsity constraints are replaced by the trace-norm (sum of singular values) and $\lVert . \rVert_{2,1}$ (sum of the column lengths) penalties. That is, they solve the following optimization problem:

$$\text{(Outlier Pursuit): } \min \lVert L \rVert_* + \lambda \lVert C \rVert_{2,1} \ \ s.t \ \lVert M - L + C \rVert_F \leq \lVert N^* \rVert_F. \tag{1}$$

While outlier pursuit obtains optimal recovery guarantees in absence of any noise, a main drawback is that its computational complexity is quadratic in d, i.e., $O(nd^2)$. Moreover, in presence of noise the bounds given in Xu et al. (2012a) are $O(\sqrt{n})$ worse than our result. Extensions of Outlier-Pursuit to the partially observed setting (Chen et al. (2016)) and online setting Feng et al. (2013) have also been proposed but share the drawback of high computational complexity. Recently, Zhang et al. (2016) showed that Outlier Pursuit achieves recovery even with the fraction of outliers larger than

the information theoretic lower bound. However, this requires the outliers to be "well-spread" which in practice is restrictive; our results allow the corruption matrix to be constructed in adversarial manner although the corruptions cannot depend on the Gaussian noise in the setting (3) described in previous section.

b) The second line of approaches based on HR-PCA (Xu et al. (2013)) iteratively prune or reweigh data points which have a large influence on the singular vectors and select an estimate with the Robust Variance Estimator metric. When applied to our finite sample setting, these results cannot achieve exact recovery even in the noiseless case with a single outlier. Moreover, their running time in this setting is $O\left(n^2 dr\right)$ while ours is nearly linear in the input size. In the special case of Gaussian noise, these results incur at least a constant error $O\left(\sigma_1(L^*)\right)$ whereas our recovery guarantee scales linearly with the standard deviation of the noise $O\left(\sqrt{r}\sigma \log d\right)$, achieving exact recovery when $\sigma = 0$. Feng et al. (2012) propose a deterministic variant of HR-PCA and Yang and Xu (2015) extend HR-PCA to PCA-like algorithms like Sparse PCA and Non-Negative PCA.

In recent work, Diakonikolas et al. (2016) study the problem of robust parameter recovery which requires clean data to be sampled from a distribution. This assumption is more restrictive than the one on incoherence made in this paper which allows us to solve deterministic problems as well. In addition, our method is based on computationally efficient tools like the SVD and iterative hard thresholding. On the other hand, the methods described in Diakonikolas et al. (2016) are based on the use of the ellipsoid method which require $O(n^3 d^4)$ time which can be prohibitive for large-scale datasets.

There has been much recent work on the related problem of Robust PCA (Candès et al. (2011), Netrapalli et al. (2014), Yi et al. (2016), Cherapanamjeri et al. (2016)). In contrast to the setting considered here, the corruptions are assumed to be both row and column sparse i.e., unlike our setting no data can be corrupted in all its dimensions. This restriction allows stronger recovery guarantees but makes the results inapplicable to the setting of outlier robust PCA.

1.2. Notations

We use the following notations in this paper. For a vector v, $\|v\|$ and $\|v\|_2$ denote the ℓ_2 norm of v. For a matrix M, $\|M\|$ and $\|M\|_2$ denote the operator norm of M while $\|M\|_F$ denotes the Frobenius norm of M. $\sigma_k(M)$ denotes the kth largest singular value of M. SVD refers to singular value decomposition of a matrix. $\mathcal{SVD}_r(M)$ refers to the rank-r SVD of M. Given a matrix M, M_i denotes the ith column of M while $M_{i,:}$ denotes the ith row of M. Given a matrix $M \in \mathbb{R}^{d \times n}$ and a set $S \subseteq [n]$, M_S is defined as

$$(M_S)_i = \left\{ \begin{array}{ll} M_i & \text{for } i \in S, \\ 0 & \text{otherwise.} \end{array} \right.$$

$M_{\backslash S}$ denotes $M_{[n]\backslash S}$. $\text{Supp}(M)$ denotes column support of M, i.e., the set of indices of non-zero columns of M.

We use two hard -thresholding operators in this paper. Given a matrix R, the first hard thresholding operator, $\mathcal{HT}_\rho(R)$ denotes the set of indices j of the top ρ fraction of largest columns (in ℓ_2 norm) in R. Given a matrix R, the second hard-thresholding operator $\widetilde{\mathcal{HT}}_\zeta(R)$ is defined as,

$$\widetilde{\mathcal{HT}}_\zeta(N) = \{i : \text{ s.t. } \|N_i\|_2 \geq \zeta\}. \tag{2}$$

$P_\perp^U(M)$ denotes $(I - UU^\top)M$. For any set $\mathcal{S} \in \mathbb{R}^d$, we will use $\mathcal{P}_\mathcal{S}$ to denote the projection onto the set $\mathcal{S}$. We will also use $\mathcal{P}_U$ to denote the projection onto the column space of U for $U \in \mathbb{R}^{d \times r}$.

2. Problem Formulation

In this section, we will formally present the setting of the paper. We are given $M^* = D^* + C^* \in \mathbb{R}^{d \times n}$, where columns of D^* are inliers and C^* are outliers. Only an α fraction of the points are outliers, i.e., only $alpha$ fraction of the columns of C^* are non-zero. Broadly, we consider three scenarios:

- **OR-PCA** (Noiseless setting): The points in D^* lie *entirely* in a low-dimensional subspace i.e., $D^* = L^*$ is a rank-r matrix.

- **OR-PCAN** (Noisy setting): The points in D^* lie *approximately* in a low-dimensional subspace i.e., $D^* = L^* + N^*$ where L^* is the rank-r projection of D^* and N^* is the noise matrix.

- **OR-PCAG** (Gaussian noise setting): The points in D^* come from a low-dimensional subspace with additive Gaussian noise i.e., $D^* = L^* + N^*$ where L^* is a rank-r matrix and N^* is a Gaussian noise matrix i.e., each element of N^* is sampled iid from $\mathcal{N}(0, \sigma^2)$.

In all the above settings, the goal is to find the low dimensional subspace spanned by the columns of L^*.

This problem is in general ill-posed. Consider for instance the case, when most of the true data points L^* are zero and only an α fraction of them are non-zero. These points can either be considered inliers or outliers. In order to overcome this issue, standard assumption used in literature (Xu et al., 2012a; Feng et al., 2013) is that of incoherence. Also, incoherence is satisfied in several standard settings; for example, when the inliers are noise and uniform samples from a low-dimensional subspace.

Assumption 1 *Rank and incoherence of* L^***:** $L^* \in \mathbb{R}^{d \times n}$ *is a rank-r incoherent matrix, i.e.,* $\left\| e_i^\top V^* \right\|_2 \leq \mu \sqrt{\frac{r}{n}} \ \forall i \in [n]$, *where* $L^* = U^* \Sigma^* (V^*)^\top$ *is the SVD of* L^*.

Having already ruled out Vanilla PCA as a viable solution to the **OR-PCA** problem, we will discuss the shortcomings of an intuitive sampling based solution in order to build intuition. Consider the sampling strategy where $O(\mu^2 r)$ columns are randomly sampled from the matrix and the SVD of the submatrix so obtained used as an estimate of the singular subspace of D^*. With constant probability, this sampling scheme will avoid any corrupted points. Hence, repeating the experiment will obtain a subspace contained in the true principal subspace. Unfortunately, some singular vectors can also be completely unrepresented in such a sample and hence, a subspace which does not contain any of the top singular vectors may be obtained. Thus, the solution obtained via this method can be arbitrarily bad as the top singular vectors may contain most of the mass of the matrix. From this experiment, we conclude that for any viable pruning-based solution to **OR-PCA**, accurate representation of singular vectors and careful pruning of outliers are important. We incorporate and develop these intuitions in the design and analysis of our proposed solution to the **OR-PCA** problem and its variants.

3. Our Results

In this section, we will present our results for the three settings mentioned above.

3.1. OR-PCA – Noiseless Setting

Recall that in the noiseless setting, we observe $M^* = D^* + C^*$ where D^* is a rank-r, μ-incoherent matrix corresponding to clean data points and the column-support of C^* is at most αn. The following theorem is our main result for this setting.

Theorem 1 (Noise-less Setting) *Let M^*, D^* and C^* be as described above. If $\alpha \leq \frac{1}{128\mu^2 r}$, then Algorithm 1 run with parameters $\rho = \frac{1}{128\mu^2 r}$ and $T = \log \frac{10n\|M^*\|_2}{\epsilon}$, returns a subspace U such that,*

$$\left\| (I - UU^\top)D^* \right\|_F \leq \epsilon.$$

Remarks:

- Note that the guarantee of Theorem 1 can be right away converted to a bound on the subspace distance between U and that spanned by the columns of D^*. In particular, we obtain $\left\| (I - UU^\top)U^* \right\|_F \leq \epsilon/\sigma_r(D^*)$, where $\sigma_r(D^*)$ denotes the smallest singular value of D^* and U^* contains the singular vectors of D^*.

- Since the most time consuming step in each iteration is computing the top-r SVD of an $n \times d$ matrix, the total runtime of the algorithm is $O\left(ndr \log \frac{10n\|M^*\|_2}{\epsilon}\right)$.

- The above assumption on the column sparsity of C^* is tight up to constant factors i.e., we may construct an incoherent matrix L^* and column sparse matrix C^* such that it is not possible to recover the true column space of L^* when the column sparsity of C^* is larger than $\frac{1}{\mu^2 r}$.

3.2. OR-PCAN – Arbitrary Noise

We now consider the noisy setting. Here we observe $M^* = D^* + C^*$, where D^* is a near low rank matrix i.e., $D^* = L^* + N^*$ where L^* is the best rank r approximation to D^* and is a μ incoherent matrix, while N^* is a noise matrix. C^* is again column sparse with at most an α fraction of the columns being non-zero.

Theorem 2 (Arbitrary Noise) *Consider the setting above. If $\alpha \leq \frac{1}{128\mu^2 r}$, then Algorithm 2 when run with parameters $\rho = \frac{1}{128\mu^2 r}$, $\eta = 2\mu\sqrt{\frac{r}{n}}$ and $T = \log \frac{20\|M^*\|_2 \cdot n}{\epsilon}$ iterations, returns a subspace U such that:*

$$\left\| (I - UU^\top)L^* \right\|_F \leq 60\sqrt{r}\,\|N^*\|_F + \epsilon.$$

Remarks:

- The theorem shows that up to $\sqrt{r}\,\|N^*\|_F$ error, recovered directions U contains top r principle directions of inliers. We do not optimize the constants in our proof. In fact, we obtain a stronger result in Theorem 6 which for certain regime of noise N^* can lead to significantly better error bound. Note that when there is no noise i.e., $N^* = 0$, we recover Theorem 1.

- The guarantee here can again be converted to a bound on subspace distance. For instance, for any $k \leq r$, we have $\left\| (I - UU^\top)U^*_{[k]} \right\|_F \leq (60\sqrt{r}\,\|N^*\|_F + \epsilon)/\sigma_k(L^*)$, where U^*_k denotes the top-k left singular subspace of L^* and $\sigma_k(L^*)$ denotes the kth largest singular value of L^*.

7

- The total runtime of the algorithm is $O\left(ndr^2 \log \frac{10\|M^*\|_2 \cdot n}{\epsilon}\right)$. However, the outer loop over k in Algorithm 2 can be replaced by a binary search for values of k between 1 and r. This reduces the runtime to $O\left(ndr \log r \log \frac{10\|M^*\|_2 \cdot n}{\epsilon}\right)$. See Algorithm 4 for more details.

3.3. OR-PCAG– Gaussian Noise

We now consider the Gaussian noise setting. Here we observe $M^* = D^* + C^*$, where D^* is a near low rank matrix i.e., $D^* = L^* + N^*$ where L^* is a rank-r, μ incoherent matrix, while N^* is a Gaussian matrix with each entry sampled iid from $\mathcal{N}(0, \sigma^2)$. C^* is again column sparse with at most an α fraction of the columns being non-zero.

Theorem 3 (Gaussian Data) *Consider the setting mentioned above. Suppose $\alpha \leq \frac{1}{1024\mu^2 r}$. Then, Algorithm 3 stops after at most $T = \alpha n$ iterations and returns a subspace U such that:*

$$\|(I - UU^\top)L^*\|_2 \leq 4\sqrt{\log d}\|N^*\|_2.$$

with probability at least $1 - \delta$ as long as $n \geq \frac{16\mu^2 r^2 d}{c_1}\left[\log\left(\frac{1}{3\delta}\right) + d\log(80d)\right]$ for some absolute constants c_1 and c_2.

Remarks:

- Data points coming from a low-dimensional subspace with additive Gaussian noise is a standard statistical model that is used to justify PCA. Though this can be seen as a special case of arbitrary noise model, we get a much tighter bound than that obtained from Theorem 2.

- While Theorem 2 gives an asymptotic error bound of $\left\|(I - UU^\top)L^*\right\|_F \leq 60\sqrt{r}\|N^*\|_F$, Theorem 3 gives an asymptotic error bound of $\left\|(I - UU^\top)L^*\right\|_F \leq 4\sqrt{r\log d}\|N^*\|_2$. Note that the right hand sides above refer to Frobenius and operator norms respectively.

- The improvement mentioned above is obtained by carefully leveraging the fact that Gaussian random vectors are spread uniformly in all directions and that there is a small fraction of vectors which is correlated. However, in order to make this argument, we need $n = O\left(d^2\right)$. It is an open problem to get rid of this assumption.

- Note also that our result is tight in the sense that as $\sigma \to 0$, we recover the result of Theorem 1. However, the running time of the algorithm is $O(n^2 d)$ which is significantly worse than that of **TORP**. We leave design/analysis of a more efficient algorithm that achieves similar error bounds as Theorem 3 as an open problem.

- We can obtain the above result even when each column in N^* is drawn from a sub-Gaussian distribution rather than each entry being iid $\mathcal{N}(0, \sigma^2)$.

4. Outlier Robust PCA: Noiseless Setting

In this section, we present our algorithm **TORP**(Algorithm 1) that applies to the special case of noise-less data, i.e., when $M^* = L^* + C^*$, L^* is rank-r, μ-incoherent matrix. While restrictive, this setting allows us to illustrate the main ideas behind our algorithm approach and the analysis techniques in a relatively simpler fashion.

Recall that the goal is to estimate U^*, the left singular vectors of L^*. However, SVD of M^* can lead to singular vectors arbitrary far from U^*, because a few column of C^* can be so large that they can bias entire singular vectors in their direction.

Our algorithm instead tries to exploit two key structural properties of the problem: sparsity of C^* and incoherence of L^*. Our algorithm maintains a column-sparse estimate $C^{(t)}$ of C^*. Each iteration of the algorithm computes a low-rank approximation of an estimate of the inliers $M^* - C^{(t)} = L^* + C^* - C^{(t)}$. Note that if $(I - U^*(U^*)^\top)C^* = (I - U^*(U^*)^\top)C^{(t)}$, then left singular vectors of $M^* - C^{(t)}$ will be U^*.

Our next step finds *residual* length of each column M_i^* when projected on to the orthogonal subspace to $U^{(t)}$. If length of each outlier is smaller compared to the smallest singular value of L^*, then using sparsity of $C^{(t)}$ and C^*, we can show that $U^{(t)}$ is "close" to U^* in all directions. So, the residual of some of the outliers will stand out and those columns can be removed. This is achieved by the hard-thresholding step 5, 8 of Algorithm 1.

A big challenge in this scheme is that if a column of the perturbation matrix $C^* - C^{(t)}$ is "very" long compared to smaller singular values of L^*, then they can perturb some directions of U^* significantly. This will lead to a failure of the above thresholding approach. However, in such a case, some of the columns of $C^* - C^{(t)}$ will be close to a few spurious singular vectors in $U^{(t)}$ (our current estimate of U^*). Hence, projection of such outliers along $U^{(t)}$ will be inordinately long. On the other hand, due to incoherence of L^*, inliers' projection along $U^{(t)}$ can be bounded in magnitude. So, we can safely threshold out certain outliers. Steps 6, 8 of Algorithm 1 perform this thresholding operation.

In summary, our algorithm computes low-rank approximation of $M^* - C^{(t)}$ and uses the obtained singular vectors $U^{(t)}$ to threshold out a few columns of $C^{(t)}$ to obtain next estimate $C^{(t+1)}$ of C^*. See Algorithm 1 for a pseudo-code of our approach.

Time Complexity: Note that the computationally most expensive operation in each iteration is that of SVD which requires $O(ndr)$ time. So, the overall time complexity of the algorithm is $O(ndr \cdot T)$. As we show in Section 7, as long as C^* is column-sparse, $T \approx \log \frac{1}{\epsilon}$ suffices to obtain an ϵ approximation to U^*. So, the overall complexity of the algorithm is $O(ndr \cdot \log \frac{20\|M^*\|_2}{\epsilon})$. Note that typically SVD computation is approximate, while all three of our algorithms and analyses assumes exact SVD. However, extension of our analysis to allow for small additive error is straightforward and we ignore it in favor of simplicity and readability.

Parameters: The algorithm requires an estimate of rank r and threshold parameter ρ which in turn depends on estimate of incoherence μ of L^*. We propose to set these parameters via cross-validation. Note that setting rank to be any value larger than rank of L^* will lead to recovery of U^*, as long as C^* is sparse enough. Similarly, if estimation of μ is larger than incoherence of L^*, then it only effects number of corrupted columns in C^* that can be allowed. So, a simple cross-validation approach with appropriately chosen grid-size leads to recovery of U^* as long as C^* is sparse enough (as specified in Theorem 1).

5. Outlier Robust PCA: General Noise

In this section, we introduce our algorithm for the general case of Outlier Robust PCA with arbitrary inlier data $D^* = L^* + N^*$, i.e., the noise matrix N^* is arbitrary. Recall that the goal is to recover left singular vectors of L^*.

Algorithm 1 Thresholding based Outlier Robust PCA (**TORP**)

1: **Input**: Data $M^* \in \mathbb{R}^{d \times n}$, Target rank r, Threshold fraction ρ, Number of iterations T
2: $C^{(0)} \leftarrow 0$
3: **for** Iteration $t = 0$ to $t = T$ **do**
4:     $[U^{(t)}, \Sigma^{(t)}, V^{(t)}] \leftarrow \mathcal{SVD}_r \left(M^* - C^{(t)} \right) ; \; L^{(t)} \leftarrow U^{(t)} \Sigma^{(t)} (V^{(t)})^\top$ $\}$ Projection onto space of low rank matrices
5:     $R \leftarrow (I - U^{(t)} (U^{(t)})^\top) M^*$ /* Compute residual */
6:     $E \leftarrow (\Sigma^{(t)})^{-1} (U^{(t)})^\top M^*$ /* Compute incoherence */   Projection onto space of column sparse matrices
7:     $\mathcal{CS}^{(t+1)} \leftarrow \mathcal{HT}_\rho (R) \cup \mathcal{HT}_\rho (E)$
8:     $C^{(t+1)} \leftarrow M^*_{\mathcal{CS}^{(t+1)}}$
    /* Threshold points with high coherence or high residual*/
9: **end for**
10: $[U, \Sigma, V] \leftarrow \mathcal{SVD}_r \left(M^* - C^{(T+1)} \right)$
11: **Return:** U

Our algorithm for the general **OR-PCA** problem builds upon the **TORP** algorithm but with added complexity due to the presence of noise matrix N^*. That is the algorithm alternately updates estimate of the outliers $C^{(t)}$ and the principal direction $U^{(t)}$ using two thresholding operators along with SVD. However due to noise N^*, our estimate of $U^{(t)}$ gets perturbed furthermore leading to arbitrary perturbation of the singular vectors of U^* corresponding to smaller eigenvalues of L^* and hence cannot be recovered. To alleviate this concern, our **TORP-N** algorithm proceeds via a pair of nested loops:

Outer Iteration on k: The outer loop iterates over the rank-variable k which represents the rank of the principal subspace we wish to estimate.

Inner Iteration on t: The inner loop iteratively revises estimates of the principal subspace and a set of outliers until a stopping criteria is triggered.

Intuitively, as in Algorithm 1, each inner iteration of Algorithm 2 obtains a better estimate of C^* and top-k singular components of L^*. That is, the k-th outer iteration after several of such inner iterations estimates U^* up to $\approx \sigma_k(L^*)$. But when the noise $\|N^*\|_F$ becomes comparable to the kth singular value of L^*, then the algorithm terminates (Line 14, Algorithm 2) as at that point it may not be possible to estimate the remaining singular vectors of L^*. As we don't know $\|N^*\|_F$ explicitly, we detect this event based on the number of data points which have a large influence on the estimated singular vectors (see lines 11, 12 of Algorithm 2).

Roughly, our stopping criterion allows us to make two statements regarding the termination of the algorithm:

1. When the algorithm terminates, the outlier columns that we have not thresholded will only have small influence on the estimated principal vectors. This is because all points with large influence will be thresholded before the estimate is computed.

2. The algorithm will not terminate if $\sigma_k^* >> \|N^*\|_F$: The bound on $\|N^*\|_F$ ensures that not many inlier points can have large influence on estimate of the kth singular vector.

By using the above two claims, our analysis shows that **TORP-N** recovers U^* up to $\sim \|N^*\|_F$ error.

Time Complexity: Time complexity of each inner iteration of **TORP-N** is $O(ndk)$. Hence, overall time complexity is $O(ndr^2)$, as k can be as large as r. However, using a slightly more complicated algorithm and analysis (see Algorithm 4), we can search for appropriate k using binary search, so the time complexity of the algorithm can be improved to $O(ndr \log r)$.

Parameter Estimation: The algorithm requires 3 parameters: rank r, threshold ρ which depends on incoherence μ of L^* and expressivity parameter η. We can search for these parameters using a coarse-grid search as estimates of these parameters up to constants are enough for our algorithm to succeed albeit with a slightly stricter restriction (by constant factors) on the number of corrupted data points.

Algorithm 2 Thresholding based Noisy Outlier Robust PCA (**TORP-N**)

1: **Input**: Corrupted matrix $M^* \in \mathbb{R}^{d \times n}$, Target rank r, Expressivity parameter η, Threshold fraction ρ, Inner iterations T
2: **for** $k = 1$ to $k = r$ **do**
3:    $C^{(0)} \leftarrow 0, \tau \leftarrow false$
4:    **for** $t = 0$ to $t = T$ **do**
5:       $[U^{(t)}, \Sigma^{(t)}, V^{(t)}] \leftarrow \mathcal{SVD}_k \left(M^* - C^{(t)}\right), L^{(t)} \leftarrow U^{(t)} \Sigma^{(t)} (V^{(t)})^\top$ Projection onto space of low rank matrices
6:       $E \leftarrow (\Sigma^{(t)})^{-1}(U^{(t)})^\top M^*$ /* *Compute Incoherence* */
7:       $R \leftarrow (I - U^{(t)}(U^{(t)})^\top)M^*$ /* *Compute residual* */
8:       $\mathcal{CS}^{(t+1)} \leftarrow \mathcal{HT}_{2\rho}(M^*, E) \cup \mathcal{HT}_\rho(M^*, R)$ Projection onto space of column sparse matrices
9:       $C^{(t+1)} \leftarrow M^*_{\mathcal{CS}^{(t+1)}}$
10:      $n_{\text{thres}} \leftarrow |\{i : \|E_i\| \geq \eta\}|$ /* *Compute high incoherence points* */
11:      $\tau \leftarrow \tau \vee (n_{\text{thres}} \geq 2\rho n)$ /* *Check termination conditions* */
12:    **end for**
13:    **if** τ **then**
14:       $break$
15:    **end if**
16:    $[U, \Sigma, V] \leftarrow \mathcal{SVD}_k \left(M^* - C^{(T+1)}\right)$
17: **end for**
18: **Return:** U

---

6. Outlier Robust PCA: Gaussian Noise

In this section, we present our algorithm for the special case of the Outlier Robust PCA problem when inlier points are generated using a standard Gaussian noise model. That is, when $D^* = L^* + N^* \in \mathbb{R}^{d \times n}$ where each entry of the noise matrix N^* is sampled i.i.d. from $\mathcal{N}(0, \sigma^2)$. Our result for arbitrary N^* (Theorem 2) estimates U^* up to $\sim \|N^*\|_F$ error, which is $\Omega(\sigma\sqrt{dn})$ for Gaussian noise. However, using a slight variant of Algorithm 2 and exploiting the noise structure, Algorithm 3 is able to estimate U^* up to $\sigma\sqrt{n \log d}$ error, which is better than the previous one by a factor of $O\left(\sqrt{d/\log d}\right)$.

At a high level the philosophy of our **TORP-G** algorithm is similar to **TORP**, i.e., we iteratively revise estimate of C^* and the top singular vectors U^* using SVD and thresholding. That is, we

---

Algorithm 3 Thresholding based Outlier Robust PCA with Gaussian Noise (**TORP-G**)

1: **Input**: Corrupted matrix $M^* \in \mathbb{R}^{d \times n}$, Target rank r, Incoherence Parameter μ, Noise Level σ

2: $M \leftarrow M^*, \tau \leftarrow true$

3: $\zeta_1 \leftarrow \sigma \left(\frac{5}{4} \mu \sqrt{r} + d^{\frac{1}{2}} + 2d^{\frac{1}{4}} \sqrt{\log \left(\frac{\mu^2 r}{c_2} \right)} \right), \zeta_2 \leftarrow \sigma \sqrt{2r} \left(\frac{5}{4} \mu + 2 \sqrt{\log \left(\frac{\mu^2 r^2 d}{c_1} \right)} \right)$

4: $C^{(0)} \leftarrow 0, \mathcal{CS}^{(0)} \leftarrow \{\}, \mathcal{CS}^{(-1)} \leftarrow \{0\}, t \leftarrow 0$

5: **while** $\mathcal{CS}^{(t)} \neq \mathcal{CS}^{(t-1)}$ **do**

6: $\quad [U^{(t)}, \Sigma^{(t)}, V^{(t)}] \leftarrow \mathcal{SVD}_{r+1}(M^* - C^{(t)}), \quad L^{(t)} \leftarrow U^{(t)}\Sigma^{(t)}(V^{(t)})^\top$ $\left. \begin{array}{c} \\ \\ \end{array} \right\}$ Projection onto space of low rank matrices

7: $\quad \mathcal{E}^{(t)} \leftarrow \{x : x = U^{(t)}\Sigma^{(t)}y \text{ for some } \|y\| \leq 2\mu\sqrt{r/n}\}$

8: $\quad \widehat{L}^{(t)} \leftarrow \mathcal{P}_{\mathcal{E}^{(t)}}(L^{(t)})$ */* Projection onto incoherent matrices */

9: $\quad \mathcal{I} \leftarrow \left\{ i : \left\| L_i^{(t)} - \widehat{L}_i^{(t)} \right\| > \zeta_2 \right\}$ */* Points with large influence */

10: $\quad \mathcal{CS}^{(t+1)} \leftarrow \mathcal{CS}^{(t)} \cup \widetilde{\mathcal{HT}_{\zeta_1}} \left(L^{(t)} - \widehat{L}^{(t)} \right)$

   / Updating support of outliers */

11: $\quad$ **if** $|I| \geq \frac{24nc_1}{\mu^2 dr}$ **then**

12: $\quad\quad \mathcal{CS}^{(t+1)} \leftarrow \mathcal{CS}^{(t+1)} \cup \widetilde{\mathcal{HT}_{\zeta_2}} \left(L^{(t)} - \widehat{L}^{(t)} \right)$

   / Update support of outliers */

13: $\quad$ **end if**

14: $\quad C^{(t+1)} \leftarrow M^*_{\mathcal{CS}^{(t+1)}}$

   / Compute Sparse Projection */

15: $\quad t \leftarrow t + 1$

16: **end while**

17: **Return:** U

Projection onto space of column sparse matrices

---

iteratively threshold columns of M^*, that we estimate are corrupted. However, due to Gaussian noise structure our thresholding step is significantly different than that of **TORP** or **TORP-N**.

In particular, the choice of our thresholding criteria (see lines 10, 12 of Algorithm 3) uses the following two insights:

1. Length based thresholding (with respect to ζ_1—line 10 of Algorithm 3): This thresholding step is used to ensure that the noise in each data point is at most $O\left(\sigma\sqrt{d}\right)$. As the length of random Gaussian vector is at most $O\left(\sigma\sqrt{d}\right)$ with high probability, only a small number of inliers are thresholded in this step (Lemma 16).

2. Projection based thresholding (with respect to ζ_2—line 12 of Algorithm 3): In this step, we threshold points that have large projection along the estimated principal subspace. Note that out of n columns of N^*, at most $O\left(\frac{1}{\mu^2 rd}\right)$ fraction of points have projected lengths greater than $O\left(\sigma\sqrt{\log(d)}\right)$ along any direction (Lemma 19). Thus, chances of a inliers being thresholded in this step is low. On the other hand, any outlier that heavily influences a principal direction will be thresholded by this step.

Algorithm 3 provides a detailed pseudo-code of **TORP-G**. Step 6 of the algorithm computes **rank-($\mathbf{r+1}$)** SVD of the estimate of inlier matrix $M^* - C^*$. Step 7 defines a set of vectors, whose projec-

tion onto singular vectors of $L^{(t)}$ is "typical" for an inlier which is composed of a low-dimensional point perturbed by Gaussian noise vector of length $O(\sigma\sqrt{d})$.

This set is used in step 10 to threshold outliers using the hard-thresholding operator $\widetilde{\mathcal{HT}_\zeta}$ as defined in (2). Next, the set $\mathcal{I}$ consists of points which have a large influence on the estimated principal components. In the absence of outliers, the size of this set is bounded by $\frac{12nc_1}{\mu^2 dr}$ with high probability. A large deviation in the size of this set indicates the presence of of outliers and the entire set is thresholded.

Note on Approximate Computation: We would like to note that the projection operator defined in step 8 of the algorithm can be computed efficiently to arbitrary accuracy. A pseudo-code for computing the required projection can be found in Algorithm 5. Algorithm 5 reduces the problem to the univariate problem of finding the root of a monotonically decreasing function in a bounded interval which can be found efficiently via binary search. For the sake of simplicity, we assume that the projection step and the $\mathcal{SVD}$ are computed exactly. Our analysis can be extended to the case where the projection and $\mathcal{SVD}$ are computed approximately with some added technical difficulty.

7. Proof Overview

In this section, we provide a brief overview of our analysis for the three main results.

7.1. Noiseless Setting—Theorem 1

In this section, we present the proof of Theorem 1. Recall that we are given $M^* = D^* + C^*$, where $D^* = L^*$ is a rank-r, μ-incoherent matrix and C^* has at most a fraction of ρ non-zero columns. We can assume with out loss of generality that D^* and C^* have disjoint column supports as we can rewrite M_i^*, for $i \in \text{Supp}(C^*)$, as $M_i^* = D_i^* + C_i^* = 0 + (C_i^* + D_i^*)$ thus absorbing D_i^* in C_i^* itself.

Our proof consists of three main steps. Given any set of columns S and letting $[U_{\backslash S}, \Sigma_{\backslash S}, V_{\backslash S}]$ be the top-r SVD of $M_{\backslash S}^*$, we establish the following:

Step 1: Every non-zero column of D^* has significantly smaller residual when projected onto subspace orthogonal to $U_{\backslash S}$ than the norm of corrupted columns of $M_{\backslash S}^*$ (Lemma 4), so its likelihood of being thresholded (Line 6, 8 of Algorithm 1) is small,

Step 2: Every non-zero column of D^* has small incoherence with respect to $[U_{\backslash S}, \Sigma_{\backslash S}, V_{\backslash S}]$ (Lemma 5), i.e., its projection onto $U_{\backslash S}$ cannot be "too large". Hence, its likelihood of being thresholded (Line 7,8 of Algorithm 1) is also small,

Step 3: Any non-zero column of C^* which has small residual and incoherence compared to those of a non-zero column of D^* and hence won't be thresholded by Algorithm 1, has small residual when projected onto U^*. That is, the column itself is close to subspace spanned by U^* and hence does not effect estimation of U^* (Proof of Theorem 1).

That is, either a corrupted column will be thresholded or it is close to U^* while inliers (D^*) have little likelihood of being thresholded (step 1,2 above). We now present the formal statements and their proofs. We start with two lemmata establishing Steps 1,2 above. Detailed proofs of the lemmata are given in Appendix B.1 and B.2, respectively.

Lemma 4 *Consider the setting of Theorem 1. Let $S \subset [n]$ denote a subset of columns of M^* such that $|S| \leq 2\rho n$. Let $M^*_{\setminus S}$ $(L^*_{\setminus S})$ be obtained from M^* (L^*) by setting the columns corresponding to indices specified in S to 0. Let $U_{\setminus S}\Sigma_{\setminus S}(V_{\setminus S})^\top$ $(U^*_{\setminus S}\Sigma^*_{\setminus S}(V^*_{\setminus S})^\top)$ be the rank-r $\mathcal{SVD}$ of $M^*_{\setminus S}$ $(L^*_{\setminus S})$, then $\forall i$:*

$$\left\| (I - U_{\setminus S}(U_{\setminus S})^\top) L_i^* \right\| \leq \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left\| (I - U^*(U^*)^\top) M^*_{\setminus S} \right\|$$

Lemma 5 *Under the setting of Lemma 4, we have for every i:*

$$\left\| \Sigma^{-1}_{\setminus S} U^\top_{\setminus S} L_i^* \right\| \leq \frac{33}{32} \mu \sqrt{\frac{r}{n}}.$$

We now present the proof of Theorem 1 where we illustrate Step 3:

Proof We will start by showing the quantity $\left\| (I - U^*(U^*)^\top) M^{(t+1)} \right\|_F$ decreases at a geometric rate, where $M^{(t+1)} = M^* - C^{(t+1)}$. Let $Q^{(t)}$ denote the columns of C^* that are not thresholded in iteration t. Also let $S^{(t)}$ denote the columns of L^* that are thresholded in iteration t. Let $\widetilde{L}^{(t+1)} := L^*_{\setminus S^{(t)}}$, $\widetilde{C}^{(t+1)} := C^*_{Q^{(t)}}$, and $P_\perp^U(M) = (I - U(U)^\top)M$. Then, we have:

$$\left\| P_\perp^{U^*}(M^{(t+1)}) \right\|_F^2 = \left\| P_\perp^{U^*}(\widetilde{L}^{(t+1)} + \widetilde{C}^{(t+1)}) \right\|_F^2 = \left\| P_\perp^{U^*}(\widetilde{C}^{(t+1)}) \right\|_F^2$$

$$= \sum_{j \in Q^{(t)}} \left\| P_\perp^{U^*}(U^{(t)}\Sigma^{(t)}W_j^{(t)} + R_j^{(t)}) \right\|^2 \leq 2 \sum_{j \in Q^{(t)}} \left\| P_\perp^{U^*}(U^{(t)})\Sigma^{(t)}W_j^{(t)} \right\|^2 + \left\| R_j^{(t)} \right\|^2, \quad (3)$$

where $W_j^{(t)} = (\Sigma^{(t)})^{-1}(U^{(t)})^T C_j^*$ and $R_j^{(t)} = P_\perp^{U^{(t)}}(C_j^*)$, $\forall j \in Q^{(t)}$. The last inequality follows from triangle inequality and the fact that $(a + b)^2 \leq 2(a^2 + b^2)$.

Recall, that we threshold a particular column l in iteration t based on $\left\| P_\perp^{U^{(t)}}(M_l^*) \right\|$ and $\left\| (\Sigma^{(t)})^{-1}(U^{(t)})^\top M_l^* \right\|$. For a particular $j \in Q^{(t)}$ that wasn't thresholded in iteration t, we know that there exists a column i_j such that $\left\| (\Sigma^{(t)})^{-1}(U^{(t)})^\top L_{i_j}^* \right\| \geq \left\| (\Sigma^{(t)})^{-1}(U^{(t)})^\top C_j^* \right\|$. Similarly, there exists a column k_j such that $\left\| P_\perp^{U^{(t)}}(L_{k_j}^*) \right\| \geq \left\| P_\perp^{U^{(t)}}(C_j^*) \right\|$. From Lemmas 5 and 4, we have:

$$\left\| W_j^{(t)} \right\| \leq \frac{33}{32} \mu \sqrt{\frac{r}{n}} \qquad \left\| R_j^{(t)} \right\| \leq \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left\| P_\perp^{U^*}(M^{(t)}) \right\| \qquad (4)$$

Using (3) and (4), we have:

$$\left\| P_\perp^{U^*}(M^{(t+1)}) \right\|_F^2 \leq 2 \sum_{j \in Q^{(t)}} \left(\frac{33}{32}\right)^2 \mu^2 \frac{r}{n} \left\| P_\perp^{U^*}(U^{(t)})\Sigma^{(t)} \right\|^2 + \left(\frac{33}{32}\right)^2 \mu^2 \frac{r}{n} \left\| P_\perp^{U^*}(M^{(t)}) \right\|^2$$

$$\leq 4 \cdot \frac{9}{8} \cdot \frac{\mu^2 r}{n} \cdot \rho n \cdot \| P_\perp^{U^*}(M^{(t)}) \|^2 \leq \frac{1}{4} \left\| P_\perp^{U^*}(M^{(t)}) \right\|^2,$$

where second last inequality follows from $|S^{(t)}| \leq \rho n$ and the last inequality follows from $\rho \leq \alpha \leq \frac{1}{128\mu^2 r}$. By recursively applying the above inequality, we obtain:

$$\left\| P_\perp^{U^*}(M^{(T+1)}) \right\|_F \leq \frac{\epsilon}{20n}. \qquad (5)$$

Also, note that using variational characterization of SVD, we have
$\|P_\perp^U(M^{(T+1)})\|_F \le \|P_\perp^{U^*}(M^{(T+1)})\|_F$. Theorem now follows from the following argument:

$$\left\|P_\perp^U(L^*)\right\|_F^2 = \left\|P_\perp^U(\widetilde{L}^{(T+1)})\right\|_F^2 + \sum_{i \in S^{(T)}} \left\|P_\perp^U(L_i^*)\right\|^2$$

$$\le \left\|P_\perp^U(M^{(T+1)})\right\|_F^2 + \sum_{i \in S^{(T)}} \frac{33^2}{32^2} \mu^2 \frac{r}{n} \left\|P_\perp^U(\widetilde{L}^{(T+1)})\right\|^2$$

$$\le \left\|P_\perp^U(M^{(T+1)})\right\|_F^2 + 2\rho n \left(\frac{33}{32}\right)^2 \mu^2 \frac{r}{n} \left\|P_\perp^U(M^{(T+1)})\right\|_F^2 \le \frac{\epsilon}{10n},$$

where the first inequality follows from Lemma 11 and using $M^{(T+1)} = \widetilde{L}^{(T+1)} + \widetilde{C}^{(T+1)}$, and the fact that $\widetilde{L}^{(T+1)}$ and $\widetilde{C}^{(T+1)}$ have different support. The second inequality follows from the fact that at most $2\rho \cdot n$ points can be thresholded and then using (5). ∎

7.2. Arbitrary Noise—Theorem 2

We now briefly discuss the proof of Theorem 2. In fact, we prove a stronger result:

Theorem 6 *Let $M^* = L^* + C^* + N^*$ such that L^* satisfies Assumption 1 and C^* has column sparsity $\alpha \le \frac{1}{128\mu^2 r}$. Furthermore, suppose that $\|N^*\|_F \le \frac{\sigma_k^*}{16}$ for some $k \in [r]$. Then, Algorithm 2 run with $\rho = \frac{1}{128\mu^2 r}$ and η set to $2\mu\sqrt{\frac{r}{n}}$ with $T = \log \frac{20\|M^*\|_2 \cdot n}{\epsilon}$, returns a subspace U such that:*

$$\left\|(I - UU^\top)L^*\right\|_F \le 3\left\|(I - U_k^*(U_k^*)^\top)L^*\right\|_F + 9\|N^*\|_F + \frac{\epsilon}{10n}.$$

Intuitively, the proof of Theorem 6 proceeds along the same lines as that of Theorem 1 but requires significantly more careful analysis due to presence of noise and due to the outer loop. For example, due to the presence of noise, we cannot guarantee that Lemma 5, that was critical to proof of Theorem 1, holds for all columns i. We show instead that the number of data points which have a large influence on the top-k singular vectors is bounded (see Lemma 17). This ensures that the algorithm at least reaches the kth stage of the outer iteration before terminating. Similarly, we generalize Lemma 4 to handle N^* (see Lemma 18). Finally, we present the key lemma that shows that if the algorithm does not terminate in the kth outer iteration, then it would have obtained a good approximation to the top-k principal subspace of L^*.

Lemma 7 *Asume the conditions of Theorem 2. Furthermore, assume that Algorithm 2 has not terminated during the kth outer iteration. Then, the iterate U at the end of the kth outer iteration satisfies:*

$$\left\|(I - UU^\top)L^*\right\|_F \le 3\left\|(I - U_{1:k}^*(U_{1:k}^*)^\top)L^*\right\|_F + 9\|N^*\|_F + \frac{\epsilon}{10n},$$

when Algorithm 2 has been run with parameters $\rho = \frac{1}{128\mu^2 r}$ and $\eta = 2\mu\sqrt{\frac{r}{n}}$.

See Appendix B.5 for a detailed proof. We can now prove Theorem 6 as follows:

Proof Note that by Lemma 17, the algorithm does not terminate before the completion of kth outer iteration. Now, suppose that the algorithm terminates at some iteration $k' > k$. Then, by Lemma 7, we have:

$$\left\|P_\perp^U(L^*)\right\| \leq 3\left\|P_\perp^{U^*_{1:k'-1}}(L^*)\right\|_F + 9\left\|N^*\right\|_F + \frac{\epsilon}{10n} \leq 3\left\|P_\perp^{U^*_{1:k}}(L^*)\right\|_F + 9\left\|N^*\right\|_F + \frac{\epsilon}{10n}.$$

This concludes the proof of the Theorem. ∎

7.3. Gaussian Noise—Theorem 3

Our analysis of **TORP-G** show that the algorithm maintains the following critical invariant with high probability:

Invariant 1 *We assume that the following hold with respect to the two thresholding steps used in Algorithm 3.*

1. *With respect to ζ_1: If a column $i \notin Supp(C^*)$ is thresholded, then the following condition holds:*
$$\|N_i^*\| \geq \sigma\left(\sqrt{d} + 2d^{\frac{1}{4}}\sqrt{\log\left(\frac{\mu^2 r}{c_2}\right)}\right).$$
and consequently only $\frac{3nc_2}{2\mu^2 r}$ points are removed in this step.

2. *With respect to ζ_2: If a thresholding step occurs due to the second thresholding step with ζ_2, then at least half the points thresholded in this step are corrupted points.*

Lemma 8 *Assume the conditions of Theorem 3. Then, Invariant 1 holds at any point in the running of Algorithm 3 with probability at least $1 - \delta$.*

See Appendix B.8 for a detailed proof.

Our proof then uses the above invariant along with a careful analysis of each of the two thresholding steps (Line 10, 12) to obtain the desired result. See Appendix C for a detailed proof.

8. Conclusions and Future Works

In this paper, we studied the outlier robust PCA problem. We proposed a novel thresholding based approach that, under standard regularity conditions, can accurately recover the top principal directions of the clean data points, as long as the number of outliers is less than $O(1/r)$ which is information theoretically tight up to constant factors. For noiseless or arbitrary noise case, our algorithms are based on two thresholding operators to detect outliers and leads to better recovery compared to existing methods in essentially the same time as that taken by vanilla PCA. For Gaussian noise, we obtain improved recovery guarantees but at a cost of higher run time.

Though our bounds have significant improvement over existing ones, they are still weaker than guarantees obtained by vanilla PCA (with out outliers). For instance, for arbitrary noise, our errors are bounded in the Frobenius norm. In contrast, in absence of outliers, SVD can estimate the

principal directions in operator norm. A challenging and important open problem is if the principal directions can be estimated in operator norm even in the presence of outliers.

Similarly, for Gaussian noise, where each entry has variance σ^2, our result obtains an error bound of $O(\sigma\sqrt{n})$ which is significantly better than the Frobenius norm bound we get for arbitrary noise. Unfortunately, our algorithm for the Gaussian setting is nearly a factor of n slower than that for vanilla PCA. In order for this to be practical, it is very important to design an algorithm for this setting with nearly the same runtime as that of vanilla PCA.

References

Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

Yudong Chen, Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Matrix completion with column manipulation: Near-optimal sample-robustness-rank tradeoffs. *IEEE Trans. Information Theory*, 62(1):503–526, 2016. doi: 10.1109/TIT.2015.2499247. URL http://dx.doi.org/10.1109/TIT.2015.2499247.

Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain. Nearly-optimal robust matrix completion. *CoRR*, abs/1606.07315, 2016. URL http://arxiv.org/abs/1606.07315.

Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 81–90, 2013. doi: 10.1145/2488608.2488620. URL http://doi.acm.org/10.1145/2488608.2488620.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 655–664, 2016. doi: 10.1109/FOCS.2016.85. URL https://doi.org/10.1109/FOCS.2016.85.

Jiashi Feng, Huan Xu, and Shuicheng Yan. Robust PCA in high-dimension: A deterministic approach. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012. URL http://icml.cc/2012/papers/136.pdf.

Jiashi Feng, Huan Xu, and Shuicheng Yan. Online robust PCA via stochastic optimization. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 404–412, 2013. URL http://papers.nips.cc/paper/5131-online-robust-pca-via-stochastic-optimization.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3rd edition, 1996.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Praneeth Netrapalli, Niranjan U N, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1107–1115. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5430-non-convex-robust-pca.pdf.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010. URL http://arxiv.org/abs/1011.3027.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. *IEEE Trans. Information Theory*, 58(5):3047–3064, 2012a. doi: 10.1109/TIT.2011.2173156. URL http://dx.doi.org/10.1109/TIT.2011.2173156.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012b.

Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust PCA: the high-dimensional case. *IEEE Trans. Information Theory*, 59(1):546–572, 2013. doi: 10.1109/TIT.2012.2212415. URL http://dx.doi.org/10.1109/TIT.2012.2212415.

Wenzhuo Yang and Huan Xu. A unified framework for outlier-robust pca-like algorithms. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 484–493, 2015. URL http://jmlr.org/proceedings/papers/v37/yangc15.html.

Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4152–4160, 2016. URL http://papers.nips.cc/paper/6445-fast-algorithms-for-robust-pca-via-gradient-descent.

Hongyang Zhang, Zhouchen Lin, and Chao Zhang. Completing low-rank matrices with corrupted samples from few coefficients in general basis. *IEEE Trans. Information Theory*, 62(8):4748–4768, 2016. doi: 10.1109/TIT.2016.2573311. URL http://dx.doi.org/10.1109/TIT.2016.2573311.

Appendix A. Supplementary Results and Preliminaries

Here, we will state and prove a few results useful in proving our main theorems. We will start by restating Weyl's perturbation inequality from Bhatia (1997).

Theorem 9 *Let $A \in \mathbb{R}^{d \times n}$. Furthermore, let $B = A + E$ for some matrix E. Then, we have that:*

$$|\sigma(A)_i - \sigma(B)_i| \leq \|E\| \quad \forall i \in min(d,n)$$

In the next lemma, we show that the singular values of the sum of two matrices with disjoint column supports are greater than either of the two matrices individually.

Lemma 10 *Let $A \in \mathbb{R}^{d \times n}$ and $B \in \mathbb{R}^{d \times n}$ be two matrices with disjoint column support. Then, we have $\forall i \in \min(d, n)$:*

$$\max(\sigma_i(A), \sigma_i(B)) \leq \sigma_i(A + B)$$

Proof Let the SVD of A and B be $U_A \Sigma_A V_A^\top$ and $U_B \Sigma_B V_B^\top$ respectively. The lemma holds for $i = 0$ as $\left\| v^\top(A + B) \right\| \geq \max(\left\| v^\top A \right\|, \left\| v^\top B \right\|)$. For any matrix M, $\sigma_i(M) = \min\limits_{U \in \mathbb{R}^{d \times (i-1)}} \left\| (I - UU^\top)M \right\| \forall i > 1$. For any U, there exist v_1 and v_2 in $Span((U_A)_{[i]})$ and $Span((U_B)_{[i]})$ respectively and $v_1^\top U = v_2^\top U = 0$. This is because the rank of $Span(U)$ is at most $(i - 1)$ and $Span((U_A)_{[i]})$ and $Span((U_B)_{[i]})$ are both rank-i subspaces. Now, we have $\left\| v_1^\top(A + B) \right\| \geq \left\| v_1^\top A \right\| \geq \sigma_i A$ and $\left\| v_1^\top(A + B) \right\| \geq \left\| v_1^\top B \right\| \geq \sigma_i(B)$. The lemma by using either v_1 or v_2 for any U. ■

The next lemma shows that an incoherent matrix remains incoherent even if a small number of columns have been set to 0.

Lemma 11 *Let $L \in \mathbb{R}^{d \times n}$ be a μ-column-incoherent, rank-r matrix. Let $S \subset [n]$ such that $|S| \leq \frac{n}{32\mu^2 r}$. Let $[U, \Sigma, V]$ and $[U_{\backslash S}, \Sigma_{\backslash S}, V_{\backslash S}]$ denote the SVDs of L and $L_{\backslash S}$ respectively. Then, the following hold $\forall i \in [n]$:*

$$\text{Claim 1: } \left\| e_i^\top V_{\backslash S} \right\| \leq \frac{33}{32}\mu\sqrt{\frac{r}{n}} \qquad \text{Claim 2: } \frac{31}{32}\sigma_i(L) \leq \sigma_i(L_{\backslash S}) \leq \sigma_i(L)$$

Furthermore, each column $L_i \forall i \in [n]$ can be expressed as:

$$\text{Claim 3: } L_i = U_{\backslash S}\Sigma_{\backslash S}w_i \text{ with } \|w_i\| \leq \frac{33}{32}\mu\sqrt{\frac{r}{n}}$$

Proof Let T be defined as the matrix V with the rows in set S set to 0. We will first begin by proving that T is full rank. Let $u \in \mathbb{R}^r$ and $\|u\| = 1$:

$$1 = \|Vu\| \geq \|Tu\| = \left(\sum_{i=1}^{n} \langle u, V_{i,:} \rangle^2 - \sum_{j \in S} \langle u, V_{j,:} \rangle^2 \right)^{\frac{1}{2}} \geq \left(1 - \sum_{j \in S} \|V_{j,:}\|^2 \right)^{\frac{1}{2}} \geq \left(1 - \frac{1}{32} \right)^{\frac{1}{2}}$$

where the third inequality is obtained from the bound on $|S|$ and $\|V_{j,:}\|$. Since T and $V_{\backslash S}$ have the same column space, there exists a matrix $R \in \mathbb{R}^{r \times r}$ such that $TR = V_{\backslash S}$. We know that R is full rank. We will now prove bounds on the singular values of R. For any $u \in \mathbb{R}^r$ and $\|u\| = 1$

$$\|Ru\| = \|VRu\| \geq \|TRu\| = \|V_{\backslash S}u\| = 1 = \|V_{\backslash S}u\| = \|TRu\| \geq \left(1 - \frac{1}{32} \right)^{\frac{1}{2}} \|Ru\|$$

From this, we obtain the following inequality:

$$1 \leq \|Ru\| \leq \left(1 - \frac{1}{32}\right)^{-\frac{1}{2}}$$

From this, we have the first claim of the lemma as $TR = V_{\backslash S}$. We also know that $U_{\backslash S}\Sigma_{\backslash S}V_{\backslash S}^\top = U\Sigma T^\top$. Writing $V_{\backslash S}$ as TR, we have $U_{\backslash S}\Sigma_{\backslash S}R^\top T^\top = U\Sigma T^\top$. Using the fact that $T^\top$ is full rank, we have $U_{\backslash S}\Sigma_{\backslash S}R^\top = U\Sigma$. From this we have that $L = U\Sigma V^\top = U_{\backslash S}\Sigma_{\backslash S}R^\top V^\top$. Choosing $w_i = (R^\top V^\top)_i$, the second claim of the lemma follows.

For the final claim of the lemma, note that the singular values of $L_{\backslash S}$ are the same as the singular values of $U\Sigma(R^\top)^{-1}$. We know that $\sigma_{k+1}(L_{\backslash S}) = \min\limits_{Q \in \mathbb{R}^{d \times k}} \left\|(I - QQ^\top)U\Sigma(R^\top)^{-1}\right\|$. The upper bound follows from setting Q to be the first k singular vectors of L and our bound on the singular values of R. For the lower bound, consider any $Q \in \mathbb{R}^{d \times k}$. $Span(Q)$ is a subspace of rank at most k. Therefore, there exists $v \in Span(U_1)$ such that $\|v\| = 1$ and $v^\top Q = 0$. We now have

$$\left\|v^\top(I - QQ^\top)U\Sigma(R^\top)^{-1}\right\| = \left\|v^\top U\Sigma(R^\top)^{-1}\right\| \geq \frac{\sigma_{k+1}(L)}{\|R\|} \geq \frac{31}{32}\sigma_{k+1}(M)$$

Where the last inequality follows from our bounds on the singular values of R and noting that the singular values of R^{-1} are the inverses of the singular values of R. This proves the third claim of the lemma. $\blacksquare$

We begin by stating a lemma used for bounding the length of Gaussian random vectors from Laurent and Massart (2000):

Lemma 12 *Let $Y_1, Y_2, \cdots, Y_d$ be i.i.d Gaussian random variables with mean 0 and variance 1. Let $Z = \sum\limits_{i=1}^{d} \left(Y_i^2 - 1\right)$. Then the following inequality holds for any positive x:*

$$\mathbb{P}\left(Z \geq 2\sqrt{dx} + 2x\right) \leq \exp(-x)$$

We will now state the famous Bernstein's Inequality from Boucheron et al. (2013).

Theorem 13 *Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that there exist positive real numbers ν and c such that $\sum\limits_{i=1}^{n} \mathbb{E}\left[X_i^2\right] \leq \nu$ and*

$$\sum\limits_{i=1}^{n} \mathbb{E}\left[(X_i)_+^q\right] \leq \frac{q!}{2}\nu c^{q-2} \, \forall \, q \geq 3,$$

where $x_+ = \max(x, 0)$.
If $S = \sum\limits_{i=1}^{n}\left(X_i - \mathbb{E}[X_i]\right)$, then $\forall t \geq 0$, we have:

$$\mathcal{P}\left(S \geq \sqrt{2\nu t} + ct\right) \leq \exp(-t)$$

We will now restate a lemma for controlling the singular values of a matrix with Gaussian random entries from Vershynin (2010).

Lemma 14 *Let $A \in \mathbb{R}^{d \times n}$ be a random matrix whose entries are independent standard normal random variables. Then, for every $t \geq 0$, with probability at least $1 - 2 \exp\left(-t^2/2\right)$, we have:*

$$\sqrt{n} - \sqrt{d} - t \leq \sigma_{min}(A) \leq \sigma_{max}(A) \leq \sqrt{n} + \sqrt{d} + t$$

Corollary 15 *Let $A \in \mathbb{R}^{d \times n}$ be a random matrix whose entries are independent standard normal random variables. For $n \geq 200\left(d + 2\log\left(\frac{2}{\delta}\right)\right)$, we have:*

$$0.9\sqrt{n} \leq \sigma_{min}(A) \leq \sigma_{max}(A) \leq 1.1\sqrt{n}$$

with probability at least $1 - \delta$

Lemma 16 *Let $Y_1, Y_2, \ldots, Y_n$ be iid d-dimensional random vectors such that $Y_i \sim \mathcal{N}(0, I) \, \forall i \in [n]$. Then, we have for any $c_2 \leq 1$:*

$$\mathcal{P}\left(\left|\left\{i : \|Y_i\| \geq d^{\frac{1}{2}} + 2d^{\frac{1}{4}}\left(\log\left(\frac{1}{c_2}\right) + \log\left(\mu^2 r\right)\right)^{\frac{1}{2}}\right\}\right| \geq \frac{3c_2 n}{2\mu^2 r}\right) \leq \beta$$

when $n \geq \frac{16\mu^2 r}{c_2}\log\left(\frac{1}{\beta}\right)$.

Proof Let $Y_1, \ldots, Y_n$ be iid random vectors such that $Y_i \sim \mathcal{N}(0, I) \, \forall i \in [n]$. From Lemma 12, we have that:

$$\mathcal{P}\left(\|Y_i\| \geq d^{1/2} + 2d^{\frac{1}{4}}\left(\log\left(\frac{1}{c_2}\right) + \log\left(\mu^2 r\right)\right)^{1/2}\right) \leq \frac{c_2}{\mu^2 r}$$

Let $p := \mathcal{P}\left(\|Y_i\| \geq d^{1/2} + 2d^{\frac{1}{4}}\left(\log\left(\frac{1}{c_2}\right) + \log\left(\mu^2 r\right)\right)^{1/2}\right)$. Consider random variables $Z_i \forall i \in [n]$ be defined such that $Z_i = \mathbb{I}\left[\|Y_i\| \geq d^{1/2} + 2d^{\frac{1}{4}}\left(\log\left(\frac{1}{c_2}\right) + \log\left(\mu^2 r\right)\right)^{1/2}\right]$. Note that Z_i satisfy the conditions of Theorem 13 with $\nu = np$ and $c = 1$. We can now bound the probability that $\sum_{i=1}^{n} Z_i$ is large by setting $t = \frac{nc_2}{16\mu^2 r}$:

$$\mathcal{P}\left(\sum_{i=1}^{n} Z_i \leq \frac{3nc_2}{2\mu^2 r}\right) \leq \mathcal{P}\left(\sum_{i=1}^{n} Z_i \leq \sum_{i=1}^{n} \mathbb{E}[Z_i] + \frac{nc_2}{2\mu^2 r}\right)$$

$$\leq \mathcal{P}\left(\sum_{i=1}^{n} Z_i \leq np + \sqrt{2npt} + t\right) \leq \exp(-t) \quad (6)$$

For our choice of n, the above inequality implies the lemma. ∎

Appendix B. Proof of Technical Lemmas

B.1. Proof of Lemma 4

Proof We prove the lemma through a series of inequalities:

$$\left\| (I - U_{\setminus S} U_{\setminus S}^\top) L_i^* \right\| \overset{(\zeta_1)}{\leq} \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left\| (I - U_{\setminus S} U_{\setminus S}^\top) L_{\setminus S}^* \right\| \overset{(\zeta_2)}{\leq} \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left\| (I - U_{\setminus S} (U_{\setminus S})^\top) M_{\setminus S}^* \right\|$$

$$\leq \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left\| (I - U^* (U^*)^\top) M_{\setminus S}^* \right\|,$$

where (ζ_1) holds from Lemma 11 and (ζ_2) follows by using the fact that $L_{\setminus S}^*$ can be obtained by setting a few columns of $M_{\setminus S}^*$ to 0. The last inequality follows from the fact that $U_{\setminus S}$ contains the top-r singular vectors of $M_{\setminus S}^*$. ∎

B.2. Proof of Lemma 5

Proof The lemma can be proved through the following set of inequalities:

$$\left\| \Sigma_{\setminus S}^{-1} U_{\setminus S}^\top L_i^* \right\| \overset{(\zeta_1)}{\leq} \left\| \Sigma_{\setminus S}^{-1} U_{\setminus S}^\top U_{\setminus S}^* \Sigma_{\setminus S}^* w \right\| \leq \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left\| \Sigma_{\setminus S}^{-1} U_{\setminus S}^\top L_{\setminus S}^* \right\|$$

$$\overset{(\zeta_2)}{\leq} \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left\| \Sigma_{\setminus S}^{-1} U_{\setminus S}^\top M_{\setminus S}^* \right\| \leq \frac{33}{32} \mu \sqrt{\frac{r}{n}},$$

where (ζ_1) holds with $\|w\| \leq \frac{33}{32} \mu \sqrt{\frac{r}{n}}$ from Lemma 11 and (ζ_2) follows from the fact that $L_{\setminus S}^*$ can be obtained from $M_{\setminus S}^*$ by setting some columns in $M_{\setminus S}^*$ to 0. ∎

B.3. Lemma 17

Lemma 17 *Consider the setting of Theorem 2. Let $S \subset [n]$ denote any subset of the columns of M^* such that $|S| \leq 3\rho n$. Furthermore, suppose that $\|N^*\|_F \leq \frac{\sigma_k(L^*)}{16}$ for some $k \in [r]$. Let $M_{\setminus S}^* (L_{\setminus S}^*, N_{\setminus S}^*, C_{\setminus S}^*)$ denote the matrix $M^*(L^*, N^*, C^*)$ projected onto the columns not in S. Let $U_{\setminus S} \Sigma_{\setminus S} V_{\setminus S}^\top$ denote the rank-k' SVD of $M_{\setminus S}^*$ for some $k' \leq k$. Then, we have:*

$$\# \left(i : \|E_i\| \geq 2\mu \sqrt{\frac{r}{n}} \right) \leq 2\rho n,$$

where $E = \Sigma_{\setminus S}^{-1} U_{\setminus S}^\top M^*$

Proof From Lemma 11, we get that $\sigma_{k'}(L_{\setminus S}^*) \geq \frac{31}{32} \sigma_{k'}(L^*)$. Along with Theorem 9, we conclude that $\sigma_{k'}(L_{\setminus S}^* + N_{\setminus S}^*) \geq \frac{7}{8} \sigma_{k'}(L^*)$. Since the column supports of $L^* + N^*$ and C^* are disjoint, we have that $\sigma_{k'}(L_{\setminus S}^* + N_{\setminus S}^*) \leq \sigma_{k'}(M_{\setminus S}^*)$. That is,

$$\sigma_{k'}(M_{\setminus S}^*) \geq \frac{7}{8} \sigma_{k'}(L^*). \tag{7}$$

We first bound the quantity $\left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} L_i^* \right\| \forall i \in [n]$:

$$
\begin{aligned}
\left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} L_i^* \right\| &\overset{(\zeta_1)}{\leq} \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} L_{\backslash S}^* \right\| \\
&\overset{(\zeta_2)}{\leq} \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left(\left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} (L_{\backslash S}^* + N_{\backslash S}^*) \right\| + \left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} N_{\backslash S}^* \right\| \right) \\
&\overset{(\zeta_3)}{\leq} \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left(\left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} M_{\backslash S}^* \right\| + \left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} N_{\backslash S}^* \right\| \right) \\
&\overset{(\zeta_4)}{\leq} \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left(1 + \frac{1}{14} \right) \leq \frac{9}{8} \mu \sqrt{\frac{r}{n}}, \quad (8)
\end{aligned}
$$

where ζ_1 follows from Lemma 11, ζ_2 using triangle inequality, ζ_3 using the above given fact that $\mathcal{SVD}_r(M_{\backslash S}^*) = U_{\backslash S} \Sigma_{\backslash S} (V_{\backslash S})^{\top}$, ζ_4 follows from using (7) with bound on $\|N^*\|_F$.

Suppose $\left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} (L_i^* + N_i^*) \right\| \geq 2\mu\sqrt{\frac{r}{n}}$ for some i. We now have:

$$
\left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} N_i^* \right\| \geq \frac{7}{8} \mu \sqrt{\frac{r}{n}}.
$$

Similarly, using (7), we get that:

$$
\|N_i^*\| \geq \frac{3}{4} \mu \sqrt{\frac{r}{n}} \sigma_{k'}(L^*).
$$

Let $\Gamma := \left\{ i : \left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} (L_i^* + N_i^*) \right\| \geq 2\mu\sqrt{\frac{r}{n}} \right\}$. Let N_Γ^* denote the matrix N^* restricted to the set Γ. Then we have,

$$
\sqrt{|\Gamma|} \frac{3}{4} \mu \sqrt{\frac{r}{n}} \sigma_{k'}(L^*) \leq \|N_\Gamma^*\|_F \leq \|N^*\|_F \leq \frac{1}{16} \sigma_{k'}(L^*).
$$

This implies that $|\Gamma| \leq \frac{n}{144\mu^2 r} \leq \rho n$. Also, by our assumption, $\alpha \leq \rho$, i.e., number of non-zero C_i^* is less than ρn. That is, the set $\left\{ i : \left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} C_i^* \right\| \geq 2\mu\sqrt{\frac{r}{n}} \right\}$ is of size at most ρn. Using the fact that support of C^* and $L^* + N^*$ is disjoint, we have that the set $\left\{ i : \left\| \Sigma_{\backslash S}^{-1} U_{\backslash S}^{\top} M_i^* \right\| \geq 2\mu\sqrt{\frac{r}{n}} \right\}$ is of size at most $2\rho n$. ∎

B.4. Proof of Lemma 18

Lemma 18 *Assume the setting of Lemma 17. Let $U_{\backslash S} \Sigma_{\backslash S} V_{\backslash S}^{\top}$ ($U_{\backslash S}^* \Sigma_{\backslash S}^* (V_{\backslash S}^*)^{\top}$) be the rank-k $\mathcal{SVD}$ of $M_{\backslash S}^*$ ($L_{\backslash S}^*$), then the following holds $\forall i, 1 \leq i \leq n$:*

$$
\left\| (I - U_{\backslash S} U_{\backslash S}^{\top})(L_i^* + N_i^*) \right\| \leq \frac{33}{32} \mu \sqrt{\frac{r}{n}} \left(\left\| (I - U_{1:k}^*(U_{1:k}^*)^{\top}) M_{\backslash S}^* \right\| + \|N^*\| \right) + \|N_i^*\|.
$$

Proof

$$\left\| (I - U_{\backslash S}U_{\backslash S}^{\top})(L_i^* + N_i^*) \right\| \le \|N_i^*\| + \left\| (I - U_{\backslash S}U_{\backslash S}^{\top})L_i^* \right\|$$

$$\overset{(\zeta_1)}{\le} \|N_i^*\| + \frac{33}{32}\mu\sqrt{\frac{r}{n}} \left\| (I - U_{\backslash S}U_{\backslash S}^{\top})L_{\backslash S}^* \right\|$$

$$\le \|N_i^*\| + \frac{33}{32}\mu\sqrt{\frac{r}{n}} \left(\left\| (I - U_{\backslash S}U_{\backslash S}^{\top})(L_{\backslash S}^* + N_{\backslash S}^*) \right\| + \left\| N_{\backslash S}^* \right\| \right)$$

$$\le \|N_i^*\| + \frac{33}{32}\mu\sqrt{\frac{r}{n}} \left(\left\| (I - U_{\backslash S}U_{\backslash S}^{\top})M_{\backslash S}^* \right\| + \|N^*\| \right)$$

$$\le \|N_i^*\| + \frac{33}{32}\mu\sqrt{\frac{r}{n}} \left(\left\| (I - U_{1:k}^*(U_{1:k}^*)^{\top})M_{\backslash S}^* \right\| + \|N^*\| \right)$$

where (ζ_1) follows from Lemma 11 and the fact that only $3\rho n$ columns are ever thresholded at any stage of the algorithm. The remaining inequalities follow using triangle inequality and $M_{\backslash S}^* = L_{\backslash S}^* + N_{\backslash S}^*$ along with $\mathcal{SVD}_r(M_{\backslash S}^*) = U_{\backslash S}\Sigma_{\backslash S}(V_{\backslash S})^{\top}$. ∎

B.5. Proof of Lemma 7

Proof Let $S^{(t)}$ denote the columns of C^* that are not thresholded in the tth inner iteration. For each $j \in S^{(t)}$, we know that $\left\| (\Sigma^{(t)})^{-1}(U^{(t)})^{\top}C_j^* \right\| \le 2\mu\sqrt{\frac{r}{n}}$ from our assumption on the termination of the algorithm. Furthermore, since C_j^* is not thresholded, we can associate a unique column i_j which is thresholded and $i_j \notin \text{Supp}(C^*)$ such that $\left\| I - U^{(t)}(U^{(t)})^{\top}M_{i_j}^* \right\| \ge \left\| I - U^{(t)}(U^{(t)})^{\top}M_j^* \right\|$. Let $y^{i,t} := (U^{(t)})^{-1}(\Sigma^{(t)})^{-1}M_i^*$ and $r^{i,t} := (I - U^{(t)}(U^{(t)})^{\top})M_i^*, \forall i$. Thus we have:

$$\left\| y^{j,t} \right\| \le 2\mu\sqrt{\frac{r}{n}}, \qquad \left\| r^{j,t} \right\| \le \left\| r^{i_j,t} \right\|.$$

Let $Q^{(t)}$ denote the columns of L^* that have been thresholded in the tth iteration. Furthermore, we definite the matrices $\widetilde{L}^{(t+1)} := L_{\backslash Q^{(t)}}^*$, $\widetilde{N}^{(t+1)} := N_{\backslash Q^{(t)}}^*$ and $\widetilde{C}^{(t+1)} := C_{S^{(t)}}^*$. Recall the notation, $P_{\perp}^U(M) = (I - UU^T)M$. We now have for any $t \ge 0$:

$$\left\| P_{\perp}^{U^{(t+1)}}(L^*) \right\|_F = \left\| P_{\perp}^{U^{(t+1)}}(\widetilde{L}^{(t+1)} + (L^* - \widetilde{L}^{(t+1)})) \right\|_F \tag{9}$$

$$\le \left\| P_{\perp}^{U^{(t+1)}}(\widetilde{L}^{(t+1)}) \right\|_F + \left\| P_{\perp}^{U^{(t+1)}}(L^* - \widetilde{L}^{(t+1)}) \right\|_F$$

$$\le \left\| P_{\perp}^{U^{(t+1)}}(\widetilde{L}^{(t+1)}) \right\|_F + \left(\sum_{i \in Q^{(t)}} \left\| P_{\perp}^{U^{(t+1)}}(L_i^*) \right\|^2 \right)^{\frac{1}{2}} \tag{10}$$

$$\overset{(\zeta_1)}{\le} \left\| P_{\perp}^{U^{(t+1)}}(\widetilde{L}^{(t+1)}) \right\|_F \left(1 + \sqrt{3\rho n}\frac{33}{32}\mu\sqrt{\frac{r}{n}} \right)$$

$$\overset{(\zeta_2)}{\le} \frac{5}{4} \left(\left\| P_{\perp}^{U^{(t+1)}}(\widetilde{L}^{(t+1)} + \widetilde{N}^{(t+1)}) \right\|_F + \left\| P_{\perp}^{U^{(t+1)}}(\widetilde{N}^{(t+1)}) \right\|_F \right)$$

$$\overset{(\zeta_3)}{\leq} \frac{5}{4}\left(\left\|P_\perp^{U^{(t+1)}}(\widetilde{L}^{(t+1)} + \widetilde{N}^{(t+1)} + \widetilde{C}^{(t+1)})\right\|_F + \left\|\widetilde{N}^{(t+1)}\right\|_F\right)$$

$$\overset{(\zeta_4)}{\leq} \frac{5}{4}\left(P_\perp^{U_{1:k}^*}(\widetilde{L}^{(t+1)} + \widetilde{N}^{(t+1)} + \widetilde{C}^{(t+1)})\right) + \left\|\widetilde{N}^{(t+1)}\right\|_F \overset{(\zeta_5)}{\leq} \frac{5}{4}\left(\left\|P_\perp^{U_{1:k}^*}(M^{(t+1)})\right\|_F + \|N^*\|_F\right) \tag{11}$$

where (ζ_1) follows from Lemma 11, (ζ_2) from triangle inequality and bound over ρ, (ζ_3) from Lemma 10, (ζ_4) from the properties of the SVD and (ζ_5) from Lemma 10.

We will now show that $\left\|P_\perp^{U_{1:k}^*}(M^{(t+1)})\right\|_F$ decreases at a geometric rate:

$$\left\|P_\perp^{U_{1:k}^*}(M^{(t+1)})\right\|_F \overset{(\zeta_6)}{\leq} \left\|P_\perp^{U_{1:k}^*}(\widetilde{L}^{(t+1)})\right\|_F + \|N^*\|_F + \left(\left\|P_\perp^{U_{1:k}^*}(\widetilde{C}^{(t+1)})\right\|_F\right)$$

$$\overset{(\zeta_7)}{\leq} \left\|P_\perp^{U_{1:k}^*}(L^*)\right\|_F + \|N^*\|_F + \left(\sum_{j \in S^{(t-1)}} \left\|P_\perp^{U_{1:k}^*}((\widetilde{C}^{(t+1)})_j)\right\|^2\right)$$

$$\overset{(\zeta_8)}{\leq} \left\|P_\perp^{U_{1:k}^*}(L^*)\right\|_F + \|N^*\|_F + \left(2\sum_{j \in S^{(t)}} \left\|P_\perp^{U_{1:k}^*}(U^{(t)}\Sigma^{(t)}y^{j,t}\right\|^2 + \left\|P_\perp^{U_{1:k}^*}(r^{j,t})\right\|^2\right)^{\frac{1}{2}}$$

$$\leq \left\|P_\perp^{U_{1:k}^*}(L^*)\right\|_F + \|N^*\|_F + \left(8\rho n\frac{\mu^2 r}{n}\left\|P_\perp^{U_{1:k}^*}(U^{(t)}\Sigma^{(t)})\right\|^2 + 2\sum_{j \in S^{(t)}}\left\|P_\perp^{U_{1:k}^*}(r^{j,t})\right\|^2\right)^{\frac{1}{2}}$$

$$\overset{(\zeta_9)}{\leq} \left\|P_\perp^{U_{1:k}^*}(L^*)\right\|_F + \|N^*\|_F + \left(\frac{1}{8}\left\|P_\perp^{U_{1:k}^*}(U^{(t)}\Sigma^{(t)})\right\|^2 + 2\sum_{j \in S^{(t)}}\left\|r^{i_j,t}\right\|^2\right)^{\frac{1}{2}}$$

$$\overset{(\zeta_{10})}{\leq} \left\|P_\perp^{U_{1:k}^*}(L^*)\right\|_F + \|N^*\|_F + \Bigg(\frac{1}{8}\underbrace{\left\|P_\perp^{U_{1:k}^*}(U^{(t)}\Sigma^{(t)})\right\|^2}_{Term\ 1}$$

$$+4\underbrace{\sum_{j \in S^{(t)}}\left\|N_{i_j}^*\right\|^2 + \left(\frac{33}{32}\right)^2\mu^2\frac{r}{n}\left(\left\|P_\perp^{U^{(t)}}(\widetilde{L}^{(t)} + \widetilde{N}^{(t)})\right\|\right)^2}_{Term\ 2}\Bigg)^{\frac{1}{2}}, \tag{12}$$

where (ζ_6) follows from triangle inequality, (ζ_7) follows from Lemma 10, (ζ_8) follows from triangle inequality and the fact that $(a + b)^2 \leq 2a^2 + 2b^2$, (ζ_9) from our previous observations about $y^{j,r}$ and $r^{j,t}$ and (ζ_{10}) from Lemma 17.

We will now proceed to bound Term 1 as follows:

$$\left\|P_\perp^{U_{1:k}^*}(U^{(t)}\Sigma^{(t)})\right\|_F^2 \overset{(\zeta_{11})}{\leq} \left\|P_\perp^{U_{1:k}^*}(M^{(t)})\right\|_F^2, \tag{13}$$

where (ζ_{11}) follows from considering the full SVD of $M^{(t)}$ and Lemma 10.

25

We now proceed to bound Term 2 as:

$$\sum_{j \in S^{(t)}} \left\| N^*_{i_j} \right\|^2 + \left(\frac{33}{32} \right)^2 \mu^2 \frac{r}{n} \left(\left\| P^{U^{(t)}}_\perp (\widetilde{L}^{(t)} + \widetilde{N}^{(t)}) \right\| \right)^2 \leq \|N^*\|_F^2 + \frac{9}{8} \rho n \frac{\mu^2 r}{n} \left\| P^{U^{(t)}}_\perp (M^{(t)}) \right\|_F^2$$

$$\leq \|N^*\|_F^2 + \frac{1}{32} \left\| P^{U^*_{1:k}}_\perp (M^{(t)}) \right\|_F^2, \quad (14)$$

where the first inequality follows from Lemma 10 and the second inequality from the fact that $U^{(t+1)}$ are top-k left singular vectors of $M^{(t)}$.

Using (12), (13), (14), we have:

$$\left\| P^{U^*_{1:k}}_\perp (M^{(t+1)}) \right\|_F$$

$$\leq \left\| P^{U^*_{1:k}}_\perp (L^*) \right\|_F + \|N^*\|_F + \left(\frac{1}{8} \left\| P^{U^*_{1:k}}_\perp (M^{(t)}) \right\|_F^2 + 4 \|N^*\|_F^2 + \frac{1}{8} \left\| P^{U^*_{1:k}}_\perp (M^{(t)}) \right\|_F^2 \right)^{\frac{1}{2}}$$

$$\leq \left\| P^{U^*_{1:k}}_\perp (L^*) \right\|_F + 3 \|N^*\|_F + \frac{1}{2} \left\| P^{U^*_{1:k}}_\perp (M^{(t)}) \right\|_F$$

where the first inequality follows from Lemma 10 and considering the full SVD of $M^{(t)}$ and the last inequality follows from the fact that $\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$.

By recursively applying the above inequality, we have:

$$\left\| P^{U^*_{1:k}}_\perp (M^{(T+1)}) \right\|_F \leq 2 \left\| P^{U^*_{1:k}}_\perp (L^*) \right\|_F + 6 \|N^*\|_F + \frac{\epsilon}{20n}.$$

Lemma now follows using (11) with the above equation. ∎

B.6. Lemma 19

Lemma 19 *Let $Y_1, Y_2, \ldots, Y_n$ be iid d-dimensional random vectors such that $Y_i \sim \mathcal{N}(0, I) \, \forall i \in [n]$. Then, we have for any $c_1 \leq 1$:*

$$\mathcal{P} \left(\exists v \in \mathbb{R}^d, \ \|v\| = 1 \ s.t \ \left| \left\{ i : \langle v, Y_i \rangle \geq 2 \left(\log \left(\frac{\mu^2 rd}{c_1} \right) \right)^{\frac{1}{2}} \right\} \right| \geq \frac{3c_1 n}{\mu^2 rd} \right) \leq \beta,$$

when $n \geq \frac{16\mu^2 rd}{c_1} \left[\log \left(\frac{1}{\beta} \right) + d \log (80d) \right]$.

Proof Let $v \in \mathbb{R}^d$ s.t $\|v\| = 1$. We define the set $\mathcal{S}_{v,\theta}$ as follows:

$$\mathcal{S}_{v,\theta} = \{u : u \in \mathbb{R}^d \land \|u\| = 1 \land \langle u, v \rangle \geq \cos(\theta)\}.$$

We now define the set $\mathcal{T}(v, \theta, \delta)$ as:

$$\mathcal{T}(v, \theta, \delta) = \{x : x \in \mathbb{R}^d \land \exists u \in \mathcal{S}_{v,\theta} \quad s.t \quad \langle u, x \rangle \geq \delta\}.$$

Now, let $y \sim \mathcal{N}(0, I)$. Using spherical symmetry of the Gaussian, w.l.o.g. $v = e_1$. We now define the complementary sets $\mathcal{Q}(\nu)$ and $\mathcal{R}(\nu)$ as:

$$\mathcal{Q}(\nu) = \{x : x \in \mathbb{R}^d \wedge x_1 < \nu\}, \qquad \mathcal{R}(\nu) = \{x : x \in \mathbb{R}^d \wedge x_1 \geq \nu\}.$$

We will now bound the probability that $y \in \mathcal{T}(v, \theta, \delta)$ for $\delta = 2\left(\log\left(\frac{\mu^2 rd}{c_1}\right)\right)^{1/2}$ and $\theta = \csc^{-1}\left(10(d-1)^{1/2}\right)$.

$$\mathcal{P}(y \in \mathcal{T}(e_1, \theta, \delta))$$

$$= \int_{\mathcal{T}(e_1,\theta,\delta)} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left(-\frac{\|y\|^2}{2}\right) dy$$

$$= \int_{\mathcal{T}(e_1,\theta,\delta)\cap\mathcal{Q}(\delta/\sqrt{2})} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left(-\frac{\|y\|^2}{2}\right) dy + \int_{\mathcal{T}(e_1,\theta,\delta)\cap\mathcal{R}(\delta/\sqrt{2})} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left(-\frac{\|y\|^2}{2}\right) dy$$

$$\leq \int_{\mathcal{R}(\delta/\sqrt{2})} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left(-\frac{\|y\|^2}{2}\right) dy + \int_{\mathcal{T}(e_1,\theta,\delta)\cap\mathcal{Q}(\delta/\sqrt{2})} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left(-\frac{\|y\|^2}{2}\right) dy$$

$$\overset{(\zeta_1)}{\leq} \frac{c_1}{\mu^2 rd} + \int_{\mathcal{T}(e_1,\theta,\delta)\cap\mathcal{Q}(\delta/\sqrt{2})} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left(-\frac{\|y\|^2}{2}\right) dy,$$

where (ζ_1) follows from the fact that for $t \geq 1$, $\int_t^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \leq \exp\left(-\frac{t^2}{2}\right)$.

We will use $\mathcal{M}(\theta, \delta, \gamma)$ to denote the set $\{z : z \in \mathbb{R}^{d-1} \wedge (\gamma, z) \in \mathcal{T}(e_1, \theta, \delta) \cap \mathcal{Q}(\delta/\sqrt{2})\}$. We can bound the second term as follows:

$$\int_{\mathcal{T}(e_1,\theta,\delta)\cap\mathcal{Q}(\delta/\sqrt{2})} \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left(-\frac{\|y\|^2}{2}\right) dy$$

$$\leq \int_{-\infty}^{\delta/\sqrt{2}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_1^2}{2}\right) \left(\int_{\mathcal{M}(\theta,\delta,y_1)} \frac{1}{\left(\sqrt{2\pi}\right)^{d-1}} \exp\left(-\frac{\|z\|^2}{2}\right) dz\right) dy_1. \quad (15)$$

Now, let $z \in \mathbb{R}^d$ be such that $z_1 = y_1 \wedge z_{2:d} \in \mathcal{M}(\theta, \delta, y_1)$ for some $y_1 \in [-\infty, \delta/\sqrt{2}]$. Therefore, $\exists w \in \mathcal{S}_{v,\theta}$ such that $\langle w, z \rangle \geq \delta$. We can decompose w into its components along v and orthogonal to it, $w = \cos(\theta')v + \sin(\theta')v^\perp$ for some unit vector $v^\perp$ orthogonal to v and some $\theta' \in [0, \theta]$. We know that $\langle w, z \rangle \geq \delta$ and that $\langle w, v \rangle \leq \delta/\sqrt{2}$. From these two inequalities and using the fact that $v = e_1$, we get:

$$\sin(\theta')\|z_{2:d}\| \geq \sin(\theta')\langle v^\perp, z \rangle \geq \delta - \cos(\theta')\langle v, z \rangle \geq \delta - \cos(\theta')\frac{\delta}{\sqrt{2}} \geq \left(1 - \frac{1}{\sqrt{2}}\right)\delta.$$

This allows us to lower bound the length of $z_{2:d}$ by $10(d-1)^{1/2}\left(1 - \frac{1}{\sqrt{2}}\right)\delta$. For our choice of δ and θ and using Lemma 12, we now get that the inner integration in equation 15 is atmost $\frac{c_1}{\mu^2 r d}$. Thus, we have the following bound on $\mathcal{P}\left(y \in \mathcal{T}\left(e_1, \theta, \delta\right)\right)$:

$$\mathcal{P}\left(y \in \mathcal{T}\left(e_1, \theta, \delta\right)\right) \leq \frac{2c_1}{\mu^2 r d}. \tag{16}$$

Let p be used to denote the value $\mathcal{P}\left(y \in \mathcal{T}\left(e_1, \theta, \delta\right)\right)$. Now, assume $Y_1, \dots, Y_n$ are iid random vectors with $Y_i \sim \mathcal{N}\left(0, I\right) \, \forall i \in [n]$. Now let Z_i be defined such that $Z_i = \mathbb{I}\left[Y_i \in \mathcal{T}\left(e_i, \theta, \delta\right)\right] \forall i \in [n]$. Note that Z_i is a Bernoulli random variable which is 1 with probability p. It can be seen that Z_i satisfy satisfy the conditions of 13 with $\nu = np$ and $c = 1$. Therefore, setting $t = \frac{nc_1}{16\mu^2 r d}$ in Theorem 13, we get:

$$\mathcal{P}\left(\sum_{i=1}^{n} Z_i \geq \frac{3nc_1}{\mu^2 r d}\right) \leq \mathcal{P}\left(\sum_{i=1}^{n} Z_i \geq np + \frac{nc_1}{\mu^2 r d}\right)$$

$$\leq \mathcal{P}\left(\sum_{i=1}^{n} Z_i \geq np + \sqrt{2\nu t} + t\right) \leq \exp\left(-t\right). \tag{17}$$

Now, consider the subset $\mathcal{K} := \{x : x \in \mathbb{R}^d \wedge |x_i| \leq 1 \forall i \in [d]\}$. Consider a partitioning of $\mathcal{K}$ into subsets $\mathcal{K}\left(\epsilon, j\right) = \{x : x \in \mathcal{K} \wedge \forall i \in [d] j_i \epsilon - 1 \leq x_i \leq (j_i + 1)\epsilon - 1\}$ where $j \in \mathcal{J}$ is an index for each of these subsets. Note that for any ϵ, at most $\left(\left\lceil \frac{2}{\epsilon} \right\rceil\right)^d$ such indices are required to ensure that $\mathcal{K} \subseteq \bigcup_{j \in \mathcal{J}} \mathcal{K}\left(\epsilon, j\right)$. Setting $\epsilon = \frac{1}{40d}$, we have for any two unit vectors v_1 and v_2 such that $v_1, v_2 \in \mathcal{K}\left(\epsilon, j\right)$ for some j, $\|v_1 - v_2\| \leq \frac{1}{40d^{1/2}}$. From this fact, it can be seen that Equation 17 holds for all unit vectors in $\mathcal{K}\left(\epsilon, j\right)$ with any unit vector $v \in \mathcal{K}\left(\epsilon, j\right)$. Therefore, we choose for each subset $\mathcal{K}\left(\epsilon, j\right)$, which contains a unit vector, a unit vector v and take an union bound over all such subsets $\mathcal{K}\left(\epsilon, j\right)$. After doing so we get the following bound:

$$\mathcal{P}\left(\exists v \in \mathbb{R}^d \wedge \|v\| = 1 \text{ s.t } \sum_{i=1}^{n} Z_i \geq \frac{3nc_1}{\mu^2 r d}\right) \leq (80d)^d \exp\left(-t\right) \overset{(\zeta_2)}{\leq} \beta, \tag{18}$$

where (ζ_2) follows from the conditions of the theorem. Thus, we have proved the theorem. ∎

B.7. Lemma 20

Lemma 20 *Assume the conditions of Theorem 3. Let $S \subset [n]$ denote any subset such that $|S| \leq \frac{1}{64\mu^2 r}$. Let $M_{\backslash S}^*(L_{\backslash S}^*, N_{\backslash S}^*, C_{\backslash S}^*)$ denote the matrices M^* (L^*, N^*, C^*) restricted to the columns not in S. Let $[U_{\backslash S}, \Sigma_{\backslash S}, V_{\backslash S}] = \mathcal{SVD}_{r+1}(M_{\backslash S}^*)$. Furthermore, let $\mathcal{E} = \{x : x = U_{\backslash S}\Sigma_{\backslash S}y$ for some $\|y\| \leq 2\mu\sqrt{r/n}\}$. Then $\forall i$, we have:*

$$\left\|\mathcal{P}_{U_{\backslash S}}\left(L_i^* + N_i^*\right) - \mathcal{P}_{\mathcal{E}}\left(L_i^* + N_i^*\right)\right\| \leq \frac{33}{32}\mu\sqrt{\frac{r}{n}}\|N^*\| + \left\|\mathcal{P}_{U_{\backslash S}}(N_i^*)\right\|,$$

where $\mathcal{P}_{U_{\backslash S}}\left(M\right) = U_{\backslash S}U_{\backslash S}^\top M$.

Proof Let $[U^*_{\backslash S}, \Sigma^*_{\backslash S}, V^*_{\backslash S}] = \mathcal{SVD}_r(L^*_{\backslash S})$. Using Lemma 11, $L^*_i = L^*_{\backslash S} V^*_{\backslash S} w_i$ for some $\|w_i\| \leq \frac{33}{32}\mu\sqrt{\frac{r}{n}}$. Now consider the vector $y_i := U_{\backslash S}\Sigma_{\backslash S} V^\top_{\backslash S} V^*_{\backslash S} w_i$. Note that $\left\|(\Sigma_{\backslash S})^{-1} U^\top_{\backslash S} y_i\right\| \leq \frac{33}{32}\mu\sqrt{\frac{r}{n}}$. Now, using definition of $\mathcal{P}_\mathcal{E}$:

$$\left\|\mathcal{P}_{U_{\backslash S}}\left(L^*_i + N^*_i\right) - \mathcal{P}_\mathcal{E}\left(L^*_i + N^*_i\right)\right\| \leq \left\|\mathcal{P}_{U_{\backslash S}}\left(L^*_i + N^*_i\right) - y_i\right\|$$

$$\leq \left\|\mathcal{P}_{U_{\backslash S}}(N^*_i)\right\| + \left\|\mathcal{P}_{U_{\backslash S}}\left(L^*_{\backslash S} V^*_{\backslash S} w_i\right) - U_{\backslash S}\Sigma_{\backslash S} V^\top_{\backslash S} V^*_{\backslash S} w_i\right\|$$

$$\overset{(\zeta_1)}{\leq} \left\|\mathcal{P}_{U_{\backslash S}}(N^*_i)\right\| + \frac{33}{32}\mu\sqrt{\frac{r}{n}}\left\|\mathcal{P}_{U_{\backslash S}}\left(L^*_{\backslash S}\right) V^*_{\backslash S} - \mathcal{P}_{U_{\backslash S}}\left(L^*_{\backslash S} + N^*_{\backslash S}\right) V^*_{\backslash S}\right\|$$

$$\leq \left\|\mathcal{P}_{U_{\backslash S}}(N^*_i)\right\| + \frac{33}{32}\mu\sqrt{\frac{r}{n}}\left\|\mathcal{P}_{U_{\backslash S}}\left(N^*_{\backslash S}\right)\right\|,$$

where (ζ_1) holds from Lemma 11 and the fact that $V^*_{\backslash S}$ has zeros in the rows corresponding to the support of C^*. ∎

B.8. Proof of Lemma 8

Proof Note that under the conditions of Theorem 3, the following three conditions hold with probability at least $1 - \delta$.

$$\#i\left(\|N^*_i\| \geq \sigma\left(\sqrt{d} + 2d^{\frac{1}{4}}\sqrt{\log\left(\frac{\mu^2 r}{c_2}\right)}\right)\right) \leq \frac{3c_2 n}{\mu^2 r}$$

$$\forall \|v\| = 1 \qquad \#i\left(|\langle v, N^*_i\rangle| \geq 2\sigma\sqrt{\log\left(\frac{\mu^2 r^2 d}{c_1}\right)}\right) \leq \frac{6c_1 n}{\mu^2 r^2 d}$$

$$0.9\sigma\sqrt{n} \leq \|N^*\| \leq 1.1\sigma\sqrt{n}$$

Also, for $c_1 = \frac{1}{12288}$ and $c_2 = \frac{1}{1536}$, the invariant implies at most $\frac{n}{512\mu^2 r}$ clean data points are thresholded at any stage. This allows us to apply Lemma 20 in the subsequent steps.

We will prove the lemma by induction on the number of thresholding steps completed so far. Let t denote the number of thresholding steps executed so far.

Base Case (t = 0): The invariant trivially holds before any data points have been thresholded.

Induction Step (t = k + 1): Assuming that the invariant holds after the kth thresholding step, we now have two cases for the $(k + 1)^{th}$ thresholding step:

Case 1: Thresholding with respect to ζ_1. For a column $i \notin \text{Supp}(C^*)$ thresholded with respect to ζ_1, we have from Lemma 20 that:

$$\|N^*_i\| \geq \zeta_1 - \frac{33}{32}\mu\sqrt{\frac{r}{n}}\|N^*\| = \sigma\left(\frac{5}{4}\mu\sqrt{r} + d^{\frac{1}{2}} + 2d^{\frac{1}{4}}\sqrt{\log\left(\frac{\mu^2 r}{c_2}\right)}\right) - \frac{5}{4}\sigma\mu\sqrt{r}$$

$$\geq \sigma\left(d^{\frac{1}{2}} + 2d^{\frac{1}{4}}\sqrt{\log\left(\frac{\mu^2 r}{c_2}\right)}\right). \quad (19)$$

From our choice of ζ_1, there are only $\frac{n}{1024\mu^2 r}$ clean points which satisfy 19.

Case 2: Thresholding with respect to ζ_2. For a column $i \notin \text{Supp}\,(C^*)$, it can only be thresholded if the number of columns to be thresholded exceeds $\frac{24c_1 n}{\mu^2 rd}$. Note that:

$$\|\mathcal{P}_{U^{(k+1)}}(N_i^*)\| \geq \zeta_2 - \frac{5}{4}\sigma\mu\sqrt{r} = 2\sigma\sqrt{2r}\sqrt{\log\left(\frac{\mu^2 r^2 d}{c_1}\right)}.$$

Note that $U^{(k+1)}$ is at most a rank $2r$ subspace. Therefore, $\exists j$ such that:

$$\left|\left\langle N_i^*, U_j^{(k+1)} \right\rangle\right| \geq 2\sigma\sqrt{\log\left(\frac{\mu^2 r^2 d}{c_1}\right)}. \tag{20}$$

Taking a union bound over all $j \in [r+1]$ and using Lemma 19 for both positive and negative inner product values, we get that at most $\frac{12c_1 n}{\mu^2 rd}$ clean points satisfy 20. Since, we threshold at least $\frac{24c_1 n}{\mu^2 rd}$ points, at least half of them must be outliers and hence the invariant holds in the next iteration.

$\blacksquare$

Appendix C. Gaussian Noise: Proof of Theorem 3

Proof We will prove the theorem for $c_1 = \frac{1}{12288}$ and $c_2 = \frac{1}{1536}$. For our choices of c_1 and c_2 and n, we have that:

$$\#i\left(\|N_i^*\| \geq \sigma\left(\sqrt{d} + 2d^{\frac{1}{4}}\sqrt{\log\left(\frac{\mu^2 r}{c_2}\right)}\right)\right) \leq \frac{3c_2 n}{\mu^2 r}$$

$$\forall \|v\| = 1 \qquad \#i\left(|\langle v, N_i^*\rangle| \geq 2\sigma\sqrt{\log\left(\frac{\mu^2 r^2 d}{c_1}\right)}\right) \leq \frac{6c_1 n}{\mu^2 r^2 d}$$

$$0.9\sigma\sqrt{n} \leq \|N^*\| \leq 1.1\sigma\sqrt{n}$$

with probability at least $1 - \delta$ from Lemmas 19, 16 and Corollary 15 along with our choice of n.

From Lemma 8, we know that Invariant 1 holds at the termination of the algorithm. Therefore, at most $\frac{n}{512\mu^2 r}$ inliers are removed (The number of inliers removed is at most $\alpha n + \frac{n}{1024\mu^2 r}$).

Suppose the algorithm terminated in the Tth iteration. Let $M := M^* - C^{(T)}$. We will start by making a few observations. The algorithm terminates when no data point is thresholded. Let $[U, \Sigma, V] = \mathcal{SVD}_{r+1}(M)$. Furthermore, define $\mathcal{E} = \{x : x = U\Sigma y \text{ for some } \|y\| \leq 2\mu\sqrt{r/n}\}$. Now, define set A as:

$$A := \left\{i : \|\mathcal{P}_U((M)_i) - \mathcal{P}_{\mathcal{E}}((M)_i)\| \geq \sigma\sqrt{2r}\left(\frac{5}{4}\mu + 2\log^{\frac{1}{2}}\left(\frac{\mu^2 r^2 d}{c_1}\right)\right)\right\},$$

and B as:

$$B := \left\{i : \|\mathcal{P}_U(M_i) - \mathcal{P}_{\mathcal{E}}(M_i)\| \geq \sigma\left(d^{\frac{1}{2}} + 2d^{\frac{1}{4}}\left(\log^{\frac{1}{2}}\left(\frac{\mu^2 r}{c_2}\right)\right)\right) + \sigma\frac{5}{4}\mu\sqrt{r}\right\}.$$

Recall that we will threshold the columns in A if $|A| \geq \frac{24c_1 n}{\mu^2 rd}$ and the columns in B if B is not empty. Therefore, we know that:

$$\forall i \in [n] \, \|\mathcal{P}_U(M_i) - \mathcal{P}_{\mathcal{E}}(M_i)\| \leq \sigma \left(d^{\frac{1}{2}} + 2d^{\frac{1}{4}} \left(\log^{\frac{1}{2}} \left(\frac{\mu^2 r}{c_2} \right) \right) \right) + \sigma \frac{5}{4} \mu \sqrt{r}, \, |A| \leq \frac{24c_1 n}{\mu^2 rd}. \tag{21}$$

Let S denote the set of data points that have been thresholded when the algorithm terminated, i.e $S = \mathcal{CS}^{(T)}$. Let $L = L^*_{\backslash S}$, $N = N^*_{\backslash S}$ and $C = C^*_{\backslash S}$. Additionally, let $[U^*_{\backslash S}, \Sigma^*_{\backslash S}, V^*_{\backslash S}] = \mathcal{SVD}(L)$. Similar to the proofs of Theorems 1 and 2, we start as follows:

$$\left\| P_\perp^{U_{1:r}}(L^*) \right\| \leq \left(\left\| P_\perp^{U_{1:r}}(L) \right\|^2 + \sum_{i \in S} \left\| P_\perp^{U_{1:r}}(U^*_{\backslash S} \Sigma^*_{\backslash S} w_i) \right\|^2 \right)^{\frac{1}{2}}$$

$$\overset{\zeta_1}{\leq} \left(\left\| P_\perp^{U_{1:r}}(L) \right\|^2 + \sum_{i \in S} \frac{9}{8} \mu^2 \frac{r}{n} \left\| P_\perp^{U_{1:r}}(U^*_{\backslash S} \Sigma^*_{\backslash S}) \right\|^2 \right)^{\frac{1}{2}} \tag{22}$$

$$\overset{\zeta_2}{\leq} \left(\left\| P_\perp^{U_{1:r}}(L) \right\|^2 + 3\rho n \frac{9}{8} \mu^2 \frac{r}{n} \left\| P_\perp^{U_{1:r}}(L) \right\|^2 \right)^{\frac{1}{2}}$$

$$\leq \frac{5}{4} \left\| P_\perp^{U_{1:r}}(L) \right\| \leq \frac{5}{4} \left(\|N\| + \left\| P_\perp^{U_{1:r}}(L + N) \right\| \right)$$

$$\overset{\zeta_3}{\leq} \frac{5}{4} \left(\|N^*\| + \left\| P_\perp^{U_{1:r}}(M) \right\| \right) = \frac{5}{4} \left(\|N^*\| + \left\| P_\perp^{U_{1:r}}(\mathcal{P}_U(M)) \right\| \right), \tag{23}$$

ζ_1 follows using Lemma 11, ζ_2 follows using $|S| \leq 2\rho n$, ζ_3 follows using Lemma 10 where the last equality follows from the fact that U consists of the top $r + 1$ singular vectors of M.

Now, let Y be an orthogonal basis of the subspace spanned by $\mathcal{P}_U(L)$. Note that the subspace spanned by Y is at most rank-r. Let O denote the set of corrupted columns that haven't been thresholded at the termination of the algorithm and let $O_l := O \cap A$ and $O_s := O \backslash O_l$. We can now bound $\left\| P_\perp^{U_{1:r}}(\mathcal{P}_U(M)) \right\|$ as follows:

$$\left\| P_\perp^{U_{1:r}}(\mathcal{P}_U(M)) \right\| \leq \left\| P_\perp^Y(\mathcal{P}_U(L + N + C)) \right\| \leq \left\| P_\perp^Y(\mathcal{P}_U(N + C)) \right\| \leq \|N\| + \left\| P_\perp^Y(\mathcal{P}_U(C)) \right\|$$

$$\leq \|N^*\| + \left(\underbrace{\sum_{i \in O_l} \left\| P_\perp^Y(\mathcal{P}_U(C_i)) \right\|^2}_{\text{Term 1}} + \underbrace{\sum_{j \in O_s} \left\| P_\perp^Y(\mathcal{P}_U(C_j)) \right\|^2}_{\text{Term 2}} \right)^{\frac{1}{2}}, \tag{24}$$

where first inequality follows from the fact that $U_{1:r}$ are top singular vectors of $\mathcal{P}_U(M)$ and second inequality follows from definition of Y.

We can now bound Term 1 as follows:

$$\sum_{i \in O_l} \left\| P_\perp^Y(\mathcal{P}_U(C_i)) \right\|^2 \overset{(\zeta_1)}{\leq} 2 \sum_{i \in O_l} \left\| P_\perp^Y(\mathcal{P}_{\mathcal{E}}(C_i)) \right\|^2 + \left\| P_\perp^Y((\mathcal{P}_U(C_i) - \mathcal{P}_{\mathcal{E}}(C_i))) \right\|^2$$

$$\leq 2 \sum_{i \in O_l} \frac{4\mu^2 r}{n} \left\| P_\perp^Y(U\Sigma) \right\|^2 + \|\mathcal{P}_U(C_i) - \mathcal{P}_{\mathcal{E}}(C_i)\|^2 \text{ (From Definition of } \mathcal{E})$$

31

$$\overset{(\varsigma_2)}{\le} \frac{48c_1 n}{\mu^2 r d} \left(\frac{4\mu^2 r}{n} \left\| P_\perp^Y(U\Sigma) \right\|^2 + \sigma^2 \left(4\mu^2 r + 2\left(\sqrt{d} + 2d^{\frac{1}{4}} \sqrt{\log\left(\frac{\mu^2 r}{c_2} \right)} \right)^2 \right) \right)$$

$$\overset{(\varsigma_3)}{\le} \frac{48c_1 n}{\mu^2 r d} \left(\frac{4\mu^2 r}{n} \left\| P_\perp^Y(U\Sigma) \right\|^2 + \sigma^2 \left(4\mu^2 r + 2\left(2d + 8d^{\frac{1}{2}} \log\left(\frac{\mu^2 r}{c_2} \right) \right) \right) \right)$$

$$\le \frac{48c_1 n}{\mu^2 r d} \left(\frac{4\mu^2 r}{n} \left\| P_\perp^Y(U\Sigma) \right\|^2 + \sigma^2 \left(4\mu^2 r + 4d + 16d^{\frac{1}{2}} \log\left(\frac{\mu^2 r}{c_2} \right) \right) \right)$$

$$\overset{(\varsigma_4)}{\le} \frac{48c_1 n}{\mu^2 r d} \left(\frac{4\mu^2 r}{n} \left\| P_\perp^Y(U\Sigma) \right\|^2 + \sigma^2 \left(4\mu^2 r + 4d + 16d^{\frac{1}{2}} \log(\mu^2 r) + 16d^{\frac{1}{2}} \log(1536) \right) \right)$$

$$\overset{(\varsigma_5)}{\le} \frac{48c_1 n}{\mu^2 r d} \left(\frac{4\mu^2 r}{n} \left\| P_\perp^Y(U\Sigma) \right\|^2 + \sigma^2 \left(24\mu^2 r d + 120d \right) \right)$$

$$\le \frac{\left\| P_\perp^Y(M) \right\|^2}{32} + \frac{48c_1 n}{\mu^2 r d} \sigma^2 \left(144\mu^2 r d \right) \text{ (From Lemma 10)}$$

$$\le \frac{\left\| P_\perp^Y(M) \right\|^2}{32} + 0.563\sigma^2 n$$

where (ζ_1) and (ζ_3) follow from $(a+b)^2 \le 2a^2 + 2b^2$, ζ_2 follows from (21). (ζ_4) and (ζ_5) follow from $\log(x) \le x$ and $\sqrt{x} \le x$ for all $x \ge 1$.

We now bound Term 2 as:

$$\sum_{j \in O_s} \left\| P_\perp^Y(\mathcal{P}_U(C_j)) \right\|^2 \overset{(\zeta_6)}{\le} 2 \sum_{j \in O_s} \left\| P_\perp^Y(\mathcal{P}_\mathcal{E}(C_j)) \right\|^2 + \left\| P_\perp^Y((\mathcal{P}_U(C_j) - \mathcal{P}_\mathcal{E}(C_j))) \right\|^2$$

$$\le 2 \sum_{i \in O_s} \frac{4\mu^2 r}{n} \left\| P_\perp^Y(U\Sigma) \right\|^2 + \left\| \mathcal{P}_U(C_j) - \mathcal{P}_\mathcal{E}(C_j) \right\|^2 \text{ (Definition of } \mathcal{E})$$

$$\overset{(\zeta_7)}{\le} 2\alpha n \left(\frac{4\mu^2 r}{n} \left\| P_\perp^Y(U\Sigma) \right\|^2 + 2\sigma^2 \left(4\mu^2 r + 8r \log\left(\frac{\mu^2 r^2 d}{c_1} \right) \right) \right)$$

$$\overset{(\zeta_8)}{\le} \frac{\left\| P_\perp^Y(M) \right\|^2}{32} + 4\alpha n \sigma^2 \left(4\mu^2 r + 8r \log(\mu^2) + 16r \log(r) + 8r \log(d) + 8r \log(12288) \right)$$

$$\overset{(\zeta_9)}{\le} \frac{\left\| P_\perp^Y(M) \right\|^2}{32} + 4\alpha n \sigma^2 \left(12\mu^2 r + 24r \log(d) + 8r \log(12288) \right)$$

$$\overset{(\zeta_{10})}{\le} \frac{\left\| P_\perp^Y(M) \right\|^2}{32} + 0.435\sigma^2 n \log(d),$$

where (ζ_6) and (ζ_7) follow from $(a+b)^2 \le 2a^2 + 2b^2$ along with (21). (ζ_8) follows from Lemma 10, (ζ_9) follows from the fact that $r \le d$ and $\log(x) \le x$ and (ζ_{10}) follows from assuming $\log(d) \ge 1$ and $\mu \ge 1$.

From our bounds on Term 1 and Term 2 and (24), we have:

$$\left\| P_\perp^{U_{1:r}}(\mathcal{P}_U(M)) \right\| \le \left\| P_\perp^Y(\mathcal{P}_U(M)) \right\| \le \frac{4}{3}\sigma\sqrt{n \log(d)}.$$

Theorem now follows by using the above observation with (23):

$$\left\| P_\perp^{U_{1:r}}(L^*) \right\| \le \frac{6}{4}\sigma\sqrt{n} + \frac{5}{3}\sigma\sqrt{n \log(d)} \le 4\sigma\sqrt{n \log(d)}.$$

Appendix D. TORP-BIN

In this section, we propose an improvement to Algorithm 2 which uses binary search instead of a linear scan in the outer iteration. This improves the running time on Algorithm 2 by almost a factor of r.

D.1. Algorithm

In this section, we present our algorithm (See Algorithm 4) for **OR-PCAN** which improves the running time of Algorithm 2 by almost a factor of r. The main insight is that inner iteration of Algorithm 2 is independent of the value of k in the outer iteration save for the rank of the projection. In Algorithm 4, we use binary search on k instead of a linear scan which reduces the number of outer iterations from $O(r)$ to $O(\log r)$.

Algorithm 4 Binary search based TORP (**TORP-BIN**)

1: **Input**: Corrupted matrix $M^* \in \mathbb{R}^{d \times n}$, Target rank r, Expressivity parameter η, Threshold fraction ρ, Number of inner iterations T

2: $minK \leftarrow 1, maxK \leftarrow r$

3: **while** $minK \leq maxK$ **do**

4: $\quad k \leftarrow \lfloor \frac{minK+maxK}{2} \rfloor$

5: $\quad C^{(0)} \leftarrow 0, \tau \leftarrow false$

6: $\quad$ **for** $t = 0$ to $t = T$ **do**

7: $\quad\quad [U^{(t)}, \Sigma^{(t)}, V^{(t)}] \leftarrow \mathcal{SVD}_k \left(M^* - C^{(t)} \right), L^{(t)} \leftarrow U^{(t)} \Sigma^{(t)} (V^{(t)})^\top$ $\left.\begin{array}{l} \\ \\ \\ \\ \end{array}\right\}$ Projection onto space of low rank matrices

8: $\quad\quad E \leftarrow (\Sigma^{(t)})^{-1}(U^{(t)})^\top M^*$ /* *Compute Incoherence* */ $\left.\begin{array}{l} \\ \\ \\ \\ \end{array}\right\}$

9: $\quad\quad R \leftarrow (I - U^{(t)}(U^{(t)})^\top)M^*$ /* *Compute residual* */

10: $\quad\quad \mathcal{CS}^{(t+1)} \leftarrow \mathcal{HT}_{2\rho}(M^*, E) \cup \mathcal{HT}_\rho(M^*, R)$ Projection onto space of column sparse matrices

11: $\quad\quad C^{(t+1)} \leftarrow M^*_{\mathcal{CS}^{(t+1)}}$

12: $\quad\quad n_{\text{thres}} \leftarrow |\{i : \|E_i\| \geq \eta\}|$ /* *Compute high incoherence points* */

13: $\quad\quad \tau \leftarrow \tau \vee (n_{\text{thres}} \geq 2\rho n)$ /* *Check termination conditions* */

14: $\quad$ **end for**

15: $\quad$ **if** τ **then**

16: $\quad\quad maxK \leftarrow k - 1$

17: $\quad$ **else**

18: $\quad\quad minK \leftarrow k + 1$

19: $\quad\quad [U, \Sigma, V] \leftarrow \mathcal{SVD}_k \left(M^* - C^{(T+1)} \right)$

20: $\quad$ **end if**

21: **end while**

22: **Return:** U

D.2. Analysis

In this section, we will state and prove a theoretical guarantee for Algorithm 4.

Theorem 21 *Assume the conditions of Theorem 2. Furthermore, suppose that $\|N^*\|_F \leq \frac{\sigma_k(L^*)}{16}$ for some k. Then, for all $\alpha \leq \frac{1}{128\mu^2 r}$, Algorithm 4 when run with parameters $\rho = \frac{1}{128\mu^2 r}$, $\eta = 2\mu\sqrt{\frac{r}{n}}$ and $T = \log \frac{20n\sigma_1(M)}{\epsilon}$, returns a subspace U which satisfies:*

$$\left\|(I - UU^\top)L^*\right\|_F \leq 3\left\|(I - U_{1:k}^*(U_{1:k}^*)^\top)L^*\right\|_F + 9\|N^*\|_F + \frac{\epsilon}{10n}.$$

Proof We will begin by bounding the running time of the algorithm. Note that because of the binary search, the algorithm will run for at most $O(\log r)$ outer iterations.

Let t denote the number of outer iterations of the algorithm. Let the value of $k(maxK, minK)$ in iteration t be denoted by $k^{(t)}$ ($maxK^{(t)}, minK^{(t)}$). We will first prove the claim that at any point in the running of the algorithm, $maxK \geq k$. We will prove the claim via induction on the number of iterations t:

Base Case: $t = 0$ The base case is trivially true as $maxK = r$.

Induction Step: $t = l+1$ Assume that the claim remains true at iteration $t = l$. In the $(l+1)^{th}$ iteration, we assume two cases:

Case 1: The inner iteration finishes with $\tau = false$. In this case, $maxK$ is not updated. So, the claim remains true for $t = (l+1)$

Case 2: The inner iteration finishes with $\tau = true$. In this case, $k^{(t)} > k$ (From Lemma 17 and the termination condition of the inner iteration.). In this iteration, $maxK$ is updated to $k^{(t)} - 1 \geq k$. Thus, the claim remains true.

Therefore, at termination of the algorithm, we have $maxK \geq k$. Suppose that the algorithm terminated after iteration T. Note that $minK^{(t)} \leq k \leq maxK^{(t)} \forall 0 \leq t \leq T$. Therefore, we have $minK^{(T+1)} = maxK^{(T+1)} + 1$. For this to happen, the inner iteration must have run with $k^{(T')} = maxK^{(T+1)}$ with $\tau = false$ for some iteration T' and also that this is the last such successful iteration as $minK$ is not updated after iteration T'. Therefore, the algorithm returns the subspace corresponding to $k^{(T')} = maxK^{(T+1)} \geq k$. Since the inner iteration is successful for iteration T', the Theorem is true from the application of Lemma 7 and noting that $k^{T'} \geq k$. $\blacksquare$

Appendix E. Fast Projection Operator

In this section, we will describe a fast algorithm to compute the projection operator onto the ellipsoid in Algorithm 2. Formally, we are provided an orthogonal basis $U \in \mathbb{R}^{d \times r}$, a positive diagonal matrix Σ, a bound b and a vector x. Let $\mathcal{E} = \{y : y = U\Sigma z \text{ for some } \|z\| \leq b\}$. The goal is to compute the projection of the vector x onto the set $\mathcal{E}$.

E.1. Algorithm

In this section, we present our algorithm (Algorithm 5) to compute the projection onto the set $\mathcal{E}$. We show that the projection operation boils down to an univariate optimization problem on a monotone function. We then perform binary search on an interval in which the solution is guaranteed to lie.

Algorithm 5 $w = \textbf{FAST-PR}(U, \Sigma, b, x, \epsilon)$

1: **Input**: Orthogonal Basis $U \in \mathbb{R}^{d \times r}$, Positive Diagonal Matrix Σ, Bound b, Projection Vector x, Accuracy Parameter ϵ
2: $\sigma_{min} = \min_{i \in [r]}(\Sigma_{i,i})$, $\sigma_{max} = \max_{i \in [r]}(\Sigma_{i,i})$
3: $y \leftarrow \Sigma U^\top x$
4: $\lambda_{min} = 0$, $\lambda_{max} = \frac{\|y\|}{b}$
5: $T \leftarrow \log\left(\frac{\lambda_{max}\sqrt{r}\|x\|}{\sigma_{min}^2 \epsilon}\right)$
6: **for** Iteration $t = 0$ to $t = T$ **do**
7: $\quad \lambda^{(t)} \leftarrow \frac{\lambda_{min} + \lambda_{max}}{2}$
8: $\quad z^{(t)} \leftarrow (\lambda I + \Sigma^2)^{-1} y$
9: $\quad$ **if** $\left\| z^{(t)} \right\| \leq b$ **then**
10: $\quad\quad \lambda_{max} \leftarrow \lambda^{(t)}$
11: $\quad$ **else**
12: $\quad\quad \lambda_{min} \leftarrow \lambda^{(t)}$
13: $\quad$ **end if**
14: **end for**
15: **Return:** $U\Sigma z^{(T)}$

E.2. Analysis

Theorem 22 *Let $U \in \mathbb{R}^{d \times r}$ be an orthonormal matrix and $\Sigma \in \mathbb{R}^{r \times r}$ be a positive diagonal matrix. Then, for any $b \geq 0$, $x \in \mathbb{R}^d$ and ϵ, the vector w returned by Algorithm 5 satisfies:*

$$\|w - \mathcal{P}_{\mathcal{E}}(x)\| \leq \epsilon$$

where $\mathcal{E} := \{y : y = U\Sigma z \text{ for some } \|z\| \leq b\}$

Proof We first define the convex optimization problem corresponding to the projection operator $\mathcal{P}_{\mathcal{E}}$. Then, we have:

$$\mathcal{P}_{\mathcal{E}}(x) = \arg \min_y \|x - y\| \ \ s.t \ y \in \mathcal{E}$$

Since, $y \in \mathcal{E}$, a solution to the above optimization problem is equivalent to:

$$\mathcal{P}_{\mathcal{E}}(x) = \arg \min_z \|x - U\Sigma z\|^2 \ \ s.t \ \|z\|^2 \leq b^2 \tag{25}$$

Note that both the constraint and the objective function are convex. Therefore, we can introduce a KKT multiplier $\lambda \geq 0$ and writing down the stationarity conditions of 25, we get:

$$2\Sigma^2 z + 2\lambda z = 2\Sigma U^\top x \implies z = \left(\Sigma^2 + \lambda I\right)^{-1} \Sigma U^\top x$$

Now, we just need to ensure that $\left\| \left(\Sigma^2 + \lambda I\right)^{-1} \Sigma U^\top x \right\| \leq b$. Let

$f(\lambda) = \left\| \left(\Sigma^2 + \lambda I\right)^{-1} \Sigma U^\top x \right\|$ for $\lambda \geq 0$. Let λ^* be the solution to $f(\lambda) = \min(\left\| \Sigma^{-1} U^\top x \right\|, b)$. We will first prove that at any point in the running of the algorithm $\lambda_{max} \geq \lambda^*$ and $\lambda_{min} \leq \lambda^*$. We prove the claim by induction on the number of iterations t:

Base Case $t = 0$: Since $\lambda_{min} = 0$, the lower bound holds trivially. That $\lambda_{max} \geq \lambda^*$ can be proved as follows:

$$f(\lambda_{max}) = \left\|\left(\Sigma^2 + \lambda_{max}I\right)^{-1}\Sigma U^\top x\right\| \leq \min\left(\left\|\Sigma^{-1}U^\top x\right\|, \frac{\left\|\Sigma U^\top x\right\|}{\lambda_{max}}\right)$$
$$\leq \min\left(\left\|\Sigma^{-1}U^\top x\right\|, b\right).$$

Since, f is a monotonically decreasing function, the claim holds true in the base case.

Induction Step $t = (k+1)$: Assume that the claim holds till $t = k$. We have two cases for iteration $k + 1$:

Case 1: $\lambda_{max} \leftarrow \lambda^{(t+1)}$. In this case, $\lambda_{min} \leq \lambda^*$ still holds from the inductive hypothesis. For λ_{max}, we have:

$$f(\lambda^{(t+1)}) = \left\|\left(\Sigma^2 + \lambda^{(t+1)}I\right)^{-1}\Sigma U^\top x\right\| \leq \min\left(\left\|\Sigma^{-1}U^\top x\right\|, f(\lambda^{(t+1)})\right)$$
$$\leq \min\left(\left\|\Sigma^{-1}U^\top x\right\|, b\right),$$

where the last inequality holds from the fact that λ_{max} was updated in this iteration. From the monotonicity of f, the induction hypothesis holds in this iteration.

Case 2: $\lambda_{min} \leftarrow \lambda^{(t+1)}$. In this case, $\lambda_{max} \geq \lambda^*$ by the inductive hypothesis. In this case, we have:

$$f(\lambda^{(t+1)}) \geq b \geq f(\lambda^*).$$

From the monotonicity of f, the induction hypothesis holds in this iteration.

Note that $(\lambda_{max} - \lambda_{min})$ is halved at each iteration. Therefore, at the termination of the algorithm, we have $(\lambda_{max} - \lambda_{min}) \leq \frac{\sigma_{min}^2 \epsilon}{\sqrt{r}\|x\|}$. From our claim, this implies that $\left|\lambda^* - \lambda^{(T)}\right| \leq \frac{\sigma_{min}^2 \epsilon}{\sqrt{r}\|x\|}$. Note that we can write $\mathcal{P}_\mathcal{E}(x) = U\Sigma^2(\lambda^* I + \Sigma^2)^{-1}U^\top x$. Note that $\|\mathcal{P}_\mathcal{E}(x) - w\| = \left\|U^\top(\mathcal{P}_\mathcal{E}(x) - w)\right\|$. We will now bound the element-wise difference between $U^\top \mathcal{P}_\mathcal{E}(x)$ and $U^\top w$:

$$\left|e_i^\top \Sigma^2((\lambda^* I + \Sigma^2)^{-1} - (\lambda^{(T)}I + \Sigma^2)^{-1})U^\top x\right| \leq \|x\|\left|\sigma_i^2\left(\frac{1}{\lambda^* + \sigma_i^2} - \frac{1}{\lambda^{(T)} + \sigma_i^2}\right)\right|$$
$$\leq \|x\|\,\sigma_i^2\left|\frac{\lambda^{(T)} - \lambda^*}{(\lambda^* + \sigma_i^2)(\lambda^{(T)} + \sigma_i^2)}\right| \leq \|x\|\left|\frac{\lambda^{(T)} - \lambda^*}{\sigma_i^2}\right| \leq \frac{\epsilon}{\sqrt{r}}.$$

By applying the element-wise bound to $\Sigma^2((\lambda^* I + \Sigma^2)^{-1} - (\lambda^{(T)}I + \Sigma^2)^{-1})U^\top x$, we have:

$$\|\mathcal{P}_\mathcal{E}(x) - w\| \leq \sqrt{r}\frac{\epsilon}{\sqrt{r}} \leq \epsilon.$$

$\blacksquare$