Robust PCA via Outlier Pursuit

Huan Xu, Constantine Caramanis, *Member*, and Sujay Sanghavi, *Member*

***Abstract*— Singular Value Decomposition (and Principal Component Analysis) is one of the most widely used techniques for dimensionality reduction: successful and efficiently computable, it is nevertheless plagued by a well-known, well-documented sensitivity to outliers. Recent work has considered the setting where each point has a few arbitrarily corrupted components. Yet, in applications of SVD or PCA such as robust collaborative filtering or bioinformatics, malicious agents, defective genes, or simply corrupted or contaminated experiments may effectively yield entire points that are completely corrupted.**

We present an efficient convex optimization-based algorithm we call Outlier Pursuit, that under some mild assumptions on the uncorrupted points (satisfied, e.g., by the standard generative assumption in PCA problems) recovers the *exact* optimal low-dimensional subspace, and identifies the corrupted points. Such identification of corrupted points that do not conform to the low-dimensional approximation, is of paramount interest in bioinformatics, financial applications, and beyond. Our techniques involve matrix decomposition using nuclear norm minimization, however, our results, setup, and approach, necessarily differ considerably from the existing line of work in matrix completion and matrix decomposition, since we develop an approach to recover the correct *column space* of the uncorrupted matrix, rather than the exact matrix itself. In any problem where one seeks to recover a *structure* rather than the *exact initial matrices*, techniques developed thus far relying on certificates of optimality, will fail. We present an important extension of these methods, that allows the treatment of such problems.

I. Introduction

This paper is about the following problem: suppose we are given a large *data matrix* M, and we know it can be decomposed as

$$M = L_0 + C_0,$$

where L_0 is a low-rank matrix, and C_0 is non-zero in only a fraction of the columns. Aside from these broad restrictions, both components are arbitrary. In particular we do not know the rank (or the row/column space) of L_0, or the number and positions of the non-zero columns of C_0. Can we recover the column-space of the low-rank matrix L_0, and the identities of the non-zero columns of C_0, *exactly* and efficiently?

We are primarily motivated by Principal Component Analysis (PCA), arguably the most widely used technique for dimensionality reduction in statistical data analysis. The canonical PCA problem [2], seeks to find the best (in the least-square-error sense) low-dimensional subspace approximation to high-dimensional points. Using the Singular Value Decomposition (SVD), PCA finds the lower-dimensional approximating subspace by forming a low-rank approximation to the data matrix, formed by considering each point as a column; the output of PCA is the (low-dimensional) column space of this low-rank approximation.

It is well known (e.g., [3]–[6]) that standard PCA is extremely fragile to the presence of *outliers*: even a single corrupted point can arbitrarily alter the quality of the approximation. Such non-probabilistic or persistent data corruption may stem from sensor failures, malicious tampering, or the simple fact that some of the available data may not conform to the presumed low-dimensional source / model. In terms of the data matrix, this means that most of the column vectors will lie in a low-dimensional space – and hence the corresponding matrix L_0 will be low-rank – while the remaining columns will be outliers – corresponding to the column-sparse matrix C_0. The natural question in this setting is to ask if we can still (exactly or near-exactly) recover the column space of the uncorrupted points, and the identities of the outliers. This is precisely our problem.

Our results: We consider a novel but natural convex optimization approach to the recovery problem above. The main result of this paper is to establish that, under certain natural conditions, the optimum of this convex program will yield the column space of L_0 and the identities of the outliers (i.e., the non-zero columns of C_0). Our conditions depend on the fraction of points that are outliers (which can otherwise be completely arbitrary), and incoherence of the *row* space of L_0. The latter condition essentially requires that each direction in the column space of L_0 be represented in a sufficient number of non-outlier points; we discuss in more detail below. We note that our results do *not* require incoherence of the column space, as is done, e.g., in the papers [5], [6]. This is due to the different corruption model, our resulting alternative convex formulation, and the fact that their objective is exact recovery. We elaborate on this in Section I-A below. We note that our analytical approach that focuses only on recovery of the column space, instead of "exact recovery" of the entire L_0 matrix. This also means our method's performance is *rotation invariant* – in particular, applying the same rotation to all given points (i.e., columns) will not change its performance. Finally, we extend our analysis to the noisy case when all points – outliers or otherwise – are additionally corrupted by noise.

A. Related Work

Robust PCA has a long history (e.g., [4], [7]–[13]). Each of these algorithms either performs standard PCA on a robust

Manuscript received December 30, 2010; revised June 16, 2011. The work of H. Xu was supported in part by National University of Singapore startup grant R-265-000-384-133. The work of C. Caramanis was supported in part by US National Science Foundation (NSF) (grants EFRI-0735905, CNS-0721532, CNS-0831580) and DTRA (grant HDTRA1-08-0029). The work of S. Sanghavi was supported in part by NSF grants 0954059 and 1017525. The material in this paper was presented in part at NIPS 2010 [1].

H. Xu is with the Department of Mechanical Engineering, National University of Singapore, Singapore 117575 (email: mpexuh@nus.edu.sg)

C. Caramanis and S. Sanghavi are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (email: caramanis@mail.utexas.edu; sanghavi@mail.utexas.edu)

Communicated by J. Romberg, Associate Editor for Signal Processing.

estimate of the covariance matrix, or finds directions that maximize a robust estimate of the variance of the projected data. These algorithms seek to *approximately* recover the column space, and moreover, no existing approach attempts to identify the set of outliers. This outlier identification, while outside the scope of traditional PCA algorithms, is important in a variety of applications such as finance, bio-informatics, and more.

Many existing robust PCA algorithms suffer two pitfalls: performance degradation with dimension increase, and computational intractability. To wit, [14] shows that several robust PCA algorithms including M-estimator [15], Convex Peeling [16], Ellipsoidal Peeling [17], Classical Outlier Rejection [18], Iterative Deletion [19] and Iterative Trimming [20] have breakdown points proportional to the inverse of dimensionality, and hence are useless in the high dimensional regime we consider.

Algorithms with non-diminishing breakdown point, such as Projection-Pursuit [21] are non-convex or even combinatorial, and hence computationally difficult as the size of the problem scales (e.g., [22]). Indeed, to the best of our knowledge, there is no algorithm that exactly solves Projection Pursuit in polynomial time. In contrast to these, the performance of Outlier Pursuit does not depend on the dimension, p, and its running time scales gracefully in problem size (in particular, it can be solved in polynomial time).

Algorithms based on nuclear norm minimization to recover low rank matrices are now standard, since the seminal work [23], [24]. Recent work [5], [6] has taken the nuclear norm minimization approach to the decomposition of a low-rank matrix and an overall sparse matrix. At a high level, these papers are close in spirit to ours, as all look to recover a low-rank matrix from corruptions. However, there are *critical differences* in *(a)* the corruption model: in our paper, a few columns are completely corrupted, while in [5], [6] every column is partially corrupted, *(b)* the objective: the model in [5], [6] allows for exact recovery, as we still have enough information about every row and column, while in our paper this is impossible for the corrupted columns, and we focus on identifying which columns are corrupted, and *(c)* the optimization problem: our corruption matrix is "block sparse" (entire columns) and hence we use the $\ell_{1,2}$ norm [25] to capture our corruption structure, while [5], [6] have simply sparse corruptions, and hence use the ℓ_1 norm. These differences allow us to impose weaker conditions – we do not need incoherence of the column space, making our results *rotation invariant:* applying the same rotation to all points will not affect the performance of our method, while it significantly affects that in [5], [6].

Beyond this, our approach differs in key analysis techniques, which we believe will prove much more broadly applicable and thus of general interest. In particular, our work requires a significant extension of existing techniques for matrix decomposition, precisely because the goal is to recover the *column space* of L_0 (the principal components, in PCA), as opposed to the exact matrices. Indeed, the above works investigate *exact* signal recovery — the intended outcome is known ahead of time, and one just needs to investigate the conditions needed for success. In our setting, however, the convex optimization cannot recover L_0 itself exactly. We introduce the use of an oracle problem, defined by the structures we seek to recover (here, the true column space and the column support). This enables us to show that our convex optimization-based algorithm recovers the correct (or nearly correct, in the presence of noise) column space, as well as the identity of the corrupted points, or outliers.

We believe that this line of analysis will prove to be much more broadly applicable. Often times, exact recovery simply does not make sense under strong corruption models (such as complete column corruption) and the best one can hope for is to capture exactly or approximately, some structural aspect of the problem. In such settings, it may be impossible to follow the proof recipes laid out in works such as [5], [6], [24], [26], that essentially obtain exact recovery from their convex optimization formulations. Thus, in addition to our algorithm and our results, we consider the particular proof technique a contribution of potentially general interest.

II. Problem Setup

The precise PCA with outlier problem that we consider is as follows: we are given n points in p-dimensional space. A fraction $1-\gamma$ of the points lie on a r-dimensional *true* subspace of the ambient $\mathbb{R}^p$, while the remaining γn points are *arbitrarily* located – we call these outliers/corrupted points. We do not have any prior information about the true subspace or its dimension r. Given the set of points, we would like to learn *(a)* the true subspace and *(b)* the identities of the outliers.

As is common practice, we collate the points into a $p \times n$ *data matrix* M, each of whose columns is one of the points, and each of whose rows is one of the p coordinates. It is then clear that the data matrix can be decomposed as

$$M = L_0 + C_0.$$

Here C_0 is the column-sparse matrix ($(1-\gamma)n$ columns are zero) corresponding to the outliers, and L_0 is the matrix corresponding to the non-outliers. Thus, $rank(L_0) = r$, and we assume its columns corresponding to non-zero columns of C_0 are identically zero (whatever those columns were cannot possibly be recovered). Consider its Singular Value Decomposition (SVD)

$$L_0 = U_0 \Sigma_0 V_0^\top. \tag{1}$$

The columns of U_0 form an orthonormal basis for the r-dimensional subspace we wish to recover. C_0 is the matrix corresponding to the outliers; we will denote the set of non-zero columns of C_0 by $\mathcal{I}_0$, with $|\mathcal{I}_0| = \gamma n$. These non-zero columns are completely arbitrary.

With this notation, out intent is to *exactly* recover the column space of L_0, and the set of outliers $\mathcal{I}_0$. All we are given is the matrix M. Clearly, exact recovery is not always going to be possible (regardless of the algorithm used) and thus we need to impose a few weak additional assumptions. We develop these in Section II-A below.

We are also interested in the noisy case, where

$$M = L_0 + C_0 + N,$$

and N corresponds to any additional noise. In this case we are interested in approximate identification of both the true subspace and the outliers.

A. Incoherence: When can the column space be recovered ?

In general, our objective of recovering the "true" column-space of a low-rank matrix that is corrupted with a column-sparse matrix is not always well defined. As an extreme example, consider the case where the data matrix M is non-zero in only one column. Such a matrix is both low-rank and column-sparse, thus the problem is unidentifiable. To make the problem meaningful, we need to impose that the low-rank matrix L_0 cannot itself be column-sparse as well. This is done via the following *incoherence condition*.

Definition: A matrix $L \in \mathbb{R}^{p\times n}$ with SVD $L = U\Sigma V^\top$, and $(1-\gamma)n$ of whose columns are non-zero, is said to be *column-incoherent* with parameter μ if

$$\max_i \|V^\top \mathbf{e}_i\|^2 \leq \frac{\mu r}{(1-\gamma)n},$$

where $\{\mathbf{e}_i\}$ are the coordinate unit vectors.

Thus if V has a column aligned with a coordinate axis, then $\mu = (1-\gamma)n/r$. Similarly, if V is perfectly incoherent (e.g., if $r = 1$ and every non-zero entry of V has magnitude $1/\sqrt{(1-\gamma)n}$) then $\mu = 1$.

In the standard PCA setup, if the points are generated by some low-dimensional isometric (e.g., Gaussian) distribution, then with high probability, one will have $\mu = O(\max(1, \log(n)/r))$ [27]. Alternatively, if the points are generated by a uniform distribution over a *bounded* set, then $\mu = \Theta(1)$.

A small incoherence parameter μ essentially enforces that the matrix L_0 will have column support that is spread out. Note that this is quite natural from the application perspective. Indeed, if the left hand side is as big as 1, it essentially means that one of the directions of the column space which we wish to recover, is defined by only a single observation. Given the regime of a constant fraction of *arbitrarily chosen* and *arbitrarily corrupted* points, such a setting is not meaningful. Having a small incoherence μ is an assumption made in all methods based on nuclear norm minimization up-to-date [5], [6], [27], [28]. Also unidentifiable is the setting where a corrupted point lies in the true subspace. Thus, in matrix terms, we require that every column of C_0 does not lie in the column space of L_0.

We note that this condition is slightly different from the incoherence conditions required for matrix completion in e.g. [27]. In particular, matrix completion requires row-incoherence (a condition on U of the SVD) and joint-incoherence (a condition on the product UV) in addition to the above condition. We do not require these extra conditions because we have a more relaxed objective from our convex program – namely, we only want to recover the column space.

The parameters μ and γ are not required for the execution of the algorithm, and *do not need to be known a priori*. They only arise in the analysis of our algorithm's performance.

Other Notation and Preliminaries: Capital letters such as A are used to represent matrices, and accordingly, A_i denotes the i^{th} column vector. Letters U, V, $\mathcal{I}$ and their variants (complements, subscripts, etc.) are reserved for column space, row space and column support respectively. There are four associated projection operators we use throughout. The projection onto the column space, U, is denoted by $\mathcal{P}_U$ and given by $\mathcal{P}_U(A) = UU^\top A$, and similarly for the row-space $\mathcal{P}_V(A) = AVV^\top$. The matrix $\mathcal{P}_{\mathcal{I}}(A)$ is obtained from A by setting column A_i to zero for all $i \notin \mathcal{I}$. Finally, $\mathcal{P}_T$ is the projection to the space spanned by U and V, and given by $\mathcal{P}_T(\cdot) = \mathcal{P}_U(\cdot) + \mathcal{P}_V(\cdot) - \mathcal{P}_U\mathcal{P}_V(\cdot)$. Note that $\mathcal{P}_T$ depends on U and V, and we suppress this notation wherever it is clear which U and V we are using. The complementary operators, $\mathcal{P}_{U^\perp}, \mathcal{P}_{V^\perp}$, $\mathcal{P}_{T^\perp}$ and $\mathcal{P}_{\mathcal{I}^c}$ are defined as usual. The notation $\mathbb{S}$ is used to represent the invariant subspace (of matrices) of a projection operator: e.g., we write $A \in \mathbb{S}_U$ for any matrix A that satisfies $\mathcal{P}_U(A) = A$. Five matrix norms are used: $\|A\|_*$ is the nuclear norm, $\|A\|$ is the spectral norm, $\|A\|_{1,2}$ is the sum of ℓ_2 norm of the columns A_i, $\|A\|_{\infty,2}$ is the largest ℓ_2 norm of the columns, and $\|A\|_F$ is the Frobenius norm. The only vector norm used is $\|\cdot\|_2$, the ℓ_2 norm. Depending on the context, I is either the unit matrix, or the identity operator; $\mathbf{e}_i$ is the i^{th} standard basis vector. The SVD of L_0 is $U_0\Sigma_0 V_0$. Through out this paper, SVD always refer to rank-reduced (this) SVD. We use r to denote the rank of L_0, and $\gamma \triangleq |\mathcal{I}_0|/n$ the fraction of outliers.

III. Main Results and Consequences

While we do not recover the matrix L_0, we show that the goal of PCA can be attained: even under our strong corruption model, with a constant fraction of points corrupted, we show that we can – under mild assumptions – *exactly* recover both the column space of L_0 (i.e., the low-dimensional space the uncorrupted points lie on) and the column support of C_0 (i.e. the identities of the outliers), from M. If there is additional noise corrupting the data matrix, i.e. if we have $M = L_0 + C_0 + N$, a natural variant of our approach finds a good approximation. In the absence of noise, an easy post-processing step is in fact able to exactly recover the original matrix L_0. We emphasize, however, that the inability to do this simply via the convex optimization step, poses significant technical challenges, as we detail below.

A. Algorithm

Given the data matrix M, our algorithm, called *Outlier Pursuit*, generates *(a)* a matrix U^*, with orthonormal rows, that spans the low-dimensional true subspace we want to recover, and *(b)* a set of column indices $\mathcal{I}^*$ corresponding to the outlier points.

While in the noiseless case there are simple algorithms with similar performance[1], the benefit of the algorithm, and of the analysis, is extension to more realistic and interesting situations where in addition to gross corruption of some

[1]For example, one method is to find a maximal linear independent set of the samples, and remove it from the sample set. Repeat this process. Since the number of outliers is relatively small, eventually they all get removed, and the column space of true samples is recovered.

Algorithm 1 Outlier Pursuit

Find (L^*, C^*), the optimum of the following convex optimization program

$$\begin{array}{ll} \text{Minimize:} & \|L\|_* + \lambda\|C\|_{1,2} \\ \text{Subject to:} & M = L + C \end{array} \tag{2}$$

Compute SVD $L^* = U_1\Sigma_1 V_1^\top$ and output $U^* = U_1$.
Output the set of non-zero columns of C^*, i.e. $\mathcal{I}^* = \{j : c_{ij}^* \neq 0 \text{ for some } i\}$

samples, there is additional noise. Adapting the Outlier Pursuit algorithm, we have the following variant for the noisy case.

Noisy Outlier Pursuit:

$$\begin{array}{ll} \text{Minimize:} & \|L\|_* + \lambda\|C\|_{1,2} \\ \text{Subject to:} & \|M - (L + C)\|_F \leq \varepsilon \end{array} \tag{3}$$

Outlier Pursuit (and its noisy variant) is a convex surrogate for the following natural (but combinatorial and intractable) first approach to the recovery problem:

$$\begin{array}{ll} \text{Minimize:} & \mathrm{rank}(L) + \lambda\|C\|_{0,c} \\ \text{Subject to:} & M = L + C \end{array} \tag{4}$$

where $\|\cdot\|_{0,c}$ stands for the number of non-zero columns of a matrix.

B. Performance

We show that under rather weak assumptions, Outlier Pursuit exactly recovers the column space of the low-rank matrix L_0, and the identities of the non-zero columns of outlier matrix C_0. The formal statement appears below.

Theorem 1 (Noiseless Case): Suppose we observe $M = L_0 + C_0$, where L_0 has rank r and incoherence parameter μ. Suppose further that C_0 is supported on at most γn columns. Any output to Outlier Pursuit recovers the column space exactly, and identifies exactly the indices of columns corresponding to outliers not lying in the recovered column space, as long as the fraction of corrupted points, γ, satisfies

$$\frac{\gamma}{1-\gamma} \leq \frac{c_1}{\mu r}, \tag{5}$$

where $c_1 = \frac{9}{121}$. This can be achieved by setting the parameter λ in the Outlier Pursuit algorithm to be $\frac{3}{7\sqrt{\gamma n}}$ – in fact it holds for any λ in a specific range which we provide below.

Note that we only need to know an upper bound on the number of outliers. This is because the success of Outlier Pursuit is monotonic: if it can recover the column space of L_0 with a certain set of outliers, it will also recover it when an arbitrary subset of these points are converted to non-outliers (i.e., they are replaced by points in the column space of L_0).

For the case where in addition to the corrupted points, we have noisy observations, $\tilde{M} = M + N$, we have the following result.

Theorem 2 (Noisy Case): Suppose we observe $\tilde{M} = M + N = L_0 + C_0 + N$, where

$$\frac{\gamma}{1-\gamma} \leq \frac{c_2}{\mu r}, \tag{6}$$

with $c_2 = \frac{9}{1024}$, and $\|N\|_F \leq \varepsilon$. Let the output of Noisy Outlier Pursuit be L', C'. Then there exists $\tilde{L}, \tilde{C}$ such that $M = \tilde{L} + \tilde{C}$, $\tilde{L}$ has the correct column space, and $\tilde{C}$ the correct column support, and

$$\|L' - \tilde{L}\|_F \leq 20\sqrt{n}\varepsilon; \quad \|C' - \tilde{C}\|_F \leq 18\sqrt{n}\varepsilon.$$

The conditions in this theorem are essentially tight in the following scaling sense (i.e., up to universal constants). If there is no additional structure imposed beyond what we have stated above, then up to scaling, in the noiseless case, Outlier Pursuit can recover from as many outliers (i.e., the same fraction) as any algorithm of possibly arbitrary complexity. In particular, it is easy to see that if the rank of the matrix L_0 is r, and the fraction of outliers satisfies $\gamma \geq 1/(r+1)$, then the problem is not identifiable, i.e., no algorithm can separate authentic and corrupted points. In the presence of stronger assumptions (e.g., isometric distribution) on the authentic points, better recovery guarantees are possible [29].

C. Novelty in Analysis

The main new ingredient in our analysis of the algorithm, is the introduction of an oracle problem. Past matrix recovery papers, including [5], [6], [27], seek exact recovery of the *ground truth*, in our case (L_0, C_0). As such, the generic (and successful) roadmap for the proof technique identifies the first-order necessary and sufficient conditions for the ground truth to be optimal, and then shows that a subgradient certifying optimality of the desired solution exists under the given assumptions. In our setting this is not possible, as the optimum L^* of (2) will be non-zero in every column of C_0 that is not *orthogonal* to L_0's column space. Thus a dual certificate certifying optimality of (L_0, C_0) cannot exist. In terms of recovering the pair (L_0, C_0), this is irrelevant: all we require is for C^* to have the correct column support; given this, recovery of (L_0, C_0) from (L^*, C^*) is immediate – we simply extract the offending columns. Thus, all we need is a dual certificate of optimality for *any feasible pair* $(\hat{L}, \hat{C})$ where $\hat{C}$ has the the correct column support. The challenge is that we do not know, *a priori*, what that pair will be.

We identify this pair using a so-called *oracle problem*, characterizing the pair as the solution to an optimization problem with two additional side constraints: that L have the same column space as L_0, and C have the same column support as C_0. The idea of using an oracle problem appeared previously in analyzing support-recovery property of Lasso and basis-pursuit (see, e.g., [30]–[32]). There, the authors consider an optimal solution directly requiring that it have the correct signed support. There are some significant challenges in our

matrix setting that are not present in the support-recovery problem. Indeed, in the case of support recovery, analysis of the solution is straightforward, because of a special property of the structure being recovered (namely, the support): when the signed support is fixed, regardless of the exact value of the solution, the sub-gradient (of the ℓ_1 norm) is known. This is not true for recovery of more general structures, and in particular, in our setting: the subgradients of both the $\|\cdot\|_*$ and $\|\cdot\|_{1,2}$ norms critically depend on the exact value of the solution to the oracle problem. While the consequence is that more delicate technical analysis is required, one message of this paper is that oracle problems can be broadly useful whenever exact recovery of the ground truth is impossible (or not sought for), and one is only interested in recovering special structures, such as support, block support, spectral properties, and beyond.

IV. Proof of Theorem 1

In this section and the next section, we prove Theorem 1 and Theorem 2.

A. Proof Outline

The detailed proof, provided in subsequent sections, contains a number of cumbersome calculations. To facilitate the flow and highlight the intuition of the proof, we give an outline, emphasizing the novel aspects we introduce, and skip over steps that are largely similar to techniques developed and used in standard literature.

Step 1: Our first step is to construct an Oracle problem. Recall that we want the optimum of (2) to satisfy $\mathcal{P}_{U_0}(L^*) = L^*$ (correct column space) and $\mathcal{P}_{\mathcal{I}_0}(C^*) = C^*$ (correct column support, i.e., identification of the outliers). The oracle problem arises by *imposing* these as additional constraints in (2):

	Oracle Problem:
Minimize:	$\|L\|_* + \lambda\|C\|_{1,2}$
Subject to:	$M = L + C;\ \mathcal{P}_{U_0}(L) = L;\ \mathcal{P}_{\mathcal{I}_0}(C) = C.$

Let $(\hat{L}, \hat{C})$ be an optimal solution to the oracle problem. To show Outlier Pursuit succeeds, it thus suffices to show that $(\hat{L}, \hat{C})$ is also an optimal solution to Outlier Pursuit.

Step 2: The second step is standard. We write down the properties a dual certificate must satisfy to guarantee that $(\hat{L}, \hat{C})$ is optimal to Outlier Pursuit. While the step itself is standard, there is a central challenge arising from the Oracle Problem. As with all results involving low-rank matrix recovery, the left and right singular vectors are a central object of study, critically involved in optimality conditions, etc. Evidently, the side constraints of the oracle problem are not enough to guarantee that L_0 and $\hat{L}$ have the same singular vectors. This forces us to introduce quantities that can relate the two, and understand how these interact with the various projection operators required to describe the subdifferentials. As an important example of this, Lemma 5 defines $\overline{V}$ as the matrix satisfying $\hat{U}\hat{V}^\top = U_0\overline{V}^\top$, and Lemma 6 establishes that $U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top) = \lambda\mathcal{P}_{U_0}(\hat{H})$, for $\hat{H}$ an element of the subdifferential of the $\ell_{1,2}$ norm at $\hat{C}$. With these considerations, we can write down the conditions that a dual certificate Q must satisfy:

$$
\begin{aligned}
&(a) \quad \mathcal{P}_{U_0}(Q) = U_0\overline{V}^\top;\\
&(b) \quad \mathcal{P}_{\overline{V}}(Q) = U_0\overline{V}^\top;\\
&(c) \quad \mathcal{P}_{\mathcal{I}_0}(Q) = \lambda\hat{H};\\
&(d) \quad \|\mathcal{P}_{\hat{T}^\perp}(Q)\| < 1;\\
&(e) \quad \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2} < \lambda.
\end{aligned}
$$

Step 3: The third step is to construct such a Q. A first guess would be to use $Q_0 = U_0\overline{V}^\top + \lambda\hat{H}$. Indeed, this works in the special case where each corrupted column is orthogonal to each authentic one, but fails otherwise. Specifically, we have that

$$\mathcal{P}_{U_0}(Q_0) - U_0\overline{V}^\top = \lambda\mathcal{P}_{U_0}(\hat{H});\ \mathcal{P}_{\mathcal{I}_0}(Q_0) - \lambda\hat{H} = U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top).$$

Recall that $U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top) = \lambda\mathcal{P}_{U_0}(\hat{H})$, we correct Q_0 by

$$\Delta_1 \triangleq \lambda\mathcal{P}_{U_0}(\hat{H}) = U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top).$$

Notice that

$$\mathcal{P}_{\overline{V}}(Q_0 - \Delta_1) = \mathcal{P}_{\overline{V}}\mathcal{P}_{U_0^\perp}(\lambda\hat{H}).$$

Hence we want to further correct Q_0 by Δ_2 such that $\Delta_2 \in \mathbb{S}_{U_0^\perp}$, $\Delta_2 \in \mathbb{S}_{\mathcal{I}_0^c}$, and $\mathcal{P}_{\overline{V}}(\Delta_2) = \mathcal{P}_{\overline{V}}\mathcal{P}_{U_0^\perp}(\lambda\hat{H})$. Such Δ_2 can be constructed using the least-square dual-certificate approach introduced in [27], which gives

$$\Delta_2 \triangleq \mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}]^{-1}\mathcal{P}_{\overline{V}}\mathcal{P}_{U_0^\perp}(\lambda\hat{H}).$$

Lemma 7 and Lemma 8 show that this definition (i.e., the inverse) indeed is meaningful.

Finally, we check that $Q \triangleq Q_0 - \Delta_1 - \Delta_2$ satisfies (a) - (e). Most computation involved is standard, with the exception that we require an incoherence property w.r.t. $\overline{V}$ whereas we only assume an incoherence property w.r.t. V_0. Interestingly, Lemma 10 shows that the latter implies the former, and hence completes the proof.

B. Oracle Problem and Optimality Conditions

We now provide a detailed proof. The notations are heavy, and hence we provide a summary list in Appendix II for the convenience of the readers. We first list some technical preliminaries that we use multiple times in the sequel. The following lemma is well-known, and gives the subgradient of the norms we consider.

Lemma 1: For any column space U, row space V and column support $\mathcal{I}$:

1) Let the SVD of a matrix A be $U\Sigma V^\top$. Then the subgradient to $\|\cdot\|_*$ at A is $\{UV^\top + W | \mathcal{P}_T(W) = 0, \|W\| \le 1\}$ [33].
2) Let the column support of a matrix A be $\mathcal{I}$. Then the subgradient to $\|\cdot\|_{1,2}$ at A is $\{H + Z | \mathcal{P}_{\mathcal{I}}(H) = H, H_i = A_i/\|A_i\|_2;\ \mathcal{P}_{\mathcal{I}}(Z) = 0, \|Z\|_{\infty,2} \le 1\}$.
3) For any A, B, we have $\mathcal{P}_{\mathcal{I}}(AB) = A\mathcal{P}_{\mathcal{I}}(B)$; for any A, $\mathcal{P}_U\mathcal{P}_{\mathcal{I}}(A) = \mathcal{P}_{\mathcal{I}}\mathcal{P}_U(A)$.

Lemma 2: If a matrix $\tilde{H}$ satisfies $\|\tilde{H}\|_{\infty,2} \le 1$ and is supported on $\mathcal{I}$, then $\|\tilde{H}\| \le \sqrt{|\mathcal{I}|}$.

Proof: Using the variational form of the operator norm, we have

$$\|\tilde{H}\| = \max_{\|\mathbf{x}\|_2\leq 1, \|\mathbf{y}\|_2 \leq 1} \mathbf{x}^\top \tilde{H}\mathbf{y} = \max_{\|\mathbf{x}\|_2 \leq 1} \|\mathbf{x}^\top \tilde{H}\|_2$$
$$= \max_{\|\mathbf{x}\|_2\leq 1} \sqrt{\sum_{i=1}^{n} (\mathbf{x}^\top \tilde{H}_i)^2} \leq \sqrt{\sum_{i\in\mathcal{I}} 1} = \sqrt{|\mathcal{I}|}.$$

The inequality holds because $\|\tilde{H}_i\|_2 = 1$ when $i \in \mathcal{I}$, and equals zero otherwise. ∎

Lemma 3: Given a matrix $U \in \mathbb{R}^{r\times n}$ with orthonormal columns, and any matrix $\tilde{V} \in \mathbb{R}^{r\times n}$, we have that $\|U\tilde{V}^\top\|_{\infty,2} = \max_i \|\tilde{V}^\top \mathbf{e}_i\|_2$.

Proof: By definition we have

$$\|U\tilde{V}^\top\|_{\infty,2} = \max_i \|U\tilde{V}_i^\top\|_2$$
$$\overset{(a)}{=} \max_i \|\tilde{V}_i^\top\|_2 = \max_i \|\tilde{V}^\top \mathbf{e}_i\|_2.$$

Here (a) holds since U has orthonormal columns. ∎

As discussed, in general Outlier Pursuit will not recover the true solution (L_0, C_0), and hence it is not possible to construct a subgradient certifying optimality of (L_0, C_0). Instead, our goal is to recover any pair $(\hat{L}, \hat{C})$ so that $\hat{L}$ has the correct column space, and $\hat{C}$ the correct column support. Thus we need only construct a dual certificate for some such pair. We develop our candidate solution $(\hat{L}, \hat{C})$ by imposing precisely these constraints on the original optimization problem (2): the solution $\hat{L}$ should have the correct column space, and $\hat{C}$ should have the correct column support.

Let the SVD of the true L_0 be $L_0 = U_0\Sigma_0 V_0^\top$, and recall that the projection of any matrix X onto the space of all matrices with column space contained in U_0 is given by $\mathcal{P}_{U_0}(X) := U_0U_0^\top X$. Similarly for the column support $\mathcal{I}_0$ of the true C_0, the projection $\mathcal{P}_{\mathcal{I}_0}(X)$ is the matrix that results when all the columns in $\mathcal{I}_0^c$ are set to 0.

Note that U_0 and $\mathcal{I}_0$ above correspond to the *truth*. Thus, with this notation, we would like L^*, C^* the optimum of (2) to satisfy $\mathcal{P}_{U_0}(L^*) = L^*$, as this is nothing but the fact that L^* has recovered the true subspace. Similarly, having C^* satisfy $\mathcal{P}_{\mathcal{I}_0}(C^*) = C^*$ means that we have succeeded in identifying the outliers. The oracle problem arises by *imposing* these as additional constraints in (2):

Oracle Problem:

$$\begin{array}{ll} \text{Minimize:} & \|L\|_* + \lambda\|C\|_{1,2} \\ \text{Subject to:} & M = L + C;\ \mathcal{P}_{U_0}(L) = L;\ \mathcal{P}_{\mathcal{I}_0}(C) = C. \end{array} \tag{7}$$

The problem is of course bounded (by zero), and is feasible, as (L_0, C_0) is a feasible solution. Thus, an optimal solution, denoted as $\hat{L}, \hat{C}$ exists. We now show that the solution $(\hat{L}, \hat{C})$ to the oracle problem, is also an optimal solution to Outlier Pursuit. Unlike the original pair (L_0, C_0), we can certify the optimality of $(\hat{L}, \hat{C})$ by constructing the appropriate subgradient witness.

The next lemma and definition, are key to the development of our optimality conditions.

Lemma 4: Let the pair (L', C') satsify $L' + C' = M$, $\mathcal{P}_{U_0}(L') = L'$, and $\mathcal{P}_{\mathcal{I}_0}(C') = C'$. Denote the SVD of L' as $L' = U'\Sigma V'^\top$, and the column support of C' as $\mathcal{I}'$. Then $U'U'^\top = U_0U_0^\top$, and $\mathcal{I}' \subseteq \mathcal{I}_0$.

Proof: The only thing we need to prove is that L' has a rank no smaller than U_0. However, since $\mathcal{P}_{\mathcal{I}_0}(C') = C'$, we must have $\mathcal{P}_{\mathcal{I}_0^c}(L') = \mathcal{P}_{\mathcal{I}_0^c}(M)$, and thus the rank of L' is at least as large as $\mathcal{P}_{\mathcal{I}_0^c}(M)$, hence L' has a rank no smaller than U_0. ∎

Next we define two operators that are closely related to the subgradient of $\|L'\|_*$ and $\|C'\|_{1,2}$.

Definition 1: Let (L', C') satisfy $L' + C' = M$, $\mathcal{P}_{U_0}(L') = L'$, and $\mathcal{P}_{\mathcal{I}_0}(C') = C'$. We define the following:

$$\mathfrak{N}(L') \triangleq U'V'^\top;$$
$$\mathfrak{G}(C') \triangleq \Big\{ H \in \mathbb{R}^{m\times n} \,\Big|\, \mathcal{P}_{\mathcal{I}_0^c}(H) = 0;\ \forall i \in \mathcal{I}' :\ H_i = \frac{C'_i}{\|C'_i\|_2};$$
$$\forall i \in \mathcal{I}_0 \cap (\mathcal{I}')^c :\ \|H_i\|_2 \leq 1 \Big\},$$

where the SVD of L' is $L' = U'\Sigma V'^\top$, and the column support of C' is $\mathcal{I}'$. Further define the operator $\mathcal{P}_{T(L')}(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ as

$$\mathcal{P}_{T(L')}(X) = \mathcal{P}_{U'}(X) + \mathcal{P}_{V'}(X) - \mathcal{P}_{U'}\mathcal{P}_{V'}(X).$$

Now we present and prove the optimality condition (to Outlier Pursuit) for solutions (L, C) that have the correct column space and support for L and C, respectively.

Theorem 3: Let (L', C') satisfy $L' + C' = M$, $\mathcal{P}_{U_0}(L') = L'$, and $\mathcal{P}_{\mathcal{I}_0}(C') = C'$. Then (L', C') is an optimal solution of Outlier Pursuit if there exists a matrix $Q \in \mathbb{R}^{m\times n}$ that satisfies

$$\begin{array}{ll} (a) & \mathcal{P}_{T(L')}(Q) = \mathfrak{N}(L'); \\ (b) & \|\mathcal{P}_{T(L')^\perp}(Q)\| \leq 1; \\ (c) & \mathcal{P}_{\mathcal{I}_0}(Q)/\lambda \in \mathfrak{G}(C'); \\ (d) & \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2} \leq \lambda. \end{array} \tag{8}$$

If both inequalities are strict (dubbed *Q strictly satisfies (8)*), and $\mathbb{S}_{\mathcal{I}_0} \cap \mathbb{S}_{V'} = \{0\}$, then any optimal solution will have the right column space, and column support.

Proof: By standard convexity arguments [34], a feasible pair (L', C') is an optimal solution of Outlier Pursuit, if there exists a Q' such that

$$Q' \in \partial\|L'\|_*; \quad Q' \in \lambda\partial\|C'\|_{1,2}.$$

Note that (a) and (b) imply that $Q \in \partial\|L'\|_*$. Furthermore, letting $\mathcal{I}'$ be the support of C', then by Lemma 4, $\mathcal{I}' \subseteq \mathcal{I}_0$. Therefore (c) and (d) imply that

$$Q_i = \frac{\lambda C'_i}{\|C'_i\|_2}; \quad \forall i \in \mathcal{I}';$$

and

$$\|Q_i\|_2 \leq \lambda; \quad \forall i \notin \mathcal{I}',$$

which implies that $Q \in \lambda\partial\|C'\|_{1,2}$. Thus, (L', C') is an optimal solution.

The rest of the proof establishes that when (b) and (d) are strict, then any optimal solution (L'', C'') satisfies $\mathcal{P}_{U_0}(L'') = L''$, and $\mathcal{P}_{\mathcal{I}_0}(C'') = C''$. We show that for any fixed $\Delta \neq 0$, $(L' + \Delta, C' - \Delta)$ is strictly worse than (L', C'), unless $\Delta \in$

$\mathcal{P}_{U_0} \cap \mathcal{P}_{\mathcal{I}_0}$. Let W be such that $\|W\| = 1$, $\langle W, \mathcal{P}_{T(L')^\perp}(\Delta)\rangle = \|\mathcal{P}_{T(L')^\perp}\Delta\|_*$, and $\mathcal{P}_{T(L')}W = 0$. Let F be such that

$$F_i = \begin{cases} \frac{-\Delta_i}{\|\Delta_i\|_2} & \text{if } i \notin \mathcal{I}_0, \text{ and } \Delta_i \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathcal{P}_{T(L')}(Q) + W$ is a subgradient of $\|L'\|_*$ and $\mathcal{P}_{\mathcal{I}_0}(Q)/\lambda + F$ is a subgradient of $\|C'\|_{1,2}$. Then we have

$$\begin{aligned}
&\|L' + \Delta\|_* + \lambda\|C' - \Delta\|_{1,2} \\
\geq &\|L'\|_* + \lambda\|C'\|_{1,2} + < \mathcal{P}_{T(L')}(Q) + W, \Delta > \\
&- \lambda < \mathcal{P}_{\mathcal{I}_0}(Q)/\lambda + F, \Delta > \\
= &\|L'\|_* + \lambda\|C'\|_{1,2} + \|\mathcal{P}_{T(L')^\perp}(\Delta)\|_* + \lambda\|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2} \\
&+ < \mathcal{P}_{T(L')}(Q) - \mathcal{P}_{\mathcal{I}_0}(Q), \Delta > \\
= &\|L'\|_* + \lambda\|C'\|_{1,2} + \|\mathcal{P}_{T(L')^\perp}(\Delta)\|_* + \lambda\|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2} \\
&+ < Q - \mathcal{P}_{T(L')^\perp}(Q) - (Q - \mathcal{P}_{\mathcal{I}_0^c}(Q)), \Delta > \\
= &\|L'\|_* + \lambda\|C'\|_{1,2} + \|\mathcal{P}_{T(L')^\perp}(\Delta)\|_* + \lambda\|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2} \\
&+ < -\mathcal{P}_{T(L')^\perp}(Q), \Delta > + < \mathcal{P}_{\mathcal{I}_0^c}(Q), \Delta > \\
\geq &\|L'\|_* + \lambda\|C'\|_{1,2} + (1 - \|\mathcal{P}_{T(L')^\perp}(Q)\|)\|\mathcal{P}_{T(L')^\perp}(\Delta)\|_* \\
&+ (\lambda - \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2})\|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2} \\
\geq &\|L'\|_* + \lambda\|C'\|_{1,2},
\end{aligned}$$

where the last inequality is strict unless

$$\|\mathcal{P}_{T(L')^\perp}(\Delta)\|_* = \|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2} = 0. \tag{9}$$

Note that (9) implies that $\mathcal{P}_{T(L')}(\Delta) = \Delta$ and $\mathcal{P}_{\mathcal{I}_0}(\Delta) = \Delta$. Furthermore

$$\begin{aligned}
\mathcal{P}_{\mathcal{I}_0}(\Delta) &= \Delta = \mathcal{P}_{T(L')}(\Delta) = \mathcal{P}_{U'}(\Delta) + \mathcal{P}_{V'}\mathcal{P}_{U'^\perp}(\Delta) \\
&= \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U'}(\Delta) + \mathcal{P}_{V'}\mathcal{P}_{U'^\perp}(\Delta),
\end{aligned}$$

where the last equality holds because we can write $\mathcal{P}_{\mathcal{I}_0}(\Delta) = \Delta$. This leads to

$$\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U'^\perp}(\Delta) = \mathcal{P}_{V'}\mathcal{P}_{U'^\perp}(\Delta).$$

Lemma 4 implies $\mathcal{P}_{U'}(\cdot) = \mathcal{P}_{U_0}(\cdot)$, which means $\mathcal{P}_{U_0^\perp}(\Delta) \in \mathbb{S}_{\mathcal{I}_0} \cap \mathbb{S}_{V'}$, and hence equal 0. Thus, $\Delta \in \mathbb{S}_{U_0}$. Recall that Equation (9) implies $\Delta \in \mathbb{S}_{\mathcal{I}_0}$, we then have $\Delta \in \mathbb{S}_{\mathcal{I}_0} \cap \mathbb{S}_{U_0}$, which completes the proof. ∎

Thus, the oracle problem determines a solution pair, $(\hat{L}, \hat{C})$, and then using this, Theorem 3 above, gives the conditions a dual certificate must satisfy. The rest of the proof seeks to build a dual certificate for the pair $(\hat{L}, \hat{C})$. To this end, The following two results are quite helpful in what follows. For the remainder of the paper, we use $(\hat{L}, \hat{C})$ to denote the solution pair that is the output of the oracle problem, and we assume that the SVD of $\hat{L}$ is given as $\hat{L} = \hat{U}\hat{\Sigma}\hat{V}^\top$.

Lemma 5: There exists an orthonormal matrix $\overline{V} \in \mathbb{R}^{n\times r}$ such that

$$\hat{U}\hat{V}^\top = U_0\overline{V}^\top.$$

In addition,

$$\begin{aligned}
\mathcal{P}_{\hat{T}}(\cdot) &\triangleq \mathcal{P}_{\hat{U}}(\cdot) + \mathcal{P}_{\hat{V}}(\cdot) - \mathcal{P}_{\hat{U}}\mathcal{P}_{\hat{V}}(\cdot) \\
&= \mathcal{P}_{U_0}(\cdot) + \mathcal{P}_{\overline{V}}(\cdot) - \mathcal{P}_{U_0}\mathcal{P}_{\overline{V}}(\cdot).
\end{aligned}$$

Proof: Due to Lemma 4, we have $U_0U_0^\top = \hat{U}\hat{U}^\top$, hence $U_0 = \hat{U}\hat{U}^\top U_0$. Letting $\overline{V} = \hat{V}\hat{U}^\top U_0$, we have $\hat{U}\hat{V}^\top = U_0\overline{V}^\top$, and $\overline{V}\overline{V}^\top = \hat{V}\hat{V}^\top$. Note that $U_0U_0^\top = \hat{U}\hat{U}^\top$ leads to $\mathcal{P}_U(\cdot) = \mathcal{P}_{\hat{U}}(\cdot)$, and $\overline{V}\overline{V}^\top = \hat{V}\hat{V}^\top$ leads to $\mathcal{P}_{\overline{V}}(\cdot) = \mathcal{P}_{\hat{V}}(\cdot)$, so the second claim follows. ∎

Since $\hat{L}, \hat{C}$ is an optimal solution to Oracle Problem (7), there exists Q_1, Q_2, A' and B' such that

$$Q_1 + \mathcal{P}_{U_0^\perp}(A') = Q_2 + \mathcal{P}_{\mathcal{I}_0^c}(B'),$$

where Q_1, Q_2 are subgradients to $\|\hat{L}\|_*$ and to $\lambda\|\hat{C}\|_{1,2}$, respectively. This means that $Q_1 = U_0\overline{V}^\top + W$ for some orthonormal $\overline{V}$ and W such that $\mathcal{P}_{\hat{T}}(W) = 0$, and $Q_2 = \lambda(\hat{H} + Z)$ for some $\hat{H} \in \mathfrak{G}(\hat{C})$, and Z such that $\mathcal{P}_{\mathcal{I}_0}(Z) = 0$. Letting $A = W + A'$, $B = \lambda Z + B'$, we have

$$U_0\overline{V}^\top + \mathcal{P}_{U_0^\perp}(A) = \lambda\hat{H} + \mathcal{P}_{\mathcal{I}_0^c}(B). \tag{10}$$

Recall that $\hat{H} \in \mathfrak{G}(\hat{C})$ means $\mathcal{P}_{\mathcal{I}_0}(\hat{H}) = \hat{H}$ and $\|\hat{H}\|_{\infty,2} \leq 1$.

Lemma 6: We have

$$U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top) = \lambda\mathcal{P}_{U_0}(\hat{H}).$$

Proof: We have

$$\begin{aligned}
&\mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(U_0\overline{V}^\top + \mathcal{P}_{U_0^\perp}(A)) \\
= \;&\mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(U_0\overline{V}^\top) + \mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(\mathcal{P}_{U_0^\perp}(A)) \\
= \;&U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top) + \mathcal{P}_{U_0}\mathcal{P}_{U_0^\perp}\mathcal{P}_{\mathcal{I}_0}(A) \\
= \;&U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top).
\end{aligned}$$

Furthermore, we have

$$\mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(\lambda\hat{H} + \mathcal{P}_{\mathcal{I}_0^c}(B)) = \lambda\mathcal{P}_{U_0}(\hat{H}).$$

The lemma follows from (10). ∎

C. Obtaining Dual Certificates for Outlier Pursuit

In this section, we complete the proof of Theorem 1 by constructing a dual certificate for $(\hat{L}, \hat{C})$ – the solution to the oracle problem – showing it is also the solution to Outlier Pursuit. The conditions the dual certificate must satisfy are spelled out in Theorem 3. It is helpful to first consider the simpler case where the corrupted columns are assumed to be orthogonal to the column space of L_0 which we seek to recover. Indeed, in that setting, we have $V_0 = \hat{V} = \overline{V}$, and moreover, straightforward algebra shows that we automatically satisfy the condition $\mathbb{S}_{\mathcal{I}_0} \cap \mathbb{S}_{V_0} = \{0\}$. (In the general case, however, we require an additional condition to be satisfied, in order to recover the same property.) Since the columns of H_0 are either zero, or defined as normalizations of the columns of matrix C_0 (i.e., normalizations of outliers), we immediately conclude that $\mathcal{P}_{U_0}(H) = \mathcal{P}_{V_0}(H) = \mathcal{P}_T(H) = 0$, and also $\mathcal{P}_{\mathcal{I}_0}(U_0V_0^\top) = 0$. As a result, it is not hard to verify that the dual certificate for the orthogonal case is:

$$Q_0 = U_0V_0^\top + \lambda H_0.$$

While not required for the proof of our main results, we include the proof of the orthogonal case in Appendix I, as there we get a stronger *necessary and sufficient* condition for recovery.

For the general, non-orthogonal case, however, this certificate does not satisfy the conditions of Theorem 3. For instance,

$\mathcal{P}_{V_0}(H_0)$ need no longer be zero, and hence the condition $\mathcal{P}_T(Q_0) = U_0V_0^\top$ may no longer hold. We correct for the effect of the non-orthogonality by modifying Q_0 with matrices Δ_1 and Δ_2, which we define below.

Recalling the definition of $\overline{V}$ from Lemma 5, define matrix $G \in \mathbb{R}^{r\times r}$ as

$$G \triangleq \mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)(\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top))^\top. \tag{11}$$

Then we have

$$\begin{aligned} G &= \sum_{i\in\mathcal{I}_0}[(\overline{V}^\top)_i][(\overline{V}^\top)_i]^\top \\ &\preceq \sum_{i=1}^{n}[(\overline{V}^\top)_i][(\overline{V}^\top)_i]^\top = \overline{V}^\top\overline{V} = I, \end{aligned}$$

where $\preceq$ is the generalized inequality induced by the positive semi-definite cone. Hence, $\|G\| \le 1$. The following lemma bounds $\|G\|$ away from 1.

Lemma 7: Let $\psi = \|G\|$. Then $\psi \le \lambda^2\gamma n$. In particular, for $\lambda \le 3/(7\sqrt{\gamma n})$, we have $\psi < 1/4$.

Proof: We have

$$\begin{aligned} \psi &= \|U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)(\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top))^\top U_0^\top\| \\ &= \|[U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)][U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)]^\top\|, \end{aligned}$$

due to the fact that U_0 is orthonormal. By Lemma 6, this implies

$$\begin{aligned} \psi &= \|[\lambda\mathcal{P}_{U_0}(\hat{H})][\lambda\mathcal{P}_{U_0}(\hat{H})]^\top\| \\ &= \lambda^2\|\sum_{i\in\mathcal{I}_0}\mathcal{P}_{U_0}(\hat{H}_i)\mathcal{P}_{U_0}(\hat{H}_i)^\top\| \\ &\le \lambda^2|\mathcal{I}_0| \\ &= \lambda^2\gamma n. \end{aligned}$$

The inequality holds because $\|\mathcal{P}_{U_0}(\hat{H}_i)\|_2 \le 1$ implies $\|\mathcal{P}_{U_0}(\hat{H}_i)\mathcal{P}_{U_0}(\hat{H}_i)^\top\| \le 1$. ■

Lemma 8: If $\psi < 1$, then the following operation $\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}$ is an injection from $\mathcal{P}_{\overline{V}}$ to $\mathcal{P}_{\overline{V}}$, and its inverse operation is $I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i$.

Proof: Fix matrix $X \in \mathbb{R}^{p\times n}$ such that $\|X\| = 1$, we have that

$$\begin{aligned} \mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}}(X) &= \mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}(X\overline{V}\overline{V}^\top) \\ &= \mathcal{P}_{\overline{V}}(X\overline{V}\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)) \\ &= X\overline{V}\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\overline{V}\overline{V}^\top \\ &= X\overline{V}(\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\overline{V})\overline{V}^\top \\ &= X\overline{V}G\overline{V}^\top, \end{aligned}$$

which leads to $\|\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}}(X)\| \le \psi$. Since $\psi < 1$, $[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i](X)$ is well defined, and has a spectral norm not larger than $1/(1-\psi)$.

Note that we have

$$\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}} = \mathcal{P}_{\overline{V}}(I - \mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}}),$$

thus for any $X \in \mathcal{P}_{\overline{V}}$ the following holds

$$\begin{aligned} &\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i](X) \\ =\ &\mathcal{P}_{\overline{V}}(I - \mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i](X) \\ =\ &\mathcal{P}_{\overline{V}}(X) = X, \end{aligned}$$

which establishes the lemma. ■

Now we define the matrices Δ_1 and Δ_2 used to construct the dual certificate. As the proof reveals, they are designed precisely as "corrections" to guarantee that the dual certificate satisfies the required constraints of Theorem 3.

Define Δ_1 and Δ_2 as follows:

$$\Delta_1 \triangleq \lambda\mathcal{P}_{U_0}(\hat{H}) = U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top); \tag{12}$$

$$\begin{aligned} \Delta_2 &\triangleq \mathcal{P}_{U_0^\perp}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}(\lambda\hat{H}) \\ &= \mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}\mathcal{P}_{U_0^\perp}(\lambda\hat{H}) \end{aligned} \tag{13}$$

The equality holds since $\mathcal{P}_{\overline{V}}, \mathcal{P}_{\mathcal{I}_0}, \mathcal{P}_{\mathcal{I}_0^c}$ are all given by right matrix multiplication, while $\mathcal{P}_{U_0^\perp}$ is given by left matrix multiplication.

Theorem 4: Assume $\psi < 1$. Let

$$Q \triangleq U_0\overline{V}^\top + \lambda\hat{H} - \Delta_1 - \Delta_2.$$

If

$$\frac{\gamma}{1-\gamma} \le \frac{(1-\psi)^2}{(3-\psi)^2\mu r},$$

and

$$\frac{(1-\psi)\sqrt{\frac{\mu r}{1-\gamma}}}{\sqrt{n}(1-\psi-\sqrt{\frac{\gamma}{1-\gamma}\mu r})} \le \lambda \le \frac{1-\psi}{(2-\psi)\sqrt{n\gamma}},$$

then Q satisfies Condition (8) (i.e., it is the dual certificate). If all inequalities hold strictly, then Q strictly satisfies (8).

Proof: Note that $\psi < 1$ implies $\mathbb{S}_{\overline{V}} \cap \mathbb{S}_{\mathcal{I}_0} = \{0\}$. Hence it suffices to show that Q simultaneously satisfies

$$\begin{aligned} &(1) \quad \mathcal{P}_{\hat{U}}(Q) = \hat{U}\hat{V}^\top; \\ &(2) \quad \mathcal{P}_{\hat{V}}(Q) = \hat{U}\hat{V}^\top; \\ &(3) \quad \mathcal{P}_{\mathcal{I}_0}(Q) = \lambda\hat{H}; \\ &(4) \quad \|\mathcal{P}_{\hat{T}^\perp}(Q)\| \le 1; \\ &(5) \quad \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2} \le \lambda. \end{aligned}$$

We prove that each of these five conditions holds, in Steps 1-5. Then in Step 6, we show that the condition on λ is not vacuous, i.e., the lower bound is strictly less than then upper bound (and in fact, we then show that $\lambda = \frac{3}{7\sqrt{\gamma n}}$ is in the specified range).

Step 1: We have

$$\begin{aligned}\mathcal{P}_{\hat{U}}(Q) &= \mathcal{P}_{U_0}(Q)\\ &= \mathcal{P}_{U_0}(U_0\overline{V}^\top + \lambda\hat{H} - \Delta_1 - \Delta_2)\\ &= U_0\overline{V}^\top + \lambda\mathcal{P}_{U_0}(\hat{H}) - \mathcal{P}_{U_0}(\Delta_1) - \mathcal{P}_{U_0}(\Delta_2)\\ &= U_0\overline{V}^\top\\ &= \hat{U}\hat{V}^\top.\end{aligned}$$

Step 2: We have

$$\begin{aligned}&\mathcal{P}_{\hat{V}}(Q) = \mathcal{P}_{\overline{V}}(Q)\\ =\ & \mathcal{P}_{\overline{V}}(U_0\overline{V}^\top + \lambda\hat{H} - \Delta_1 - \Delta_2)\\ =\ & U_0\overline{V}^\top + \mathcal{P}_{\overline{V}}(\lambda\hat{H}) - \mathcal{P}_{\overline{V}}(\lambda\mathcal{P}_{U_0}(\hat{H}))\\ & -\mathcal{P}_{\overline{V}}\{\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}\mathcal{P}_{U_0^\perp}(\lambda\hat{H})\}\\ =\ & U_0\overline{V}^\top + \mathcal{P}_{\overline{V}}(\mathcal{P}_{U_0^\perp}(\lambda\hat{H}))\\ & -\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}\mathcal{P}_{U_0^\perp}(\lambda\hat{H})\\ \overset{(a)}{=}\ & U_0\overline{V}^\top + \mathcal{P}_{\overline{V}}(\mathcal{P}_{U_0^\perp}(\lambda\hat{H})) - \mathcal{P}_{\overline{V}}(\mathcal{P}_{U_0^\perp}(\lambda\hat{H}))\\ =\ & U_0\overline{V}^\top\\ =\ & \hat{U}\hat{V}^\top.\end{aligned}$$

Here, (a) holds since on $\mathcal{P}_{\overline{V}}$, $[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]$ is the inverse operation of $\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}$.

Step 3: We have

$$\begin{aligned}&\mathcal{P}_{\mathcal{I}_0}(Q) = \mathcal{P}_{\mathcal{I}_0}(U_0\overline{V}^\top + \lambda\hat{H} - \Delta_1 - \Delta_2)\\ =\ & U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top) + \lambda\hat{H} - \mathcal{P}_{\mathcal{I}_0}(U_0\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top))\\ & -\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}\mathcal{P}_{U_0^\perp}(\lambda\hat{H})\\ =\ & \lambda\hat{H}.\end{aligned}$$

Step 4: We need a lemma first.

Lemma 9: Given $X \in \mathbb{R}^{p\times n}$ such that $\|X\| = 1$, we have $\|\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}(X)\| \le 1$.

Proof: By definition,

$$\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}(X) = X\overline{V}\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top).$$

For any $\mathbf{z} \in \mathbb{R}^n$ such that $\|\mathbf{z}\|_2 = 1$, we have

$$\begin{aligned}&\|X\overline{V}\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)\mathbf{z}\|_2 = \|X\overline{V}\overline{V}^\top\mathcal{P}_{\mathcal{I}_0^c}(\mathbf{z})\|_2\\ \le\ & \|X\|\|\overline{V}\overline{V}^\top\|\|\mathcal{P}_{\mathcal{I}_0^c}(\mathbf{z})\|_2 \le 1,\end{aligned}$$

where we use $\mathcal{P}_{\mathcal{I}_0^c}(\mathbf{z})$ to represent the vector whose coordinates $i \in \mathcal{I}_0$ are set to zero. The last inequality follows from the fact that $\|X\| = 1$. Note that this holds for any $\mathbf{z}$, hence by the definition of spectral norm (as the ℓ_2 operator norm), the lemma follows. ∎

Now we continue with Step 4. We have

$$\begin{aligned}&\mathcal{P}_{\hat{T}^\perp}(Q)\\ =&\mathcal{P}_{\hat{T}^\perp}(U_0\overline{V}^\top + \lambda\hat{H} - \Delta_1 - \Delta_2)\\ =&\mathcal{P}_{\overline{V}^\perp}\mathcal{P}_{U_0^\perp}(\lambda\hat{H})\\ &- \mathcal{P}_{\overline{V}^\perp}\mathcal{P}_{U_0^\perp}(\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}\mathcal{P}_{U_0^\perp}(\lambda\hat{H}))\\ =&\mathcal{P}_{\overline{V}^\perp}\mathcal{P}_{U_0^\perp}(\lambda\hat{H})\\ &- \mathcal{P}_{U_0^\perp}\mathcal{P}_{\overline{V}^\perp}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}(\lambda\hat{H}).\end{aligned}$$

Notice that $\|\mathcal{P}_{\overline{V}^\perp}\mathcal{P}_{U_0^\perp}(\lambda\hat{H})\| \le \|\lambda\hat{H}\|$. Furthermore, we have the following:

$$\begin{aligned}&\|\mathcal{P}_{U_0^\perp}\mathcal{P}_{\overline{V}^\perp}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}(\lambda\hat{H})\|\\ \le\ & \|\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}(\lambda\hat{H})\|\\ \le\ & \|[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}(\lambda\hat{H})\|\\ \le\ & \|\mathcal{P}_{\overline{V}}(\lambda\hat{H})\|/(1-\psi)\\ \le\ & \|\lambda\hat{H}\|/(1-\psi).\end{aligned}$$

Recall that we have shown $\|\lambda\hat{H}\| \le \lambda\sqrt{|\mathcal{I}_0|}$. Thus we have that

$$\|\mathcal{P}_{\hat{T}^\perp}(Q)\| \le \frac{2-\psi}{1-\psi}\lambda\sqrt{|\mathcal{I}_0|}.$$

From the assumptions of the theorem, we have

$$\lambda \le \frac{1-\psi}{(2-\psi)\sqrt{n\gamma}},$$

and hence

$$\|\mathcal{P}_{\hat{T}^\perp}(Q)\| \le 1.$$

The inequality will be strict if

$$\lambda < \frac{1-\psi}{(2-\psi)\sqrt{n\gamma}}.$$

Step 5: We first need a lemma that shows that the incoherence parameter for the matrix $\overline{V}$ is no larger than the incoherence parameter of the original matrix V_0.

Lemma 10: Define the incoherence of $\overline{V}$ as follows:

$$\overline{\mu} = \max_{i\in\mathcal{I}_0^c}\frac{|\mathcal{I}_0^c|}{r}\|\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)\mathbf{e}_i\|^2.$$

Then $\overline{\mu} \le \mu$.

Proof: Recall that $L_0 = U_0\Sigma_0V_0^\top$, and

$$\mu = \max_{i\in\mathcal{I}_0^c}\frac{|\mathcal{I}_0^c|}{r}\|\mathcal{P}_{\mathcal{I}_0^c}(V_0^\top)\mathbf{e}_i\|^2.$$

Thus it suffices to show that for fixed $i \in \mathcal{I}_0$, the following holds:

$$\|\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)\mathbf{e}_i\| \le \|\mathcal{P}_{\mathcal{I}_0^c}(V_0^\top)\mathbf{e}_i\|.$$

Note that $\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)$ and $\mathcal{P}_{\mathcal{I}_0^c}(V_0^\top)$ span the same row space. Thus, due to the fact that $\mathcal{P}_{\mathcal{I}_0^c}(V_0^\top)$ is orthonormal, we

conclude that $\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)$ is row-wise full rank. Since $0 \preceq \mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)^\top = I - G$, and $G \succeq 0$, there exists a symmetric, invertible matrix $Y \in \mathbb{R}^{r\times r}$, such that

$$\|Y\| \leq 1; \quad \text{and} \quad Y^2 = \mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)^\top.$$

This in turn implies that $Y^{-1}\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)$ is orthonormal and spans the same row space as $\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)$, and hence spans the same row space as $\mathcal{P}_{\mathcal{I}_0^c}(V_0^\top)$. Note that $\mathcal{P}_{\mathcal{I}_0^c}(V_0^\top)$ is also orthonormal, which implies there exists an orthonormal matrix $Z \in \mathbb{R}^{r\times r}$, such that

$$ZY^{-1}\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top) = \mathcal{P}_{\mathcal{I}_0^c}(V_0^\top).$$

We have

$$\begin{aligned} &\|\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)\mathbf{e}_i\|_2 = \|YZ^\top\mathcal{P}_{\mathcal{I}_0^c}(V_0^\top)\mathbf{e}_i\|_2 \\ \leq\ &\|Y\|\|Z^\top\|\|\mathcal{P}_{\mathcal{I}_0^c}(V_0^\top)\mathbf{e}_i\|_2 \leq \|\mathcal{P}_{\mathcal{I}_0^c}(V_0^\top)\mathbf{e}_i\|_2. \end{aligned}$$

This concludes the proof of the lemma. ∎

Now, recall from the proof of Lemma 8 that

$$\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}}(X) = X\overline{V}G\overline{V}^\top.$$

Hence, noting that $(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i = (\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^{i-1}$ and $\overline{V}^\top\overline{V} = I$, by induction we have

$$(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i(X) = X\overline{V}G^i\overline{V}^\top.$$

We use this to expand Δ_2:

$$\begin{aligned} \Delta_2 \ &= \mathcal{P}_{U^\perp}\mathcal{P}_{\mathcal{I}_0^c}\mathcal{P}_{\overline{V}}[I + \sum_{i=1}^{\infty}(\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}})^i]\mathcal{P}_{\overline{V}}(\lambda\hat{H}) \\ &= \ (I - U_0U_0^\top)(\lambda\hat{H})\overline{V}\overline{V}^\top[1 + \sum_{i=1}^{\infty}\overline{V}G^i\overline{V}^\top]\overline{V}\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top). \end{aligned}$$

Thus, we have

$$\begin{aligned} \|\Delta_2\mathbf{e}_i\|_2 \ &\leq \ \|(I - U_0U_0^\top)\|\|(\lambda H)\|\|\overline{V}\overline{V}^\top\| \\ &\quad \times\|1 + \sum_{i=1}^{\infty}\overline{V}G^i\overline{V}^\top\|\|\overline{V}\|\|\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)\mathbf{e}_i\|_2 \\ &\leq \ \|\lambda H\|\frac{1}{1-\psi}\sqrt{\frac{\overline{\mu}r}{n - |\mathcal{I}_0|}} \\ &\leq \ \frac{\lambda\sqrt{|\mathcal{I}_0|}\sqrt{\frac{\mu r}{n-|\mathcal{I}_0|}}}{1-\psi}, \end{aligned}$$

where we have used Lemma 10 in the last inequality. This now implies

$$\|\Delta_2\|_{\infty,2} \leq \frac{\lambda\sqrt{|\mathcal{I}_0|}\sqrt{\frac{\mu r}{n-|\mathcal{I}_0|}}}{1-\psi}.$$

Notice that

$$\begin{aligned} \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2} =& \|\mathcal{P}_{\mathcal{I}_0^c}(U_0\overline{V}^\top + \lambda\hat{H} - \Delta_1 - \Delta_2)\|_{\infty,2} \\ =& \|U_0\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top) - \Delta_2\|_{\infty,2} \\ \leq& \|U_0\mathcal{P}_{\mathcal{I}_0^c}(\overline{V}^\top)\|_{\infty,2} + \|\Delta_2\|_{\infty,2} \\ \leq& \sqrt{\frac{\mu r}{n - |\mathcal{I}_0|}} + \frac{\lambda\sqrt{|\mathcal{I}_0|}\sqrt{\frac{\mu r}{n-|\mathcal{I}_0|}}}{1-\psi}. \end{aligned}$$

Therefore, showing that $\|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2} \leq \lambda$ is equivalent to showing

$$\begin{aligned} &\sqrt{\frac{\mu r}{n - |\mathcal{I}_0|}} + \frac{\lambda\sqrt{|\mathcal{I}_0|}\sqrt{\frac{\mu r}{n-|\mathcal{I}_0|}}}{1-\psi} \leq \lambda \\ \iff \quad &\lambda\left(1 - \frac{\sqrt{\frac{\gamma}{1-\gamma}\mu r}}{1-\psi}\right) \geq \sqrt{\frac{\mu r}{n(1-\gamma)}} \\ \iff \quad &\lambda \geq \frac{(1-\psi)\sqrt{\frac{\mu r}{1-\gamma}}}{\sqrt{n}(1 - \psi - \sqrt{\frac{\gamma}{1-\gamma}\mu r})}, \end{aligned}$$

as long as $1 - \psi - \sqrt{\frac{\gamma}{1-\gamma}\mu r} > 0$ (which is proved in Step 6).

Step 6: We have shown that each of the 5 conditions holds. Finally, we show that the theorem's conditions on λ can be satisfied. But this amounts to a condition on γ. Indeed, we have:

$$\begin{aligned} &\frac{(1-\psi)\sqrt{\frac{\mu r}{1-\gamma}}}{\sqrt{n}(1 - \psi - \sqrt{\frac{\gamma}{1-\gamma}\mu r})} \leq \frac{1-\psi}{(2-\psi)\sqrt{n\gamma}} \\ \iff \quad &(2-\psi)\sqrt{\frac{\gamma}{1-\gamma}\mu r} \leq 1 - \psi - \sqrt{\frac{\gamma}{1-\gamma}\mu r} \\ \iff \quad &\frac{\gamma}{1-\gamma} \leq \frac{(1-\psi)^2}{(3-\psi)^2\mu r}, \end{aligned}$$

which can certainly be satisfied, since the right hand side does not depend on γ. Moreover, observe that under this condition, $1 - \psi - \sqrt{\frac{\gamma}{1-\gamma}\mu r} > 0$ holds. Note that if the last inequality holds strictly, then so does the first. ∎

We have thus shown that as long as $\psi < 1$, then for λ within the given bounds, we can construct a dual certificate. From here, the following corollary immediately establishes our main result, Theorem 1.

Corollary 1: Let $\gamma \leq \gamma^*$. Then any solution to Outlier Pursuit with $\lambda = \frac{3}{7\sqrt{\gamma^* n}}$, identifies the correct column space and support of outlier, as long as

$$\frac{\gamma^*}{1-\gamma^*} \leq \frac{9}{121\mu r}.$$

Proof: First note that $\lambda = \frac{3}{7\sqrt{\gamma^* n}}$ and $\gamma \leq \gamma^*$ together imply that

$$\lambda \leq \frac{3}{7\sqrt{\gamma n}},$$

which by Lemma 7 leads to

$$\psi \leq \lambda^2\gamma n < \frac{1}{4}.$$

Thus, it suffices to check that γ and λ satisfy the conditions of Theorem 4, namely

$$\frac{\gamma}{1-\gamma} < \frac{(1-\psi)^2}{(3-\psi)^2\mu r},$$

and

$$\frac{(1-\psi)\sqrt{\frac{\mu r}{1-\gamma}}}{\sqrt{n}(1 - \psi - \sqrt{\frac{\gamma}{1-\gamma}\mu r})} < \lambda < \frac{1-\psi}{(2-\psi)\sqrt{n\gamma}}.$$

Since $\psi < 1/4$, we have

$$\frac{\gamma}{1-\gamma} \leq \frac{\gamma^*}{1-\gamma^*} \leq \frac{9}{121\mu r} = \frac{(1-1/4)^2}{(3-1/4)^2\mu r} < \frac{(1-\psi)^2}{(3-\psi)^2\mu r},$$

which proves the first condition.

Next, observe that $\frac{(1-\psi)\sqrt{\frac{\mu r}{1-\gamma}}}{\sqrt{n}(1-\psi-\sqrt{\frac{\gamma}{1-\gamma}\mu r})}$, as a function of $\psi, \gamma, (\mu r)$ is strictly increasing in ψ, (μr), and γ. Moreover, $\mu r \leq \frac{(1-\psi)^2(1-\gamma)}{(3-\psi)^2\gamma}$, and thus

$$\begin{aligned}&\frac{(1-\psi)\sqrt{\frac{\mu r}{1-\gamma}}}{\sqrt{n}(1-\psi-\sqrt{\frac{\gamma}{1-\gamma}\mu r})} < \frac{(1-\psi)\sqrt{\frac{(1-\psi)^2}{(3-\psi)^2\gamma}}}{\sqrt{n}(1-\psi-\frac{1-\psi}{3-\psi})}\\ &= \frac{3\sqrt{1+\gamma/(1-\gamma)}}{7\sqrt{n}} \leq \frac{3\sqrt{1+\gamma^*/(1-\gamma^*)}}{7\sqrt{n}} = \lambda.\end{aligned}$$

Similarly, $\frac{1-\psi}{(2-\psi)\sqrt{n\gamma}}$ is strictly decreasing in ψ and γ, which implies that

$$\frac{1-\psi}{(2-\psi)\sqrt{n\gamma}} > \frac{1-1/4}{(2-1/4)\sqrt{n\gamma^*}} = \lambda.$$

■

V. Proof of Theorem 2: The Case of Noise

In practice, the observed matrix may be a noisy copy of M. In this section, we investigate this noisy case and show that the proposed method, with minor modification, is robust to noise. Specifically, we observe $M' = M + N$ for some unknown N, and we want to approximately recover U_0 and $\mathcal{I}_0$. This leads to the following formulation that replaces the equality constraint $M = L + C$ with a norm inequality.

$$\begin{aligned}\text{Minimize:}\quad & \|L\|_* + \lambda\|C\|_{1,2}\\ \text{Subject to:}\quad & \|M' - L - C\|_F \leq \epsilon.\end{aligned} \tag{14}$$

In fact, we show in this section that under the essentially equivalent conditions as that of the noiseless case, Noisy Outlier Pursuit succeeds. Here, we say that the algorithm "succeeds" if the optimal solution of (14) is "close" to a pair that has the correct column space and column support. To this end, we first establish the next theorem – a counterpart in the noisy case of Theorem 3 – that states that Noisy Outlier Pursuit succeeds if there exists a dual certificate (with slightly stronger requirements than the noiseless case) for decomposing the *noiseless matrix* M. Then, applying our results on constructing the dual certificate from the previous section, we have that Noisy Outlier Pursuit succeeds under the essentially equivalent conditions as that of the noiseless case.

Theorem 5: Let L', C' be an optimal solution of (14). Suppose $\|N\|_F \leq \epsilon$, $\lambda < 1$, and $\psi < 1/4$. Let $M = \hat{L} + \hat{C}$ where $\mathcal{P}_U(\hat{L}) = \hat{L}$ and $\mathcal{P}_{\mathcal{I}_0}(\hat{C}) = \hat{C}$. If there exists a Q such that

$$\begin{aligned}&\mathcal{P}_{T(\hat{L})}(Q) = \mathfrak{N}(\hat{L}); \quad \|\mathcal{P}_{T(\hat{L})^\perp}(Q)\| \leq 1/2;\\ &\mathcal{P}_{\mathcal{I}_0}(Q)/\lambda \in \mathfrak{G}(\hat{C}); \quad \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2} \leq \lambda/2,\end{aligned} \tag{15}$$

then there exists a pair $(\tilde{L}, \tilde{C})$ such that $M = \tilde{L} + \tilde{C}$, $\tilde{L} \in \mathcal{P}_{U_0}$, $\tilde{C} \in \mathcal{P}_{\mathcal{I}_0}$ and

$$\|L' - \tilde{L}\|_F \leq 20\sqrt{n}\epsilon; \quad \|C' - \tilde{C}\|_F \leq 18\sqrt{n}\epsilon.$$

Proof: Let $\overline{V}$ be as defined before. We establish the following lemma first.

Lemma 11: Recall that $\psi = \|G\|$ where $G = \mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)^\top$. We have

$$\|\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}(X)\|_F \leq \psi\|X\|_F.$$

Proof: Let $T \in \mathbb{R}^{n\times n}$ be such that

$$T_{ij} = \begin{cases} 1 & \text{if } i = j,\ i \in \mathcal{I};\\ 0 & \text{otherwise.}\end{cases}$$

We then expand $\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}(X)$, which equals

$$\begin{aligned}&XT\overline{V}\overline{V}^\top T = XT\overline{V}\overline{V}^\top T^\top\\ =\ & X(T\overline{V})(T\overline{V})^\top = X\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)^\top\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top).\end{aligned}$$

The last equality follows from $(T\overline{V})^\top = \mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)$. Since $\psi = \|G\|$ where $G = \mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)^\top$, we have

$$\|\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)^\top\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\| = \|\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)^\top\| = \psi.$$

Now consider the i^{th} row of X, denoted as $\mathbf{x}^i$. Since $\|\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)^\top\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\| = \psi$, we have

$$\|\mathbf{x}^i\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)^\top\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\|_2^2 \leq \psi^2\|\mathbf{x}^i\|_2^2.$$

The lemma holds from the following inequality.

$$\begin{aligned}&\|\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}(X)\|_F^2 = \|X\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)^\top\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\|_F^2\\ =&\sum_i \|\mathbf{x}_i\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)^\top\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)\|_2^2\\ \leq&\psi^2\sum\|\mathbf{x}^i\|_2^2 = \psi^2\|X\|_F^2.\end{aligned}$$

■

Let $N_L = L' - \hat{L}$, $N_C = C' - \hat{C}$ and $E = N_C + N_L$. Then

$$\begin{aligned}&\|E\|_F \leq \|L' + C' - M\|_F \leq \|L' + C' - (M' - N)\|_F\\ &\leq \|L' + C' - M'\|_F + \|N\|_F \leq 2\epsilon.\end{aligned}$$

Further, define $N_L^+ = N_L - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}(N_L)$, $N_C^+ = N_C - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}(N_C)$, and $E^+ = E - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}(E)$. Observe that for any A, $\|(I - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0})(A)\|_F \leq \|A\|_F$.

Choosing the same W and F as in the proof of Theorem 3, we have

$$\begin{aligned}&\|\hat{L}\|_* + \lambda\|\hat{C}\|_{1,2} \geq \|L'\|_* + \lambda\|C'\|_{1,2}\\ \geq&\|\hat{L}\|_* + \lambda\|\hat{C}\|_{1,2} + \langle\mathcal{P}_{T(\hat{L})}(Q) + W, N_L\rangle\\ &+ \lambda\langle\mathcal{P}_{\mathcal{I}_0}(Q)/\lambda + F, N_C\rangle\\ =&\|\hat{L}\|_* + \lambda\|\hat{C}\|_{1,2} + \|\mathcal{P}_{T(\hat{L})^\perp}(N_L)\|_* + \lambda\|\mathcal{P}_{\mathcal{I}_0^c}(N_C)\|_{1,2}\\ &+ \langle\mathcal{P}_{T(\hat{L})}(Q), N_L\rangle + \langle\mathcal{P}_{\mathcal{I}_0}(Q), N_C\rangle\\ =&\|\hat{L}\|_* + \lambda\|\hat{C}\|_{1,2} + \|\mathcal{P}_{T(\hat{L})^\perp}(N_L)\|_* + \lambda\|\mathcal{P}_{\mathcal{I}_0^c}(N_C)\|_{1,2}\\ &- \langle\mathcal{P}_{T(\hat{L})^\perp}(Q), N_L\rangle - \langle\mathcal{P}_{\mathcal{I}_0^c}(Q), N_C\rangle + \langle Q, N_L + N_C\rangle\\ \geq&\|\hat{L}\|_* + \lambda\|\hat{C}\|_{1,2} + (1 - \|\mathcal{P}_{T(\hat{L})^\perp}(Q)\|)\|\mathcal{P}_{T(\hat{L})^\perp}(N_L)\|_*\\ &+ (\lambda - \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2})\|\mathcal{P}_{\mathcal{I}_0^c}(N_C)\|_{1,2} + \langle Q, E\rangle\\ \geq&\|\hat{L}\|_* + \lambda\|\hat{C}\|_{1,2} + (1/2)\|\mathcal{P}_{T(\hat{L})^\perp}(N_L)\|_*\\ &+ (\lambda/2)\|\mathcal{P}_{\mathcal{I}_0^c}(N_C)\|_{1,2} - 2\epsilon\|Q\|_F.\end{aligned}$$

Note that $\|Q\|_{\infty,2} \leq \lambda$, hence $\|Q\|_F \leq \sqrt{n}\lambda$. Thus we have

$$\begin{aligned} &\|\mathcal{P}_{T(\hat{L})^\perp}(N_L)\|_F \leq \|\mathcal{P}_{T(\hat{L})^\perp}(N_L)\|_* \leq 4\lambda\sqrt{n}\epsilon; \\ &\|\mathcal{P}_{\mathcal{I}_0^c}(N_C)\|_F \leq \|\mathcal{P}_{\mathcal{I}_0^c}(N_C)\|_{1,2} \leq 4\sqrt{n}\epsilon. \end{aligned} \tag{16}$$

Furthermore,

$$\begin{aligned} &\mathcal{P}_{\mathcal{I}_0}(N_C^+) \\ =&\mathcal{P}_{\mathcal{I}_0}(N_C) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(N_C) \\ =&\mathcal{P}_{\mathcal{I}_0}(E) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})^\perp}(N_L) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}(N_L) \\ &- \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(N_C) \\ =&\mathcal{P}_{\mathcal{I}_0}(E) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})^\perp}(N_L) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}(E) \\ &+ \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}(N_C) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(N_C) \\ =&\mathcal{P}_{\mathcal{I}_0}(E) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})^\perp}(N_L) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}(E) \\ &+ \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0^c}(N_C) + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0}(N_C) \\ &- \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(N_C) \\ \overset{(a)}{=}&\mathcal{P}_{\mathcal{I}_0}(E) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})^\perp}(N_L) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}(E) \\ &+ \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0^c}(N_C) + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0}(N_C^+) \\ \overset{(b)}{=}&\mathcal{P}_{\mathcal{I}_0}(E) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})^\perp}(N_L) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}(E) \\ &+ \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0^c}(N_C) + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}(N_C^+). \end{aligned} \tag{17}$$

Here (a) holds due to the following

$$\begin{aligned} &\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0}(N_C^+) \\ =&\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0}(N_C) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0}(\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}(N_C)) \\ =&\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0}(N_C) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(N_C), \end{aligned}$$

and (b) holds since by definition, each column of N_C^+ is orthogonal to U_0, hence $\mathcal{P}_{U_0}\mathcal{P}_{\mathcal{I}_0}(N_C^+) = 0$. Thus, Equation (17) leads to

$$\begin{aligned} &\|\mathcal{P}_{\mathcal{I}_0}(N_C^+)\|_F \\ \leq&\|\mathcal{P}_{\mathcal{I}_0}(E) - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}(E)\|_F + \|\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})^\perp}(N_L)\|_F \\ &+ \|\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{T(\hat{L})}\mathcal{P}_{\mathcal{I}_0^c}(N_C)\|_F + \|\mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\overline{V}}\mathcal{P}_{\mathcal{I}_0}(N_C^+)\|_F \\ \leq&\|E\|_F + \|\mathcal{P}_{T(\hat{L})^\perp}(N_L)\|_F + \|\mathcal{P}_{\mathcal{I}_0^c}(N_C)\|_F + \psi\|\mathcal{P}_{\mathcal{I}_0}(N_C^+)\|_F \\ \leq&(2 + 4\lambda\sqrt{n} + 4\sqrt{n})\epsilon + \psi\|\mathcal{P}_{\mathcal{I}_0}(N_C^+)\|_F. \end{aligned}$$

This implies that

$$\|\mathcal{P}_{\mathcal{I}_0}(N_C^+)\|_F \leq (2 + 4\lambda\sqrt{n} + 4\sqrt{n})\epsilon/(1-\psi).$$

Now using the fact that $\lambda < 1$, and $\psi < 1/4$, we have

$$\begin{aligned} \|N_C^+\|_F &= \|\mathcal{P}_{\mathcal{I}_0^c}(N_C) + \mathcal{P}_{\mathcal{I}_0}(N_C^+)\|_F \\ &\leq \|\mathcal{P}_{\mathcal{I}_0^c}(N_C)\|_F + \|\mathcal{P}_{\mathcal{I}_0}(N_C^+)\|_F \leq 18\sqrt{n}\epsilon. \end{aligned}$$

Note that $N_C^+ = (I - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0})(C' - \hat{C}) = C' - [\hat{C} + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}(C' - \hat{C})]$. Letting $\tilde{C} = \hat{C} + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}(C' - \hat{C})$, we have $\tilde{C} \in \mathcal{P}_{\mathcal{I}_0}$ and $\|C' - \tilde{C}\|_F \leq 18\sqrt{n}\epsilon$. Letting $\tilde{L} = \hat{L} - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}(C' - \hat{C})$, we have that $\tilde{L}, \tilde{C}$ is a successful decomposition, and

$$\begin{aligned} \|L' - \tilde{L}\|_F &\leq \|L' - \hat{L} + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{U_0}(C' - \hat{C})\|_F \\ &= \|L' - \hat{L} + \tilde{C} - \hat{C}\|_F \\ &= \|(L' - \hat{L} + C' - \hat{C}) + \tilde{C} - C'\|_F \\ &\leq \|E\|_F + \|C' - \tilde{C}\|_F \leq 20\sqrt{n}\epsilon. \end{aligned}$$

■

Remark: From the proof of Theorem 4, we have that Condition (15) holds when

$$\frac{\gamma}{1-\gamma} \leq \frac{(1-\psi)^2}{(9-4\psi)^2\mu_0 r}$$

and

$$\frac{2(1-\psi)\sqrt{\frac{\mu_0 r}{1-\gamma}}}{\sqrt{n}(1-\psi-\sqrt{\frac{\gamma}{1-\gamma}\mu_0 r})} \leq \lambda \leq \frac{1-\psi}{2(2-\psi)\sqrt{n\gamma}}.$$

For example, one can take

$$\lambda = \frac{\sqrt{9 + 1024\mu_0 r}}{14\sqrt{n}},$$

and all conditions of Theorem 5 hold when

$$\frac{\gamma}{1-\gamma} \leq \frac{9}{1024\mu_0 r}.$$

This establishes Theorem 2.

Remark: Notice that the subspace of the singular vectors corresponding to the r largest singular values of L', denoted $\mathbb{S}_{U'}$, can not deviate far away from the original column space $\mathbb{S}_{U_0}$. Indeed, applying a result from [35] (see for example Theorem 4 of [36], also [37]), we have that the Canonical Angle matrix Θ (see for example [36], [37] for a definition) between $\mathbb{S}_{U'}$ and $\mathbb{S}_{U_0}$ satisfies

$$\|\sin(\Theta)\|_F \leq \frac{\sqrt{2}\|L' - \tilde{L}\|_F}{\sigma_r(\tilde{L})} \leq \frac{20\sqrt{2n}\epsilon}{\sigma_r(L_0)},$$

where $\sigma_r(\cdot)$ represents the r-th largest singular value of a matrix. Here the last inequality holds since $\mathcal{P}_{\mathcal{I}_0}(\tilde{L}) = \mathcal{P}_{\mathcal{I}_0}(\tilde{L}_0)$ and $\mathcal{P}_{\mathcal{I}}(\tilde{L}_0) = 0$, hence the singular value for the former is always larger than or equal to the latter.

VI. Implementation issues and numerical experiments

While minimizing the nuclear norm is known to be a semidefinite program, and can be solved using a general purpose SDP solver such as SDPT3 or SeDuMi, such a method does not scale well to large data-sets. In fact, the computational time becomes prohibitive even for modest problem sizes as small as hundreds of variables. Recently, a family of optimization algorithms known as *proximal gradient algorithms* have been proposed to solve optimization problems of the form

$$\text{minimize: } g(\mathbf{x}), \quad \text{subject to: } \mathcal{A}(\mathbf{x}) = \mathbf{b},$$

of which Outlier Pursuit is a special case. It is known that such algorithms converge with a rate of $O(k^{-2})$ where k is the number of variables, and significantly outperform interior point methods for solving SDPs in practice. Following this paradigm, we solve Outlier Pursuit with the following algorithm. The validity of the algorithm follows easily from [38], [39]. See also [40].

Here, $\mathfrak{L}_\epsilon(S)$ is the diagonal soft-thresholding operator: if $|S_{ii}| \leq \epsilon$, then it is set to zero, otherwise, we set $S_{ii} := S_{ii} - \epsilon \cdot \text{sgn}(S_{ii})$. Similarly, $\mathfrak{C}_\epsilon(C)$ is the column-wise thresholding

Input: $M \in \mathbb{R}^{m\times n}$, λ, $\delta := 10^{-5}$, $\eta := 0.9$, $\mu_0 := 0.99\|M\|_F$.
1) $L_{-1}, L_0 := 0^{m\times n}$; $C_{-1}, C_0 := 0^{m\times n}$, $t_{-1}, t_0 := 1$; $\bar{\mu} = \delta\mu$;
2) **while** not converged **do**
3) $Y_k^L := L_k + \frac{t_{k-1}-1}{t_k}(L_k - L_{k-1})$, $Y_k^C := C_k + \frac{t_{k-1}-1}{t_k}(C_k - C_{k-1})$;
4) $G_k^L := Y_k^L - \frac{1}{2}\big(Y_k^L + Y_k^C - M\big)$; $G_k^C := Y_k^C - \frac{1}{2}\big(Y_k^L + Y_k^C - M\big)$;
5) $(U, S, V) := \mathrm{svd}(G_k^L)$; $L_{k+1} := U\mathfrak{L}_{\frac{\mu_k}{2}}(S)V$;
6) $C_{k+1} := \mathfrak{C}_{\frac{\lambda\mu_k}{2}}(G_k^C)$;
7) $t_{k+1} := \frac{1+\sqrt{4t_k^2+1}}{2}$; $\mu_{k+1} := \max(\eta\mu_k.\bar{\mu})$; $k++$;
8) **end while**
Output: $L := L_k$, $C = C_k$.

operator: set C_i to zero if $\|C_i\|_2 \le \epsilon$, otherwise set $C_i := C_i - \epsilon C_i/\|C_i\|_2$.

We explore the performance of Outlier Pursuit on some synthetic and real-world data, and find that its performance is quite promising.[2] Our first experiment investigates the phase-transition property of Outlier Pursuit, using randomly generated synthetic data. Fix $n = p = 400$. For different r and number of outliers γn, we generated matrices $A \in \mathbb{R}^{p\times r}$ and $B \in \mathbb{R}^{(n-\gamma n)\times r}$ where each entry is an independent $\mathcal{N}(0,1)$ random variable, and then set $L^* := A \times B^\top$ (the "clean" part of M). Outliers, $C^* \in \mathbb{R}^{\gamma n\times p}$ are generated either *neutrally*, where each entry of C^* is *iid* $\mathcal{N}(0,1)$, or *adversarially*, where every column is an identical copy of a random Gaussian vector. Outlier Pursuit succeeds if $\hat{C} \in \mathcal{P}_{\mathcal{I}}$, and $\hat{L} \in \mathcal{P}_U$ with a tolerance of 0.1%, i.e., if $\|\mathcal{P}_{\mathcal{I}_0^c}(\hat{C})\|_F \le 0.001\|\mathcal{P}_{\mathcal{I}_0^c}(L_0)\|_F$, and the $r+1$-th singular value of $\hat{L}$ is small than 0.001 times the r-th singular value. The parameter value λ is set using cross-validation with the information of the correct rank and the number of outliers. We initialize λ as in Theorem 1 and perform a bisection. If the resulting $\hat{L}$ has more ranks than we expect, we decrease λ; similarly, if the number of non-zero columns of $\hat{C}$ is larger than we expect, we increase λ. At most 5 different λ are selected, before the algorithm claims failure.

Figure 1 shows the phase transition property. We represent success in gray scale, with white denoting success, and black failure. When outliers are random (easier case) Outlier Pursuit succeeds even when $r = 20$ with 100 outliers. In the adversarial case, Outlier Pursuit succeeds when $r \times \gamma \le c$, and fails otherwise, consistent with our theory's predictions. We then fix $r = \gamma n = 5$ and examine the outlier identification ability of Outlier Pursuit with noisy observations. We scale each outlier so that the ℓ_2 distance of the outlier to the span of true samples equals a pre-determined value s. Each true sample is thus corrupted with a Gaussian random vector with an ℓ_2 magnitude σ. We perform (noiseless) Outlier Pursuit on this noisy observation matrix, and claim that the algorithm successfully identifies outliers if for the resulting $\hat{C}$ matrix, $\|\hat{C}_j\|_2 < \|\hat{C}_i\|_2$ for all $j \notin \mathcal{I}$ and $i \in \mathcal{I}$, i.e., there exists a threshold value to separate out outliers. Figure 1 (c) shows the result: when $\sigma/s \le 0.3$ for the identical outlier case, and $\sigma/s \le 0.7$ for the random outlier case, Outlier Pursuit correctly identifies the outliers.

[2]We have learned that [41] has also performed some numerical experiments minimizing $\|\cdot\|_* + \lambda\|\cdot\|_{1,2}$, and found promising results.

We further study the case of decomposing M under incomplete observation, which is motivated by *robust collaborative filtering*: we generate M as before, but only observe each entry with a given probability (independently). Letting Ω be the set of observed entries, we solve

$$\begin{aligned} \text{Minimize:} \quad & \|L\|_* + \lambda\|C\|_{1,2}; \\ \text{Subject to:} \quad & \mathcal{P}_\Omega(L + C) = \mathcal{P}_\Omega(M). \end{aligned} \tag{18}$$

The same success condition is used. Figure 2 shows a very promising result: the successful decomposition rate under incomplete observation is close the the complete observation case even only 30% of entries are observed. Given this empirical result, a natural direction of future research is to understand theoretical guarantee of (18) in the incomplete observation case.

Next we report some experimental results on the USPS digit data-set. The goal of this experiment is to show that Outlier Pursuit can be used to identify anomalies within the dataset. We use the data from [42], and construct the observation matrix M as containing the first 220 samples of digit "1" and the last 11 samples of "7". The learning objective is to correctly identify all the "7's". Note that throughout the experiment, label information is unavailable to the algorithm, i.e., there is no training stage. Since the columns of digit "1" are not exactly low rank, an exact decomposition is not possible. Hence, we use the ℓ_2 norm of each column in the resulting C matrix to identify the outliers: a larger ℓ_2 norm means that the sample is more likely to be an outlier — essentially, we apply thresholding after C is obtained. Figure 3(a) shows the ℓ_2 norm of each column of the resulting C matrix. We see that all "7's" are indeed identified. However, two "1" samples (columns 71 and 137) are also identified as outliers, due to the fact that these two samples are written in a way that is different from the rest of the "1's" as shown in Figure 4. Under the same setup, we also simulate the case where only 80% of entries are observed. As Figure 3 (b) and (c) show, similar results as that of the complete observation case are obtained, i.e., all true "7's" and also "1's" No 71, No 177 are identified.

VII. Conclusion and Future Direction

This paper considers robust PCA from a matrix decomposition approach, and develops the Outlier Pursuit algorithm.

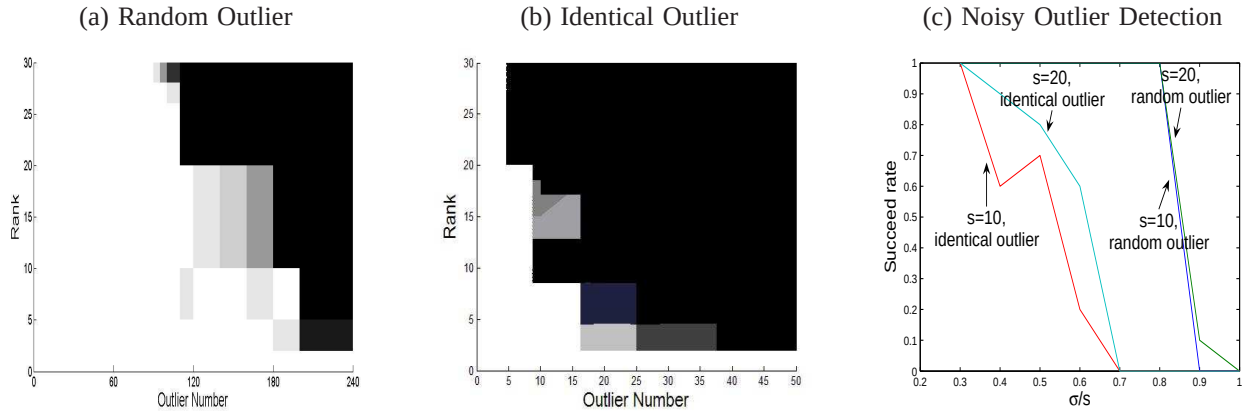

Fig. 1. This figure shows the performance of our algorithm in the case of complete observation (compare the next figure). The results shown represent an average over 10 trials.

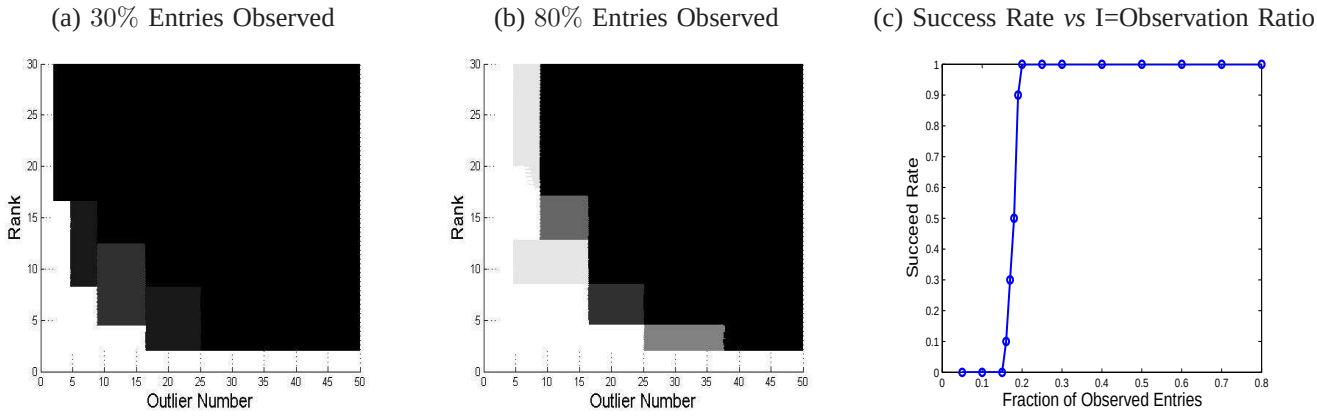

Fig. 2. This figure shows the case of partial observation, where only a fraction of the entries, sampled uniformly at random, are observed.

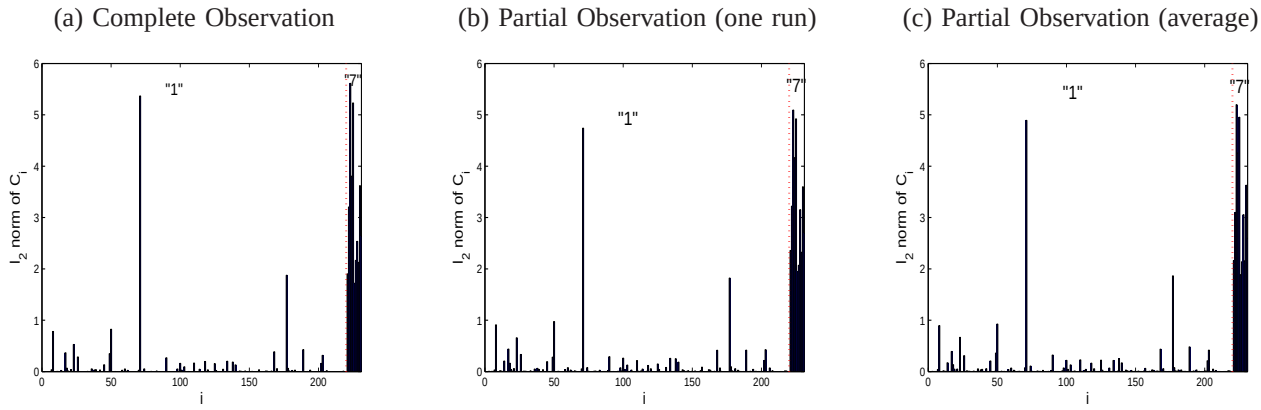

Fig. 3. This figure shows the ℓ_2 norm of each of the 220 columns of C. Large norm indicates that the algorithm believes that column is an outlier. All "7's" and two "1's" are identified as outliers.

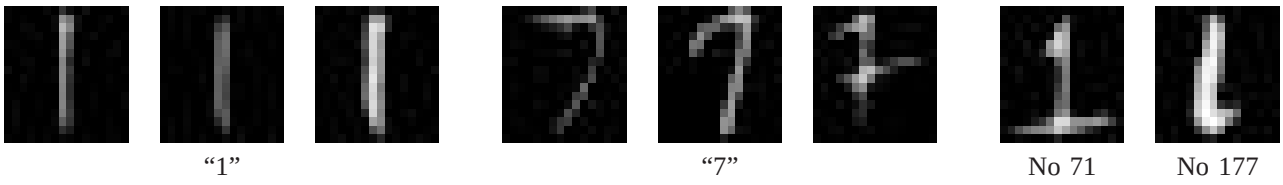

Fig. 4. This figure shows the typical "1's", the typical "7's" and also the two abnormal "1's" identified by the algorithm as outliers.

Under some mild conditions that are quite natural in most PCA settings, we show that Outlier Pursuit can exactly recover the column support, and exactly identify outliers. This result is new, differing both from results in Robust PCA, and also from results using nuclear-norm approaches for matrix completion and matrix reconstruction. One central innovation we introduce is the use of an oracle problem. Whenever the recovery concept (in this case, column space) does not uniquely correspond to a single matrix (we believe many, if not most cases of interest, fit this description), the use of such a tool will be quite useful. Immediate goals for future work include considering specific applications, in particular, robust collaborative filtering (here, the goal is to decompose a partially observed column-corrupted matrix) and also obtaining tight bounds for outlier identification in the noisy case. Indeed, in a subsequent paper [43], we, together with other co-authors, report some promising progress in the robust collaborative filtering setup, which essentially shows outlier pursuit provably succeeds in the partial observation case under reasonable technical conditions.

REFERENCES

[1] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. In *Advances in Neural Information Processing Systems*, 2010.

[2] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, Berlin: Springer, 1986.

[3] P. J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

[4] L. Xu and A. L. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, 6(1):131–143, 1995.

[5] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. ArXiv:0906.2220, 2009.

[6] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? To appear in *Journal of ACM*, 2011.

[7] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374):354–362, 1981.

[8] T. N. Yang and S. D. Wang. Robust algorithms for principal component analysis. *Pattern Recognition Letters*, 20(9):927–933, 1999.

[9] C. Croux and G. Hasebroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87(3):603–618, 2000.

[10] F. De la Torre and M. J. Black. Robust principal component analysis for computer vision. In *Proceedings of the Eighth International Conference on Computer Vision (ICCV'01)*, pages 362–369, 2001.

[11] F. De la Torre and M. J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1/2/3):117–142, 2003.

[12] C. Croux, P. Filzmoser, and M. Oliveira. Algorithms for Projection−Pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225, 2007.

[13] S. C. Brubaker. Robust PCA and clustering on noisy mixtures. In *Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, pages 1078–1087, 2009.

[14] D. L. Donoho. Breakdown properties of multivariate location estimators. Qualifying paper, Harvard University, 1982.

[15] R. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67, 1976.

[16] V. Barnett. The ordering of multivariate data. *Journal of Royal Statistics Society Series, A*, 138:318–344, 1976.

[17] D. Titterington. Estimation of correlation coefficients by ellipsoidal trimming. *Applied Statistics*, 27:227–234, 1978.

[18] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, New York, 1978.

[19] A. Dempster and M. Gasko-Green. New tools for residual analysis. *The Annals of Statistics*, 9(5):945–959, 1981.

[20] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62:531–545, 1975.

[21] G. Li and Z. Chen. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and monte carlo. *Journal of the American Statistical Association*, 80(391):759–766, 1985.

[22] C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226, 2005.

[23] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

[24] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. To appear in *SIAM Review*, 2010.

[25] S. Cotter, B. Rao, K. Engan, and K. Kreutz-delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Transaction on Signal Processing*, 2005.

[26] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[27] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

[28] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56:2053–2080, 2010.

[29] H. Xu, C. Caramanis, and S. Mannor. Principal component analysis with contaminated data: The high dimensional case. In *Proceeding of the Twenty-third Annual Conference on Learning Theory*, pages 490–502, 2010.

[30] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

[31] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[32] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

[33] G. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.

[34] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, N.J., 1970.

[35] P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[36] G. W. Stewart and G. W. Stewart. Perturbation theory for the singular value decomposition. In *in SVD and Signal Processing, II: Algorithms, Analysis and Applications*, pages 99–109. Elsevier, 1990.

[37] R. Bhatia. *Matrix Analysis*. Springer, 1997.

[38] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Submitted to SIAM Journal on Optimization, 2008.

[39] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(372-376), 1983.

[40] J-F. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20:1956–1982, 2008.

[41] M. McCoy and J. Tropp. Personal Correspondence, October 2010.

[42] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. the MIT Press, 2006.

[43] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi. Robust matrix completion with corrupted columns. In *Proceedings of Twenty-Eighth International Conference on Machine Learning*, 2011.

APPENDIX I
ORTHOGONAL CASE

This section investigates the special case where each outlier is orthogonal to the span of true samples, as stated in the following assumption.

Assumption 1: For $i \in \mathcal{I}_0$, $j \notin \mathcal{I}_0$, we have $M_i^\top M_j = 0$.

In the orthogonal case, we are able to derive a *necessary and sufficient* condition of Outlier Pursuit to succeed. Such condition is of course a necessary condition for Outlier Pursuit to succeed in the more general (non-orthogonal) case. Let

$$H_0 = \begin{cases} \frac{(C_0)_i}{\|(C_0)_i\|_2}, & \text{if } i \in \mathcal{I}_0; \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 6: Under Assumption 1, there exists a solution to Outlier Pursuit that correctly identifies the column space and outlier support, *if and only if*

$$\|H_0\| \leq 1/\lambda; \quad \|U_0 V_0^\top\|_{\infty,2} \leq \lambda. \tag{19}$$

If both inequalities hold strictly, then *any* solution to Outlier Pursuit correctly identifies the column space and outlier support.

Corollary 2: If the outliers are generated adversarial, and Assumption 1 holds, then Outlier Pursuit succeeds (for some λ^*) if and only if

$$\frac{\gamma}{1-\gamma} \leq \frac{1}{\mu r}.$$

Specifically, we can choose $\lambda^* = \sqrt{\frac{\mu r+1}{n}}$.

A. Proof of Theorem 6

The proof consists of three steps. We first show that if Outlier Pursuit succeeds, then (L_0, C_0) must be an optimal solution to Outlier Pursuit. Then using subgradient condition of optimal solutions to convex programming, we show that the necessary and sufficient condition for (L_0, C_0) being optimal solution is the existence of a dual certificate Q. Finally, we show that the existence of Q is equivalent to Condition (19) holds. We devote a subsection for each step.

1) Step 1: We need a technical lemma first.

Lemma 12: Given $A \in \mathbb{R}^{m\times n}$, we have

$$\|\mathcal{P}_{\mathcal{I}_0^c}(A)\|_* \leq \|A\|_*.$$

Proof: Fix $r \geq \mathrm{rank}(A)$. It is known that $\|A\|_*$ has the following variational form (Lemma 5.1 of [24]):

$$\begin{array}{rcll} \|A\|_* & = & \text{Minimize:}_{X\in\mathbb{R}^{m\times r}, Y\in\mathbb{R}^{n\times r}} & \frac{1}{2}(\|X\|_F^2 + \|Y\|_F^2) \\ & & \text{Subject to:} & XY^\top = A. \end{array} \tag{20}$$

Note that for any $XY^\top = A$, we have

$$X\overline{Y}^\top = X(\mathcal{P}_{\mathcal{I}_0^c}(Y^\top)) = \mathcal{P}_{\mathcal{I}^c}(A),$$

where $\overline{Y}$ is the matrix resulted by setting all *rows* of Y in $\mathcal{I}$ to zero. Thus, by variational form of $\|\mathcal{P}_{\mathcal{I}_0^c}(A)\|_*$, and note that $\mathrm{rank}(\mathcal{P}_{\mathcal{I}_0^c}(A)) \leq r$, we have

$$\|\mathcal{P}_{\mathcal{I}_0^c}(A)\|_* \leq \frac{1}{2}[\|X\|_F^2 + \|\overline{Y}\|_F^2] \leq \frac{1}{2}[\|X\|_F^2 + \|Y\|_F^2].$$

Note this holds for any X, Y such that $XY^\top = A$, the lemma follows from (20). ∎

Theorem 7: Under Assumption 1, for any L', C' such that $L' + C' = M$, $\mathcal{P}_{\mathcal{I}_0}(C') = C'$, and $\mathcal{P}_{U_0}(L') = L'$, we have

$$\|L_0\|_* + \lambda\|C_0\|_{1,2} \leq \|L'\|_* + \lambda\|C'\|_{1,2},$$

with the equality holds only when $L' = L_0$ and $C' = C_0$.

Proof: Write $L' = L_0 + \Delta$ and $C' = C_0 - \Delta$. Since $\mathcal{P}_{U_0}(L') = L'$, we have that for $i \in \mathcal{I}_0$, $\mathcal{P}_{U_0}(\Delta_i) = \Delta_i$, which implies that for $i \in \mathcal{I}_0$

$$C_{0i}^\top \Delta_i = (C_{0i}^\top U_0) U_0^\top \Delta_i = 0 \times U_0^\top \Delta_i,$$

where the last equality holds from Assumption 1 and the definition of C_0 (recall that C_{0i} is the i^{th} column of C_0). Thus, $\|C_0\|_{1,2} = \sum_{i\in\mathcal{I}} \|C_{0i}\|_2 \leq \sum_{i\in\mathcal{I}_0} \|C_{0i} + \Delta_i\|_2 \leq \sum_{i=1}^n \|C_{0i} + \Delta_i\|_2 = \|C'\|_{1,2}$, with equality only holds when $\Delta = 0$.

Further note that $\mathcal{P}_{\mathcal{I}_0}(C') = C'$ implies that $\mathcal{P}_{\mathcal{I}_0}(\Delta) = \Delta$, which by definition of L_0 leads to

$$L_0 = \mathcal{P}_{\mathcal{I}_0^c}(L').$$

Thus, Lemma 12 implies $\|L_0\|_* \leq \|L'\|_*$. The theorem thus follows. ∎

Note that Theorem 7 essentially says that in the orthogonal case, if Outlier Pursuit succeeds, i.e., it outputs a pair (L', C') such that L' has the correct column space, and C' has the correct column support, then (L_0, C_0) must be the output. This makes it possible to restrict out attention to investigate when the solution to Outlier Pursuit is (L_0, C_0).

2) Step 2:

Theorem 8: Under Assumption 1, (L_0, C_0) is an optimal solution to Outlier Pursuit if and only if there exists Q such that

$$\begin{array}{ll} (a) & \mathcal{P}_{T_0}(Q) = U_0 V_0^\top; \\ (b) & \|\mathcal{P}_{T_0^\perp}(Q)\| \leq 1; \\ (c) & \mathcal{P}_{\mathcal{I}_0}(Q) = \lambda H_0; \\ (d) & \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2} \leq \lambda. \end{array} \tag{21}$$

Here $\mathcal{P}_{T_0}(\cdot) \triangleq \mathcal{P}_{T(L_0)}(\cdot)$. In addition, if both inequalities are strict, then (L_0, C_0) is the unique optimal solution.

Proof: Standard convex analysis yields that (L_0, C_0) is an optimal solution to Outlier Pursuit if and only if there exists a dual matrix Q such that

$$Q \in \partial\|L_0\|_*; \quad Q \in \partial\lambda\|C_0\|_{1,2}.$$

Note that a matrix Q is a subgradient of $\|\cdot\|_*$ evaluated at L_0 if and only if it satisfies

$$\mathcal{P}_{T_0}(Q) = U_0 V_0^\top; \quad \text{and} \quad \|\mathcal{P}_{T_0^\perp}(Q)\| \leq 1.$$

Similarly, Q is a subgradient of $\lambda\|\cdot\|_{1,2}$ evaluated at C_0 if and only if

$$\mathcal{P}_{\mathcal{I}_0}(Q) = \lambda H_0; \quad \text{and} \quad \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2} \leq \lambda.$$

Thus, we conclude the proof of the first part of the theorem, i.e., the necessary and sufficient condition of (L_0, C_0) being an optimal solution.

Next we show that if both inequalities are strict, then (L_0, C_0) is the unique optimal solution. Fix $\Delta \neq 0$, we show that $(L_0 + \Delta, C_0 - \Delta)$ is strictly worse than (L_0, C_0). Let W be such that $\|W\| = 1$, $\langle W, \mathcal{P}_{T_0^\perp}(\Delta)\rangle = \|\mathcal{P}_{T_0^\perp}\Delta\|_*$, and $\mathcal{P}_{T_0}W = 0$. Let F be such that such that

$$F_i = \begin{cases} \frac{-\Delta_i}{\|\Delta_i\|_2} & \text{if } i \notin \mathcal{I}_0, \text{ and } \Delta_i \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then $U_0V_0^\top + W$ is a subgradient of $\|\cdot\|_*$ at L_0 and $H_0 + F$ is a subgradient of $\|\cdot\|_{1,2}$ at C_0. Then we have

$$\begin{aligned}
&\|L_0+\Delta\|_* + \lambda\|C_0-\Delta\|_{1,2}\\
\geq&\|L_0\|_* + \lambda\|C_0\|_{1,2} + < U_0V_0^\top + W, \Delta >\\
&- \lambda < H_0 + F, \Delta >\\
=&\|L_0\|_* + \lambda\|C_0\|_{1,2} + \|\mathcal{P}_{T_0^\perp}(\Delta)\|_* + \lambda\|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2}\\
&+ < U_0V_0^\top - \lambda H_0, \Delta >\\
=&\|L_0\|_* + \lambda\|C_0\|_{1,2} + \|\mathcal{P}_{T_0^\perp}(\Delta)\|_* + \lambda\|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2}\\
&+ < Q - \mathcal{P}_{T_0^\perp}(Q) - (Q - \mathcal{P}_{\mathcal{I}_0^c}(Q)), \Delta >\\
=&\|L_0\|_* + \lambda\|C_0\|_{1,2} + \|\mathcal{P}_{T_0^\perp}(\Delta)\|_* + \lambda\|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2}\\
&+ < -\mathcal{P}_{T_0^\perp}(Q), \Delta > + < \mathcal{P}_{\mathcal{I}_0^c}(Q), \Delta >\\
\geq&\|L_0\|_* + \lambda\|C_0\|_{1,2} + (1 - \|\mathcal{P}_{T_0^\perp}(Q)\|)\|\mathcal{P}_{T_0^\perp}(\Delta)\|_*\\
&+ (\lambda - \|\mathcal{P}_{\mathcal{I}_0^c}(Q)\|_{\infty,2})\|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2}\\
\geq&\|L_0\|_* + \lambda\|C_0\|_{1,2},
\end{aligned}$$

where the last inequality is strict unless

$$\|\mathcal{P}_{T_0^\perp}(\Delta)\|_* = \|\mathcal{P}_{\mathcal{I}_0^c}(\Delta)\|_{1,2} = 0. \tag{22}$$

We next show that Condition (22) also implies a strict increase of the objective function to complete the proof. Note that Equation (22) is equivalent to $\Delta = \mathcal{P}_{T_0}(\Delta) = \mathcal{P}_{\mathcal{I}_0}(\Delta)$, and note that

$$\begin{aligned}
\mathcal{P}_{U_0}(\Delta) &= \mathcal{P}_{T_0}(\Delta) - \mathcal{P}_{V_0}(\Delta) + \mathcal{P}_{U_0}\mathcal{P}_{V_0}(\Delta)\\
&= \Delta - (I - \mathcal{P}_{U_0})\mathcal{P}_{V_0}\Delta.
\end{aligned}$$

Since $\mathcal{P}_{\mathcal{I}_0}(V_0^\top) = 0$, $\mathcal{P}_{\mathcal{I}_0}(\Delta) = \Delta$ implies that $\mathcal{P}_{V_0}(\Delta) = 0$, which means

$$\Delta = \mathcal{P}_{U_0}(\Delta) = \mathcal{P}_{\mathcal{I}_0}(\Delta).$$

Thus, $\mathcal{P}_{U_0}(L_0+\Delta) = L_0+\Delta$, and $\mathcal{P}_{\mathcal{I}_0}(C_0-\Delta) = C_0-\Delta$. By Theorem 7, $\|L_0+\Delta\|_* + \lambda\|C_0-\Delta\|_{1,2} > \|L_0\|_* + \lambda\|C_0\|_{1,2}$, which completes the proof. ■

3) Step 3:

Theorem 9: Under Assumption 1, if there exists any matrix Q that satisfies Condition (21), then $U_0V_0^\top + \lambda H_0$ satisfies (21).

Proof: Denote $Q_0 \triangleq U_0V_0^\top + \lambda H_0$. We first show that the two equalities of Condition (21) hold. Note that

$$\begin{aligned}
&\mathcal{P}_{T_0}(Q_0) = \mathcal{P}_{T_0}(U_0V_0^\top) + \lambda\mathcal{P}_{T_0}(H_0)\\
&= U_0V_0^\top + \lambda[\mathcal{P}_{U_0}(H_0) + \mathcal{P}_{V_0}(H_0) - \mathcal{P}_{U_0}\mathcal{P}_{V_0}(H_0)].
\end{aligned}$$

Further note that $\mathcal{P}_{U_0}(H_0) = U_0(U_0^\top H_0) = 0$ due to Assumption 1, and $\mathcal{P}_{V_0}(H_0) = 0$ because $\mathcal{P}_{\mathcal{I}_0}(H_0) = H_0$ and $\mathcal{P}_{\mathcal{I}_0}(V_0^\top) = 0$ lead to $H_0V_0 = 0$. Hence

$$\mathcal{P}_{T_0}(Q_0) = U_0V_0^\top.$$

Furthermore,

$$\begin{aligned}
&\mathcal{P}_{\mathcal{I}_0}(Q_0) = \mathcal{P}_{\mathcal{I}_0}(U_0V_0^\top) + \lambda\mathcal{P}_{\mathcal{I}_0}(H_0)\\
&= U_0\mathcal{P}_{\mathcal{I}_0}(V_0^\top) + \lambda H_0 = \lambda H_0.
\end{aligned}$$

Here, the last equality holds because $\mathcal{P}_{\mathcal{I}_0}(V_0^\top) = 0$. Note that this also implies that

$$\mathcal{P}_{T_0^\perp}(H_0) = H_0; \quad \mathcal{P}_{\mathcal{I}_0^c}(U_0V_0^\top) = U_0V_0^\top. \tag{23}$$

Now consider any matrix Q that also satisfies the two equalities. Let $Q = U_0V_0^\top + \lambda H_0 + \Delta$, note that Q satisfies $\mathcal{P}_{\mathcal{I}_0}(Q) = \lambda H_0$ and $\mathcal{P}_{\mathcal{T}_0}(Q) = U_0V_0^\top$, which leads to

$$\mathcal{P}_{\mathcal{I}_0}(\Delta) = 0; \quad \text{and } \mathcal{P}_{\mathcal{T}_0}(\Delta) = 0.$$

Thus,

$$\mathcal{P}_{\mathcal{I}_0^c}(Q) = U_0V_0^\top + \Delta; \quad \text{and} \quad \mathcal{P}_{\mathcal{T}_0^\perp}(Q) = \lambda H_0 + \Delta.$$

Note that

$$\begin{aligned}
&\|U_0V_0^\top + \Delta\|_{\infty,2} = \max_i \|U_0(V_0^\top)_i + \Delta_i\|_2\\
&\geq \max_i \|U_0(V_0^\top)_i\|_2 = \|U_0V_0^\top\|_{\infty,2}.
\end{aligned}$$

Here, the inequality holds because $\mathcal{P}_{\mathcal{T}_0}(\Delta) = 0$ implies that Δ_i are orthogonal to the span of U_0. Note that the inequality is strict when $\Delta \neq 0$.

On the other hand

$$\begin{aligned}
\|\lambda H_0\| &= \max_{\|\mathbf{x}\|\leq 1, \|\mathbf{y}\|\leq 1} \mathbf{x}^\top(\lambda H_0)\mathbf{y}\\
&\overset{(a)}{=} \max_{\|\mathbf{x}\|\leq 1, \|\mathbf{y}\|\leq 1, \mathcal{P}_{\mathcal{I}_0^c}(\mathbf{y}^\top)=0} \mathbf{x}^\top(\lambda H_0)\mathbf{y}\\
&\overset{(b)}{=} \max_{\|\mathbf{x}\|\leq 1, \|\mathbf{y}\|\leq 1, \mathcal{P}_{\mathcal{I}_0^c}(\mathbf{y}^\top)=0} \mathbf{x}^\top(\lambda H_0 + \Delta)\mathbf{y}\\
&\leq \max_{\|\mathbf{x}\|\leq 1, \|\mathbf{y}\|\leq 1} \mathbf{x}^\top(\lambda H_0 + \Delta)\mathbf{y} = \|\lambda H_0 + \Delta\|.
\end{aligned}$$

Here, (a) holds because $\mathcal{P}_{\mathcal{I}_0}H_0 = H_0$, thus for any $\mathbf{y}$, set all $y_i = 0$ for $i \notin \mathcal{I}_0$ does not change $\mathbf{x}^\top(\lambda H_0)\mathbf{y}$; while (b) holds since $\mathcal{P}_{\mathcal{I}_0^c}\Delta = \Delta$.

Thus, if Q satisfies the two inequalities, then so does Q_0, which completes the proof. ■

Note that by Equation (23) we have

$$\mathcal{P}_{T_0^\perp}(H_0) = H_0; \quad \mathcal{P}_{\mathcal{I}_0^c}(U_0V_0^\top) = U_0V_0^\top.$$

Thus, Theorem 7, Theorem 8 and Theorem 9 together establish Theorem 6.

B. Proof of Corollary 2

Corollary 2 holds due to the following lemma that tightly bounds $\|H_0\|$ and $\|U_0V_0^\top\|_{\infty,2}$.

Lemma 13: We have (I) $\|H_0\| \leq \sqrt{\gamma n}$, and the inequality is tight. (II) $\|U_0V_0^\top\|_{\infty,2} = \max_i \|V_0^\top\mathbf{e}_i\|_2 = \sqrt{\frac{\mu r}{(1-\gamma)n}}$.

Proof: Following the variational form of the operator norm, we have

$$\begin{aligned}
\|H_0\| &= \max_{\|\mathbf{x}\|_2\leq 1, \|\mathbf{y}\|_2\leq 1} \mathbf{x}^\top H_0\mathbf{y} = \max_{\|\mathbf{x}\|_2\leq 1} \|\mathbf{x}^\top H_0\|_2\\
&= \max_{\|\mathbf{x}\|_2\leq 1} \sqrt{\sum_{i=1}^n (\mathbf{x}^\top H_i)^2} \leq \sqrt{\sum_{i\in\mathcal{I}_0} 1} = \sqrt{|\mathcal{I}_0|} = \sqrt{\gamma n}.
\end{aligned}$$

The inequality holds because $\|(H_0)_i\|_2 = 1$ when $i \in \mathcal{I}_0$, and equals zero otherwise. Note that if we let $(H_0)_i$ all be the same, such as taking identical outliers, the inequality is tight.

By definition we have $\|U_0V_0^\top\|_{\infty,2} = \max_i \|U_0(V_0^\top)_i\|_2 \overset{(a)}{=} \max_i \|(V_0^\top)_i\|_2 = \max_i \|V_0^\top\mathbf{e}_i\|_2$. Here (a) holds since U_0 is orthonormal. The second claim hence follows from definition of μ. ■

APPENDIX II
LIST OF NOTATIONS

M	The observed matrix.
p	The number of rows of M.
n	The number of columns of M.
L_0, C_0	The ground truth.
$\mathcal{I}_0$	The index of outliers (non-zero columns of C_0).
γ	Fraction of outliers, which equals $\lvert\mathcal{I}_0\rvert/n$.
U_0, V_0	The left and right singular vectors of L_0.
μ	Incoherence parameter of V_0
$\hat{L}, \hat{C}$	The optimal solution of the Oracle Problem.
$\hat{U}, \hat{V}$	The left and right singular vectors of $\hat{L}$.
$\overline{V}$	An auxiliary matrix, introduced in Lemma 5, which satisfies $\hat{U}\hat{V}^\top = U_0\overline{V}^\top$.
$\overline{\mu}$	Incoherence parameter of $\overline{V}$.
$\hat{H}$	An auxiliary matrix, introduced in Lemma 5, which satisfies $\hat{H} \in \mathfrak{G}(\hat{C})$.
$\mathfrak{N}(\cdot)$ $\mathfrak{G}(\cdot)$	Operators defined in Definition 1.
G	Auxiliary matrix defined in Equation (11), as $G \triangleq \mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top)(\mathcal{P}_{\mathcal{I}_0}(\overline{V}^\top))^\top$.
ψ	Defined in Lemma 7 as $\psi = \lVert G\rVert$.

Huan Xu received the B.Eng. degree in automation from Shanghai Jiaotong University, Shanghai, China in 1997, the M.Eng. degree in electrical engineering from the National University of Singapore in 2003, and the Ph.D. degree in electrical engineering from McGill University, Canada in 2009. From 2009 to 2010, he was a postdoctoral associate at The University of Texas at Austin.

Since 2011, he has been an assistant professor at the Department of Mechanical Engineering at the National University of Singapore. His research interests include statistics, machine learning, robust optimization, and planning and control.

Constantine Caramanis (M'06) received his Ph.D. in EECS from the Massachusetts Institute of Technology in 2006. Since then, he has been on the faculty in Electrical and Computer Engineering at The University of Texas at Austin. He received the NSF CAREER award in 2011. His current research interests include robust and adaptable optimization, machine learning and high-dimensional statistics, with applications to large scale networks.

Sujay Sanghavi (M'06) is on the faculty of Electrical Engineering at the University of Texas, Austin. He obtained his PhD in 2006 from the University of Illinois, and a postdoc from the Massachusetts Institute of Technology. His research interests are in the use of probability, optimization and algorithms for applications in large-scale networks and high-dimensional machine learning. Sujay received the NSF CAREER award in 2010.