Exact Recoverability of Robust PCA via Outlier Pursuit with Tight Recovery Bounds

Hongyang Zhang[†], Zhouchen Lin[†*], Chao Zhang[†*], Edward Y. Chang[‡]
[†]Key Lab. of Machine Perception (MOE), School of EECS, Peking University, China
[‡] HTC Research, Taiwan
{hy_zh, zlin}@pku.edu.cn, chzhang@cis.pku.edu.cn, edward_chang@htc.com

Abstract

Subspace recovery from noisy or even corrupted data is critical for various applications in machine learning and data analysis. To detect outliers, Robust PCA (R-PCA) via Outlier Pursuit was proposed and had found many successful applications. However, the current theoretical analysis on Outlier Pursuit only shows that it succeeds when the sparsity of the corruption matrix is of $O(n/r)$, where n is the number of the samples and r is the rank of the intrinsic matrix which may be comparable to n. Moreover, the regularization parameter is suggested as $3/(7\sqrt{\gamma n})$, where γ is a parameter that is not known a priori. In this paper, with incoherence condition and proposed ambiguity condition we prove that Outlier Pursuit succeeds when the rank of the intrinsic matrix is of $O(n/\log n)$ and the sparsity of the corruption matrix is of $O(n)$. We further show that the orders of both bounds are tight. Thus R-PCA via Outlier Pursuit is able to recover intrinsic matrix of higher rank and identify much denser corruptions than what the existing results could predict. Moreover, we suggest that the regularization parameter be chosen as $1/\sqrt{\log n}$, which is definite. Our analysis waives the necessity of tuning the regularization parameter and also significantly extends the working range of the Outlier Pursuit. Experiments on synthetic and real data verify our theories.

Introduction

It is well known that many real-world datasets, e.g., motion (Gear 1998; Yan and Pollefeys 2006; Rao et al. 2010), face (Liu et al. 2013), and texture (Ma et al. 2007), can be approximately characterized by low-dimensional subspaces. So recovering the intrinsic subspace that the data distribute on is a critical step in many applications in machine learning and data analysis. There has been a lot of work on robustly recovering the underlying subspace. Probably the most widely used one is Principal Component Analysis (PCA). Unfortunately, the standard PCA is known to be sensitive to outliers: even a single but severe outlier may degrade the effectiveness of the model. Note that such a type of data corruption commonly exists because of sensor failures, uncontrolled environments, etc.

To overcome the drawback of PCA, several efforts have been devoted to robustifying PCA (Croux and Haesbroeck 2000; De La Torre and Black 2003; Huber 2011), among which Robust PCA (R-PCA) (Wright et al. 2009; Candès et al. 2011; Chen et al. 2011; Xu, Caramanis, and Sanghavi 2012) is probably the most attractive one due to its theoretical guarantee. It has been shown (Wright et al. 2009; Candès et al. 2011; Chen et al. 2011; Xu, Caramanis, and Sanghavi 2012) that under mild conditions R-PCA exactly recovers the ground-truth subspace with an overwhelming probability. Nowadays, R-PCA has been applied to various tasks, such as video denoising (Ji et al. 2010), background modeling (Wright et al. 2009), subspace clustering (Zhang et al. 2014), image alignment (Peng et al. 2010), photometric stereo (Wu et al. 2011), texture representation (Zhang et al. 2012), and spectral clustering (Xia et al. 2014).

Related Work

R-PCA was first proposed by Wright et al. (2009) and Chandrasekaran et al. (2011). Suppose we are given a data matrix $M = L_0 + S_0 \in \mathbb{R}^{m\times n}$, where each column of M is an observed sample vector and L_0 and S_0 are the intrinsic and the corruption matrices, respectively, R-PCA recovers the ground-truth structure of the data by decomposing M into a low rank component L and a sparse one S. R-PCA achieves this goal by Principal Component Pursuit, formulated as follows (please refer to Table 1 for explanation of notations):

$$\min_{L,S} ||L||_* + \lambda||S||_1, \text{ s.t. } M = L + S. \quad (1)$$

Candès et al. (2011) proved that when the locations of nonzero entries (a.k.a. support) of S_0 are uniformly distributed, the rank of L_0 is no larger than $\rho_r n_{(2)}/(\log n_{(1)})^2$, and the number of nonzero entries (a.k.a. sparsity) of S_0 is less than $\rho_s mn$, Principal Component Pursuit with a regularization parameter $\lambda = 1/\sqrt{n_{(1)}}$ exactly recovers the ground truth matrices L_0 and S_0 with an overwhelming probability, where ρ_r and ρ_s are both numerical constants.

Although Principal Component Pursuit has been applied to many tasks, e.g., face repairing (Candès et al. 2011) and photometric stereo (Wu et al. 2011), it breaks down when the noises or outliers are distributed columnwise, i.e., large errors concentrate only on a number of columns of S_0 rather than scattering uniformly across S_0. Such a situation commonly occurs, e.g., in abnormal hand writing (Xu, Caramanis, and Sanghavi 2012) and traffic anomalies data (Liu et

*Corresponding authors.

al. 2013). To address this issue, Chen et al. (2011), McCoy and Tropp (2011), and Xu, Caramanis, and Sanghavi (2012) proposed replacing the ℓ_1 norm in model (1) with the $\ell_{2,1}$ norm, resulting in the following Outlier Pursuit model:

$$\min_{L,S} ||L||_* + \lambda ||S||_{2,1}, \text{ s.t. } M = L + S. \tag{2}$$

The formulation of model (2) looks similar to that of model (1). However, theoretical analysis on Outlier Pursuit is more difficult than that on Principal Component Pursuit. The most distinct difference is that we cannot expect Outlier Pursuit to exactly recover L_0 and S_0. Rather, only the *column space* of L_0 and the *column support* of S_0 can be exactly recovered (Chen et al. 2011; Xu, Caramanis, and Sanghavi 2012). This is because a corrupted sample can be addition of any vector in the column space of L_0 and another appropriate vector. This ambiguity cannot be resolved if no trustworthy information from the sample can be utilized. Furthermore, theoretical analysis on Outlier Pursuit (Chen et al. 2011; Xu, Caramanis, and Sanghavi 2012) imposed weaker conditions on the incoherence of the matrix, which makes the proofs harder to complete.

Nowadays, a large number of applications have testified to the validity of Outlier Pursuit, e,g., meta-search (Pan et al. 2013), image annotation (Dong et al. 2013), and collaborative filtering (Chen et al. 2011). However, current results for Outlier Pursuit are not satisfactory in some aspects:

- Chen et al. (2011) and Xu, Caramanis, and Sanghavi (2012) proved that Outlier Pursuit exactly recovers the column space of L_0 and identifies the column support of S_0 when the column sparsity (i.e., the number of the corrupted columns) of S_0 is of $O(n/r)$, where r is the rank of the intrinsic matrix. When r is comparable to n, the working range of Outlier Pursuit is limited.
- McCoy and Tropp (2011) suggested to choose λ within the range $[\sqrt{T/n}, 1]$ and Xu, Caramanis, and Sanghavi (2012) suggested to choose λ as $3/(7\sqrt{\gamma n})$, where unknown parameters, i.e., the number of principal components T and the outlier ratio γ, are involved. So a practitioner has to tune the unknown parameters by cross validation, which is inconvenient and time consuming.

Our Contributions

This paper is concerned with the exact recovery problem of Outlier Pursuit. Our contributions are as follows:

- With incoherence condition on the low rank term and proposed ambiguity condition on the column sparse term, we prove that Outlier Pursuit succeeds at an overwhelming probability when the rank of L_0 is no larger than $\tilde{\rho_r} n_{(2)}/\log n$ and the column sparsity of S_0 is no greater than $\tilde{\rho_s} n$, where $\tilde{\rho_r}$ and $\tilde{\rho_s}$ are numerical constants. We also demonstrate that the orders of both bounds are tight.
- We show that in theory $\lambda = 1/\sqrt{\log n}$ is suitable for model (2). This result is useful because it waives the necessity of tuning the regularization parameter in the existing work. We will show by experiments that our choice of λ not only works well but also extends the working range of Outlier Pursuit (see Figure 1).

Table 1: Summary of main notations used in this paper.

Notations	Meanings
m, n	Size of the data matrix M.
$n_{(1)}, n_{(2)}$	$n_{(1)} = \max\{m, n\}$, $n_{(2)} = \min\{m, n\}$.
$\Theta(n)$	Grows in the same order of n.
$O(n)$	Grows equal to or less than the order of n.
e_i	Vector whose ith entry is 1 and others are 0s.
$M_{:j}$	The jth column of matrix M.
M_{ij}	The entry at the ith row and jth column of M.
$\|\|\cdot\|\|_2$	ℓ_2 norm for vector, $\|\|v\|\|_2 = \sqrt{\sum_i v_i^2}$.
$\|\|\cdot\|\|_*$	Nuclear norm, the sum of singular values.
$\|\|\cdot\|\|_0$	ℓ_0 norm, number of nonzero entries.
$\|\|\cdot\|\|_{2,0}$	$\ell_{2,0}$ norm, number of nonzero columns.
$\|\|\cdot\|\|_1$	ℓ_1 norm, $\|\|M\|\|_1 = \sum_{i,j} \|M_{ij}\|$.
$\|\|\cdot\|\|_{2,1}$	$\ell_{2,1}$ norm, $\|\|M\|\|_{2,1} = \sum_j \|\|M_{:j}\|\|_2$.
$\|\|\cdot\|\|_{2,\infty}$	$\ell_{2,\infty}$ norm, $\|\|M\|\|_{2,\infty} = \max_j \|\|M_{:j}\|\|_2$.
$\|\|\cdot\|\|_F$	Frobenious norm, $\|\|M\|\|_F = \sqrt{\sum_{i,j} M_{ij}^2}$.
$\|\|\cdot\|\|_\infty$	Infinity norm, $\|\|M\|\|_\infty = \max_{ij} \|M_{ij}\|$.
$\|\|\mathcal{P}\|\|$	(Matrix) operator norm.
$\hat{U}, \hat{V}$	Left and right singular vectors of $\hat{L}$.
$\mathcal{U}_0, \hat{\mathcal{U}}, \mathcal{U}^*$	Column space of L_0, $\hat{L}$, L^*.
$\mathcal{V}_0, \hat{\mathcal{V}}, \mathcal{V}^*$	Row space of L_0, $\hat{L}$, L^*.
$\hat{\mathcal{T}}$	Space $\hat{\mathcal{T}}=\{\hat{U}X^T+Y\hat{V}^T, \forall X, Y \in \mathbb{R}^{n\times r}\}$.
$\mathcal{X}^\perp$	Orthogonal complement of the space $\mathcal{X}$.
$\mathcal{P}_{\hat{\mathcal{U}}}, \mathcal{P}_{\hat{\mathcal{V}}}$	$\mathcal{P}_{\hat{\mathcal{U}}} M = \hat{U}\hat{U}^T M$, $\mathcal{P}_{\hat{\mathcal{V}}} M = M\hat{V}\hat{V}^T$.
$\mathcal{P}_{\hat{\mathcal{T}}^\perp}$	$\mathcal{P}_{\hat{\mathcal{T}}^\perp} M = \mathcal{P}_{\hat{\mathcal{U}}^\perp} \mathcal{P}_{\hat{\mathcal{V}}^\perp} M$.
$\mathcal{I}_0, \hat{\mathcal{I}}, \mathcal{I}^*$	Index of outliers of S_0, $\hat{S}$, S^*.
$\mathcal{B}(\hat{S})$	Operator normalizing nonzero columns of $\hat{S}$, $\mathcal{B}(\hat{S})=\{H: \mathcal{P}_{\hat{\mathcal{I}}^\perp}(H)=0; H_{:j}=\frac{\hat{S}_{:j}}{\|\|\hat{S}_{:j}\|\|_2}, j \in \hat{\mathcal{I}}\}$.
$\sim$ Ber(p)	Bernoulli distribution with parameter p.
$\mathcal{N}(a, b^2)$	Gaussian distribution (mean a,variance b^2).

So in both aspects, we extend the working range of Outlier Pursuit. We validate our theoretical analysis by simulated and real experiments. The experimental results match our theoretical results nicely.

Problem Setup

In this section, we set up the problem by making some definitions and assumptions.

Exact Recovery Problem

This paper focuses on the exact recovery problem of Outlier Pursuit as defined below.

Definition 1 (Exact Recovery Problem)**.** *Suppose we are given an observed data matrix $M = L_0 + S_0 \in \mathbb{R}^{m\times n}$, where L_0 is the ground-truth intrinsic matrix and S_0 is the real corruption matrix with sparse nonzero columns, the exact recovery problem investigates whether the column space of L_0 and the column support of S_0 can be exactly recovered.*

A similar problem has been proposed for Principal Component Pursuit (Candès et al. 2011; Zhang, Lin, and Zhang 2013). However, Definition 1 for Outlier Pursuit has its own characteristic: one can only expect to recover the column space of L_0 and the column support of S_0, rather than the whole L_0 and S_0 themselves. This is because a corrupted sample can be addition of any vector in the column space of L_0 and another appropriate vector.

Incoherence Condition on Low Rank Term

In general, the exact recovery problem has an identifiability issue. As an extreme example, imagine the case where the low rank term has only one nonzero column. Such a matrix is both low rank and column sparse. So it is hard to identify whether this matrix is the low rank component or the column sparse one. Similar situation occurs for Principal Component Pursuit (Candès et al. 2011).

To resolve the identifiability issue, Candès et al. proposed the following three μ-incoherence conditions (Candès and Recht 2009; Candès and Tao 2010; Candès et al. 2011) for a matrix $L \in \mathbb{R}^{m\times n}$ with rank r:

$$\max_i ||V^T e_i||_2 \le \sqrt{\frac{\mu r}{n}}, \quad \text{(avoid column sparsity)} \tag{3a}$$

$$\max_i ||U^T e_i||_2 \le \sqrt{\frac{\mu r}{m}}, \quad \text{(avoid row sparsity)} \tag{3b}$$

$$||UV^T||_\infty \le \sqrt{\frac{\mu r}{mn}}, \tag{3c}$$

where the first two conditions are for avoiding a matrix to be column sparse and row sparse, respectively, and $U\Sigma V^T$ is the skinny SVD of L. As discussed in (Candès and Recht 2009; Candès and Tao 2010; Gross 2011), the incoherence conditions imply that for small values of μ, the singular vectors of the matrix are not sparse.

Chen et al. (2011) assumed conditions (3a) and (3b) for matrices L_0 and M. However, a row sparse matrix most likely should be the low rank component, rather than the column sparse one. So condition (3b) is actually redundant for recovering the column space of L_0, and we adopt condition (3a) and an ambiguity condition on the column sparse term (see the next subsection) instead of condition (3b). With our reasonable conditions, we are able to extend the working range of Outlier Pursuit.

Ambiguity Condition on Column Sparse Term

Similarly, the column sparse term S has the identifiability issue as well. Imagine the case where S has rank 1, $\Theta(n)$ columns are zeros, and other $\Theta(n)$ columns are nonzeros. Such a matrix is both low rank and sparse. So it is hard to identify whether S is the column sparse term or the low rank one. Therefore, Outlier Pursuit fails in such a case without any additional assumptions (Xu, Caramanis, and Sanghavi 2012). To resolve the issue, we propose the following ambiguity condition on S:

$$||\mathcal{B}(S)|| \le \sqrt{\log n}/4. \tag{4}$$

Note that the above condition is feasible, e.g., it holds if the nonzero columns of $\mathcal{B}(S)$ obey i.i.d. uniform distribution on the unit ℓ_2 sphere (Eldar and Kutyniok 2012). Moreover, the uniformity is not a necessary assumption. Geometrically, condition (4) holds as long as the directions of the nonzero columns of S scatter sufficiently randomly. Thus it guarantees that matrix S cannot be low rank when the column sparsity of S is comparable to n.

Probability Model

Our main results are based on the assumption that the column support of S_0 is uniformly distributed among all sets of cardinality s. Such an assumption is reasonable since we have no further information on the outlier positions. By the standard arguments in (Candès and Tao 2010) and (Candès et al. 2011), any guarantee proved for the Bernoulli distribution, which takes the value 0 with probability $1-p$ and the value 1 with probability p, equivalently holds for the uniform distribution of cardinality s, where $p = \Theta(1)$ implies $s = \Theta(n)$. Thus for convenience we assume $\mathcal{I}_0 \sim \text{Ber}(p)$. More specifically, we assume $[S_0]_{:j} = [\delta_0]_j [Z_0]_{:j}$ throughout our proof, where $[\delta_0]_j \sim \text{Ber}(p)$ determines the outlier positions and $[Z_0]_{:j}$ determines the outlier values. We also call any event which holds with a probability at least $1-\Theta(n^{-10})$ to happen with an overwhelming probability.

Main Results

Our theory shows that Outlier Pursuit (2) succeeds in the exact recovery problem under mild conditions, even though a fraction of the data are severely corrupted. We summarize our results in the following theorem.

Theorem 1 (Exact Recovery of Outlier Pursuit)**.** *Suppose $m = \Theta(n)$, Range(L_0) = Range$(\mathcal{P}_{\mathcal{I}_0^\perp} L_0)$, and $[S_0]_{:j} \notin$ Range(L_0) for $\forall j \in \mathcal{I}_0$. Then any solution (L_0+H, S_0-H) to Outlier Pursuit* (2) *with $\lambda = 1/\sqrt{\log n}$ exactly recovers the column space of L_0 and the column support of S_0 with a probability at least $1-cn^{-10}$, if the column support $\mathcal{I}_0$ of S_0 is uniformly distributed among all sets of cardinality s and*

$$\text{rank}(L_0) \le \rho_r \frac{n_{(2)}}{\mu \log n} \quad \text{and} \quad s \le \rho_s n, \tag{5}$$

where c, ρ_r, and ρ_s are constants, $L_0 + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0} H$ satisfies μ-incoherence condition (3a)*, and $S_0 - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0} H$ satisfies ambiguity condition* (4)*.*

The incoherence and ambiguity conditions on $\hat{L} = L_0 + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0} H$ and $\hat{S} = S_0 - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0} H$ are not surprising. Note that $\hat{L}$ has the same column space as Range(L_0) and $\hat{S}$ has the same column index as that of S_0. Also, notice that $\hat{L} + \hat{S} = M$. So it is natural to consider $\hat{L}$ and $\hat{S}$ as the underlying low-rank and sparse terms, i.e., M is constructed by $\hat{L} + \hat{S}$, and we assume incoherence and ambiguity conditions on them instead of L_0 and S_0.

Theorem 1 has several advantages: first, while the parameter λs in the previous literatures are related to some unknown parameters, such as the outlier ratio, our choice of parameter is simple and precise. Second, with incoherence or ambiguity conditions, we push the bound on the column sparsity of S_0 from $O(n/r)$ to $O(n)$, where r is the rank of L_0 (or the dimension of the underlying subspace) and may be comparable to n. The following theorem shows that our bounds in (5) are optimal, whose proof can be found in the supplementary material.

Theorem 2. *The orders of the upper bounds given by inequalities* (5) *are tight.*

Experiments also testify to the tightness of our bounds.

Outline of Proofs

In this section, we sketch the outline of proving Theorem 1. For the details of proofs, please refer to the supplementary materials. Without loss of generality, we assume $m = n$. The following theorem shows that Outlier Pursuit succeeds for easy recovery problem.

Theorem 3 (Elimination Theorem)**.** *Suppose any solution* (L^*, S^*) *to Outlier Pursuit* (2) *with input* $M = L^* + S^*$ *exactly recovers the column space of* L_0 *and the column support of* S_0*, i.e., Range*$(L^*) = \mathcal{U}_0$ *and* $\{j : S^*_{:j} \notin$ *Range*$(L^*)\} = \mathcal{I}_0$*. Then any solution* (L'^*, S'^*) *to* (2) *with input* $M'{=}L^* + \mathcal{P}_{\mathcal{I}} S^*$ *succeeds as well, where* $\mathcal{I} \subseteq \mathcal{I}^* {=} \mathcal{I}_0$*.*

Theorem 3 shows that the success of the algorithm is monotone on the cardinality of set $\mathcal{I}_0$. Thus by standard arguments in (Candès et al. 2011), (Candès, Romberg, and Tao 2006), and (Candès and Tao 2010), any guarantee proved for the Bernoulli distribution equivalently holds for the uniform distribution. For completeness, we give the details in the appendix. In the following, we will assume $\mathcal{I}_0 \sim \text{Ber}(p)$.

There are two main steps in our following proofs: 1. find dual conditions under which Outlier Pursuit succeeds; 2. construct dual certificates which satisfy the dual conditions.

Dual Conditions

We first give dual conditions under which Outlier Pursuit succeeds.

Lemma 1 (Dual Conditions for Exact Column Space)**.** *Let* $(L^*, S^*) = (L_0 + H, S_0 - H)$ *be any solution to Outlier Pursuit* (2)*,* $\hat{L} = L_0 + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0} H$ *and* $\hat{S} = S_0 - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0} H$*, where Range*$(L_0) =$ *Range*$(P_{\mathcal{I}_0^\perp} L_0)$ *and* $[S_0]_{:j} \notin$ *Range*(L_0) *for* $\forall j \in \mathcal{I}_0$*. Assume that* $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| < 1$*,* $\lambda > 4\sqrt{\mu r/n}$*, and* $\hat{L}$ *obeys incoherence* (3a)*. Then* L^* *has the same column space as that of* L_0 *and* S^* *has the same column indices as those of* S_0 *(thus* $\mathcal{I}_0 = \{j : S^*_{:j} \notin$ *Range*$(L^*)\}$*), provided that there exists a pair* (W, F) *obeying*

$$W = \lambda(\mathcal{B}(\hat{S}) + F), \tag{6}$$

with $\mathcal{P}_{\hat{\mathcal{V}}} W = 0$*,* $||W|| \le 1/2$*,* $\mathcal{P}_{\hat{\mathcal{I}}} F{=}0$ *and* $||F||_{2,\infty} {\le} 1/2$*.*

Remark 1. *There are two important modifications in our conditions compared with those of (Xu, Caramanis, and Sanghavi 2012): 1. The space* $\hat{\mathcal{T}}$ *(see Table 1) is not involved in our conclusion. Instead, we restrict* W *in the complementary space of* $\hat{\mathcal{V}}$*. The subsequent proofs benefit from such a modification. 2. Our conditions slightly simplify the constraint* $\hat{U}\hat{V}^T + W = \lambda(\mathcal{B}(\hat{S}) + F)$ *in (Xu, Caramanis, and Sanghavi 2012), where* $\hat{U}$ *is another dual certificate which needs to be constructed. Moreover, our modification enables us to build the dual certificate* W *by least squares and greatly facilitates our proofs.*

By Lemma 1, to prove the exact recovery of Outlier Pursuit, it is sufficient to find a suitable W such that

$$\begin{cases} W \in \hat{\mathcal{V}}^\perp, \\ ||W|| \le 1/2, \\ \mathcal{P}_{\hat{\mathcal{I}}} W = \lambda \mathcal{B}(\hat{S}), \\ ||\mathcal{P}_{\hat{\mathcal{I}}^\perp} W||_{2,\infty} \le \lambda/2. \end{cases} \tag{7}$$

As shown in the following proofs, our dual certificate W can be constructed by least squares.

Certification by Least Squares

The remainder of the proofs is to construct W which satisfies dual conditions (7). Note that $\hat{\mathcal{I}} = \mathcal{I}_0 \sim \text{Ber}(p)$. To construct W, we consider the method of least squares, which is

$$W = \lambda \mathcal{P}_{\hat{\mathcal{V}}^\perp} \sum_{k \ge 0} (\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}})^k \mathcal{B}(\hat{S}). \tag{8}$$

Note that we have assumed $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| < 1$. Thus $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}}|| = ||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}(\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}})|| = ||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}||^2 < 1$ and equation (8) is well defined. We want to highlight the advantage of our construction over that of (Candès et al. 2011). In our construction, we use a smaller space $\hat{\mathcal{V}} \subset \hat{\mathcal{T}}$ instead of $\hat{\mathcal{T}}$ in (Candès et al. 2011). Such a utilization significantly facilitates our proofs. To see this, notice that $\hat{\mathcal{I}} \cap \hat{\mathcal{T}} \neq \mathbf{0}$. Thus $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{T}}}|| = 1$ and the Neumann series $\sum_{k \ge 0}(\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{T}}}\mathcal{P}_{\hat{\mathcal{I}}})^k$ in the construction of (Candès et al. 2011) diverges. However, this issue does not exist for our construction. This benefits from our modification in Lemma 1. Moveover, our following theorem gives a good bound on $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}||$, whose proof takes into account that the elements in the same column of $\hat{S}$ are not independent. The complete proof can be found in the supplementary material.

Theorem 4. *For any* $\mathcal{I} \sim$ *Ber*(a)*, with an overwhelming probability*

$$||\mathcal{P}_{\hat{\mathcal{V}}} - a^{-1}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\mathcal{I}}\mathcal{P}_{\hat{\mathcal{V}}}|| < \varepsilon, \tag{9}$$

provided that $a \ge C_0 \varepsilon^{-2}(\mu r \log n)/n$ *for some numerical constant* $C_0 > 0$ *and other assumptions in Theorem 1 hold.*

By Theorem 4, our bounds in Theorem 1 guarantee that a is always larger than a constant when ρ_r is selected small enough.

We now bound $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}||$. Note $\hat{\mathcal{I}}^\perp \sim \text{Ber}(1-p)$. Then by Theorem 4, we have $||\mathcal{P}_{\hat{\mathcal{V}}} - (1-p)^{-1}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}^\perp}\mathcal{P}_{\hat{\mathcal{V}}}|| < \varepsilon$, or equivalently $(1-p)^{-1}||\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}} - p\mathcal{P}_{\hat{\mathcal{V}}}|| < \varepsilon$. Therefore, by the triangle inequality

$$\begin{aligned} ||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}||^2 &= ||\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| \\ &\le ||\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}} - p\mathcal{P}_{\hat{\mathcal{V}}}|| + ||p\mathcal{P}_{\hat{\mathcal{V}}}|| \\ &\le (1-p)\varepsilon + p. \end{aligned} \tag{10}$$

Thus we establish the following bound on $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}||$.

Corollary 1. *Assume that* $\hat{\mathcal{I}} \sim$ *Ber*(p)*. Then with an overwhelming probability* $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}||^2 \le (1-p)\varepsilon + p$*, provided that* $1 - p \ge C_0\varepsilon^{-2}(\mu r \log n)/n$ *for some numerical constant* $C_0 > 0$*.*

Note that $\mathcal{P}_{\hat{\mathcal{I}}}W = \lambda\mathcal{B}(\hat{\mathcal{S}})$ and $W \in \hat{\mathcal{V}}^{\perp}$. So to prove the dual conditions (7), it is sufficient to show that

$$\begin{aligned} &\text{(a)} \ \ ||W|| \leq 1/2, \\ &\text{(b)} \ \ ||\mathcal{P}_{\hat{\mathcal{I}}^{\perp}}W||_{2,\infty} \leq \lambda/2. \end{aligned} \tag{11}$$

Proofs of Dual Conditions

Since we have constructed the dual certificates W, the remainder is to prove that the construction satisfies our dual conditions (11), as shown in the following lemma.

Lemma 2. *Assume that $\hat{\mathcal{I}} \sim Ber(p)$. Then under the other assumptions of Theorem 1, W given by* (8) *obeys the dual conditions* (11).

The proof of Lemma 2 is in the supplementary material, which decomposes W in (8) as

$$W = \lambda\mathcal{P}_{\hat{\mathcal{V}}^{\perp}}\mathcal{B}(\hat{S}) + \lambda\mathcal{P}_{\hat{\mathcal{V}}^{\perp}}\sum_{k\geq 1}(\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}})^k(\mathcal{B}(\hat{S})), \tag{12}$$

and proves that the first and the second terms can be bounded with high probability, respectively.

Since Lemma 2 shows that W satisfies the dual conditions (11), the proofs of Theorem 1 are completed.

Experiments

This section aims at verifying the validity of our theories by numerical experiments. We solve Outlier Pursuit by the alternating direction method (Lin, Liu, and Su 2011), which is probably the most efficient algorithm for solving nuclear norm minimization problems. Details can be found in the supplementary material.

Validity of Regularization Parameter

We first verify the validity of our choice of regularization parameter $\lambda = 1/\sqrt{\log n}$. Our data are generated as follows. We construct $L_0 = XY^T$ as a product of $n \times r$ i.i.d. $\mathcal{N}(0,1)$ matrices. The nonzero columns of S_0 are uniformly selected, whose entries follow i.i.d. $\mathcal{N}(0,1)$. Finally, we construct our observation matrix $M = L_0 + S_0$. We solve the Outlier Pursuit (2) to obtain an optimal solution (L^*, S^*) and then compare with (L_0, S_0). The distance between the column spaces are measured by $||\mathcal{P}_{\mathcal{U}^*} - \mathcal{P}_{\mathcal{U}_0}||_F$ and the distance between the column supports is measured by the Hamming distance. We run the experiment by 10 times and report the average results. Table 2 shows that our choice of λ can make Outlier Pursuit exactly recover the column space of L_0 and the column support of S_0.

To verify that the success of Outlier Pursuit (2) is robust to various noise magnitudes, we test the cases where the entries of S_0 follow i.i.d. $\mathcal{N}(0,1/n)$, $\mathcal{N}(0,1)$, and $\mathcal{N}(0,100)$, respectively. We specifically adopt $n = 1,500$, while all the other settings are the same as the experiment above. Table 3 shows that Outlier Pursuit (2) could exactly recover the ground truth subspace and correctly identify the noise index, no matter what magnitude the noises are.

We also compare our parameter with those of other orders, e.g., $\lambda = c_1/\sqrt{n}$ and $\lambda = c_2/\log n$. The coefficients c_1 and c_2 are calibrated on a 200×200 matrix, i.e., $c_1/\sqrt{200} = 1/\sqrt{\log 200}$ and $c_2/\log 200 = 1/\sqrt{\log 200}$, and we fix the obtained c_1 and c_2 for other sizes of problems. For each size of problem, we test with different rank and outlier ratios. For each choice of rank and outlier ratios, we record the distances between the column spaces and between the column supports as did above. The experiment is run by 10 times, and we define the algorithm succeeds if the distance between the column spaces is below 10^{-6} and the Humming distance between the column supports is exact 0 for 10 times. As shown in Figure 1, $\lambda = 1/\sqrt{\log n}$ consistently outperforms $\Theta(1/\sqrt{n})$ and $\Theta(1/\log n)$. The advantage of our parameter is salient. Moreover, a phase transition phenomenon, i.e., when the rank and outlier ratios are below a curve Outlier Pursuit strictly succeeds and when they are above the curve Outlier Pursuit fails, can also be observed.

Table 2: Exact recovery on problems with different sizes. Here $\text{rank}(L_0) = 0.05n$, $||S_0||_{2,0} = 0.1n$, and $\lambda = 1/\sqrt{\log n}$.

n	dist(Range(L^*),Range(L_0))	dist($\mathcal{I}^*, \mathcal{I}_0$)
500	2.46×10^{-7}	0
1,000	8.92×10^{-8}	0
2,000	2.30×10^{-7}	0
3,000	3.15×10^{-7}	0

Table 3: Exact recovery on problems with different noise magnitudes. Here $n = 1,500$, $\text{rank}(L_0) = 0.05n$, $||S_0||_{2,0} = 0.1n$, and $\lambda = 1/\sqrt{\log n}$.

Magnitude	dist(Range(L^*),Range(L_0))	dist($\mathcal{I}^*, \mathcal{I}_0$)
$\mathcal{N}(0, 1/n)$	4.36×10^{-7}	0
$\mathcal{N}(0, 1)$	6.14×10^{-7}	0
$\mathcal{N}(0, 100)$	2.56×10^{-7}	0

Tightness of Bounds

We then test the tightness of our bounds. We repeat the exact recovery experiments by increasing the data size successively. Each experiment is run by 10 times, and Figure 2 plots the fraction of correct recoveries: white denotes perfect recovery in all experiments, and black represents failures in all experiments. It shows that the intersection point of the phase transition curve with the vertical axes is almost unchanged and that with the horizontal axes moves leftwards very slowly. These are consistent with our forecasted orders $O(n)$ and $O(n/\log n)$, respectively. So our bounds are tight.

Experiment on Real Data

To test the performance of Outlier Pursuit (2) on the real data, we conduct an experiment on the Hopkins 155 database[1]. The Hopkins 155 data set is composed of multiple data points drawn from two or three motion objects. The data points (trajectory) of each object lie in a single subspace, so

[1] http://www.vision.jhu.edu/data/hopkins155

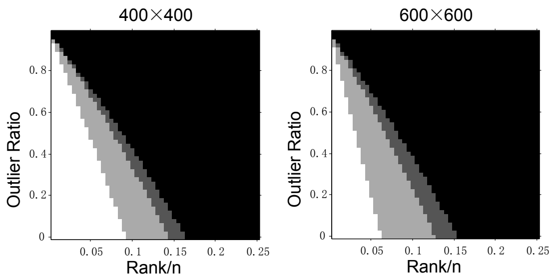

Figure 1: Exact recovery under varying orders of the regularization parameter. **White Region:** Outlier Pursuit succeeds under $\lambda = c_1/\sqrt{n}$. **White and Light Gray Regions:** Outlier Pursuit succeeds under $\lambda = c_2/\log n$. **White, Light Gray, and Dark Gray Regions:** Outlier Pursuit succeeds under $\lambda = 1/\sqrt{\log n}$. **Black Regions:** Outlier Pursuit fails. The success region of $\lambda = 1/\sqrt{\log n}$ strictly contains those of other orders.

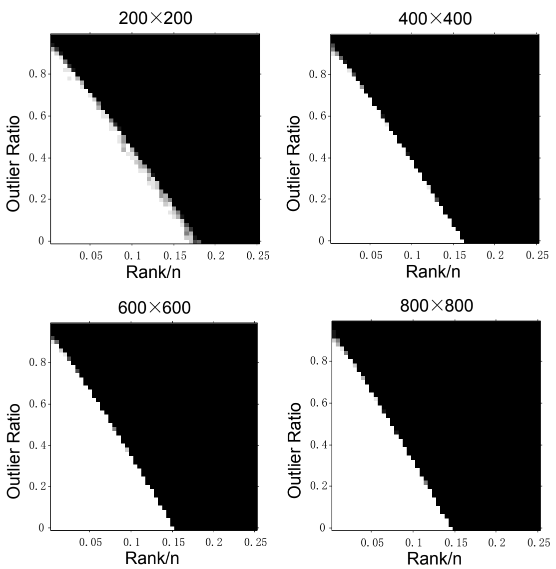

Figure 2: Exact recovery of Outlier Pursuit on random problems of varying sizes. The success regions (white regions) change very little when the data size changes.

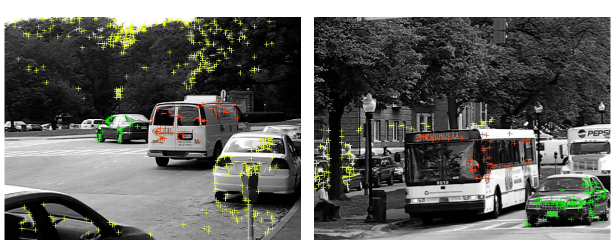

Figure 3: Examples in Hopkins 155 **(best viewed in color)**.

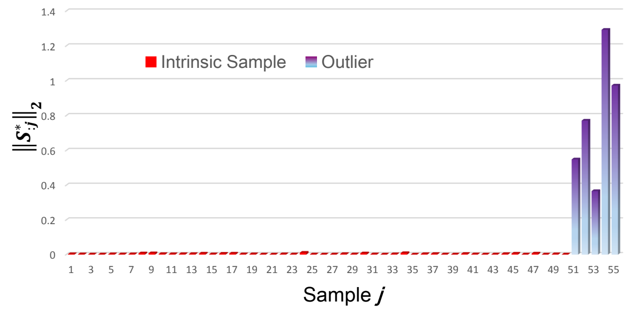

Figure 4: Results on the Hopkins 155 database. The horizontal axis represents serial number of the sample while the vertical axis represents $||S^*_{:j}||_2$, where S^* is the optimal solution obtained by the algorithm. The last five samples are the real outliers.

it is possible to separate different objects according to their underlying subspaces. Figure 3 presents some examples in the database. In particular, we select one object in the data set, which contains 50 points with 49-dimension feature, as the intrinsic sample. We then adopt another 5 data points from other objects as the outliers, so the outlier ratio is nearly 10%. Figure 4 shows $||S^*_{:j}||_2$ for each sample j, where S^* is the optimal solution obtained by the algorithm. Note that some of $||S^*_{:j}||_2$s for the intrinsic sample $j \in \{1, 2, ..., 50\}$ are not strictly zero. This is because the data is sampled from real sensors and so there are small noises therein. However, from the scale of $||S^*_{:j}||_2$, one can easily distinguish the outliers from the ground truth data. So it is clear that Outlier Pursuit (2) correctly identifies the outlier index, which demonstrates the effectiveness of the model on the real data.

Conclusion

We have investigated the exact recovery problem of R-PCA via Outlier Pursuit. In particular, we push the upper bounds on the allowed outlier number from $O(n/r)$ to $O(n)$, where r is the rank of the intrinsic matrix and may be comparable to n. We further suggest a global choice of the regularization parameter, which is $1/\sqrt{\log n}$. This result waives the necessity of tuning the regularization parameter in the previous literature. Thus our analysis significantly extends the working range of Outlier Pursuit. Extensive experiments testify to the validity of our choice of regularization parameter and the tightness of our bounds.

Acknowledgements

Zhouchen Lin is supported by NSF China (grant nos. 61272341 and 61231002), 973 Program of China (grant no. 2015CB3525), and MSRA Collaborative Research Program. Chao Zhang and Hongyang Zhang are partially supported by a NKBRPC (2011CB302400) and NSFC (61131003).

References

Candès, E. J., and Recht, B. 2009. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* 9(6):717–772.

Candès, E. J., and Tao, T. 2010. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory* 56(5):2053–2080.

Candès, E. J.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust principal component analysis? *Journal of the ACM* 58(3):11.

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

Chandrasekaran, V.; Sanghavi, S.; Parrilo, P. A.; and Willsky, A. S. 2011. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization* 21(2):572–596.

Chen, Y.; Xu, H.; Caramanis, C.; and Sanghavi, S. 2011. Robust matrix completion and corrupted columns. In *International Conference on Machine Learning*, 873–880.

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

De La Torre, F., and Black, M. J. 2003. A framework for robust subspace learning. *International Journal of Computer Vision* 54(1):117–142.

Dong, J.; Cheng, B.; Chen, X.; Chua, T.; Yan, S.; and Zhou, X. 2013. Robust image annotation via simultaneous feature and sample outlier pursuit. *ACM Transactions on Multimedia Computing, Communications, and Applications* 9(4):24.

Eldar, Y., and Kutyniok, G. 2012. *Compressed sensing: theory and applications*. Cambridge University Press.

Gear, W. 1998. Multibody grouping from motion images. *International Journal of Computer Vision* 29(2):133–150.

Gross, D. 2011. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory* 57(3):1548–1566.

Huber, P. J. 2011. *Robust statistics*. Springer.

Ji, H.; Liu, C.; Shen, Z.; and Xu, Y. 2010. Robust video denoising using low-rank matrix completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1791–1798.

Lin, Z.; Liu, R.; and Su, Z. 2011. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in Neural Information Processing Systems*, 612–620.

Liu, G.; Lin, Z.; Yan, S.; Sun, J.; and Ma, Y. 2013. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1):171–184.

Ma, Y.; Derksen, H.; Hong, W.; and Wright, J. 2007. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(9):1546–1562.

McCoy, M., and Tropp, J. A. 2011. Two proposals for robust PCA using semidefinite programming. *Electronic Journal of Statistics* 5:1123–1160.

Pan, Y.; Lai, H.; Liu, C.; Tang, Y.; and Yan, S. 2013. Rank aggregation via low-rank and structured-sparse decomposition. In *AAAI Conference on Artificial Intelligence*.

Peng, Y.; Ganesh, A.; Wright, J.; Xu, W.; and Ma, Y. 2010. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 763–770.

Rao, S.; Tron, R.; Vidal, R.; and Ma, Y. 2010. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(10):1832–1845.

Wright, J.; Ganesh, A.; Rao, S.; Peng, Y.; and Ma, Y. 2009. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in Neural Information Processing Systems*, 2080–2088.

Wu, L.; Ganesh, A.; Shi, B.; Matsushita, Y.; Wang, Y.; and Ma, Y. 2011. Robust photometric stereo via low-rank matrix completion and recovery. In *Asian Conference on Computer Vision*, 703–717.

Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI Conference on Artificial Intelligence*.

Xu, H.; Caramanis, C.; and Sanghavi, S. 2012. Robust PCA via outlier pursuit. *IEEE Transaction on Information Theory* 58(5):3047–3064.

Yan, J., and Pollefeys, M. 2006. A general framework for motion segmentation: Independent, articulated, rigid, nonrigid, degenerate and nondegenerate. In *European Conference on Computer Vision*, volume 3954, 94–106.

Zhang, Z.; Ganesh, A.; Liang, X.; and Ma, Y. 2012. TILT: Transform-invariant low-rank textures. *International Journal of Computer Vision* 99(1):1–24.

Zhang, H.; Lin, Z.; Zhang, C.; and Gao, J. 2014. Robust latent low rank representation for subspace clustering. *Neurocomputing* 145:369–373.

Zhang, H.; Lin, Z.; and Zhang, C. 2013. A counterexample for the validity of using nuclear norm as a convex surrogate of rank. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, volume 8189, 226–241.