Supplementary Material of

Exact Recoverability of Robust PCA via Outlier Pursuit
with Relatively Dense Outliers

Paper ID 1532

Robust PCA via Outlier Pursuit:

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

\(\mu\)-incoherence condition on matrix \(L = U\Sigma V^T\):

\[
\max_i \|V^T e_i\|_2 \leq \frac{\mu \sqrt{p}}{n}, \quad \text{(avoid column sparsity)} \tag{1a}
\]

\[
\max_i \|U^T e_i\|_2 \leq \frac{\mu \sqrt{m}}{n}, \quad \text{(avoid row sparsity)} \tag{1b}
\]

\[
\|UV^T\|_\infty \leq \sqrt{\frac{\mu}{mn}}. \tag{1c}
\]

Ambiguity condition on matrix \(S\):

\[
\|B(S)\| \leq \log n/4. \tag{2}
\]

Main Results:

**Theorem 1 (Exact Recovery of Outlier Pursuit).** Suppose \(m = \Theta(n)\), \(\text{Range}(L_0) = \text{Range}(P_{I_0} L_0)\), and \([S_0]_{j,j} \notin \text{Range}(L_0)\) for \(\forall j \in I_0\). Then any solution \((L_0 + H, S_0 - H)\) to Outlier Pursuit (1) 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 \(1 - e^{-c n^{-10}}\), if the column support of \(S_0\) is uniformly distributed among all sets of cardinality \(s\) and

\[
\text{rank}(L_0) \leq \rho_c \frac{n(2)}{\mu \log n} \quad \text{and} \quad s \leq \rho_s n, \tag{3}
\]

where \(c, \rho_c\), and \(\rho_s\) are constants, \(L_0 + P_{I_0} P_{I_0}^T H\) satisfies \(\mu\)-incoherence condition (1a), and \(S_0 - P_{I_0} P_{I_0}^T H\) satisfies ambiguity condition (2).

**Architecture of Proofs**

This section is devoted to proving Theorem 1. Without loss of generality, we assume \(m = n\). The following theorem presents a good characteristic of Outlier Pursuit.

**Theorem 2 (Elimination Theorem).** Suppose any solution \((L^*, S^*)\) to Outlier Pursuit (1) with input \(M = L^* + S^*\) exactly recovers the column space of \(L_0\) and the column support of \(S_0\), i.e., \(\text{Range}(L^*) = \text{Range}(L_0)\) and \([j : S_{j,j}^* \notin \text{Range}(L^*)] = I_0\). Then any solution \((L^*, S^*)\) to (1) with input \(M' = L^* + P_{I^*} S^*\) succeeds as well, where \(I \subseteq I^* = I_0\).

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<table>
<thead>
<tr>
<th>Notations</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m, n)</td>
<td>Size of the data matrix (M).</td>
</tr>
<tr>
<td>(n(1), n(2))</td>
<td>(n(1) = \max{m, n}, n(2) = \min{m, n}).</td>
</tr>
<tr>
<td>(\Theta(n))</td>
<td>(\Theta(n)) grows in the same order of (n).</td>
</tr>
<tr>
<td>(O(n))</td>
<td>(O(n)) grows equal to or less than the order of (n).</td>
</tr>
<tr>
<td>(e_i)</td>
<td>Vector whose (i)th entry is 1 and others are 0s.</td>
</tr>
<tr>
<td>(M_{ij})</td>
<td>The (j)th column of matrix (M).</td>
</tr>
<tr>
<td>(S_0)</td>
<td>The entry at the (i)th row and (j)th column of (M).</td>
</tr>
<tr>
<td>(|v|_1)</td>
<td>(|v|_1) norm for vector, (|v|_1 = \sqrt{\sum_i v_i^2}).</td>
</tr>
<tr>
<td>(|v|_\infty)</td>
<td>(|v|_\infty) norm, the sum of singular values.</td>
</tr>
<tr>
<td>(|v|_0)</td>
<td>(|v|_0) norm, number of nonzero entries.</td>
</tr>
<tr>
<td>(|v|_{2,0})</td>
<td>(|v|_{2,0}) norm, number of nonzero columns.</td>
</tr>
<tr>
<td>(\ell_1) norm</td>
<td>(\ell_1) norm, (|M|<em>1 = \sum</em>{i,j}</td>
</tr>
<tr>
<td>(\ell_{2,1}) norm</td>
<td>(\ell_{2,1}) norm, (|M|_{2,1} = \sqrt{\sum_j</td>
</tr>
<tr>
<td>(|v|_{2,\infty})</td>
<td>(|v|<em>{2,\infty}) norm, (|v|</em>{2,\infty} = \max_j</td>
</tr>
<tr>
<td>(|v|_F)</td>
<td>Frobenius norm, (|M|<em>F = \sqrt{\sum</em>{i,j} M_{ij}^2}).</td>
</tr>
<tr>
<td>(|v|_\infty)</td>
<td>Infinity norm, (|v|<em>\infty = \max</em>{i,j}</td>
</tr>
<tr>
<td>(|P|_2)</td>
<td>(Matrix) operator norm.</td>
</tr>
<tr>
<td>(L^<em>, S^</em>)</td>
<td>Optimal solutions to Outlier Pursuit.</td>
</tr>
<tr>
<td>(L_0, S_0)</td>
<td>Ground Truth.</td>
</tr>
<tr>
<td>(\hat{U}, \hat{V})</td>
<td>Left and right singular vectors of (\hat{L}).</td>
</tr>
<tr>
<td>(U_0, U, U^*)</td>
<td>Column space of (L_0, \hat{L}, L^*).</td>
</tr>
<tr>
<td>(V_0, \hat{V}, V^*)</td>
<td>Row space of (L_0, \hat{L}, L^*).</td>
</tr>
<tr>
<td>(T)</td>
<td>(T = {UX^T + Y\hat{V}^T, \forall X, Y \in \mathbb{R}^{n \times r}}).</td>
</tr>
<tr>
<td>(\hat{X})</td>
<td>Orthogonal complement of the space (\hat{X}).</td>
</tr>
<tr>
<td>(P_{U_0}', P_V)</td>
<td>(P_{U_0'} M = U\hat{U}^T M, P_V M = M\hat{V}).</td>
</tr>
<tr>
<td>(P_{\hat{U}}', P_{\hat{V}})</td>
<td>(P_{\hat{U}}' \hat{M} = \hat{P}_{\hat{U}} \hat{M}).</td>
</tr>
<tr>
<td>(I_0, I, I^*)</td>
<td>Index of outliers of (S_0, \hat{S}, S^*).</td>
</tr>
<tr>
<td>({I_0})</td>
<td>Outliers number of (S_0).</td>
</tr>
<tr>
<td>(X \in I)</td>
<td>The column support of (X) is a subset of (I).</td>
</tr>
<tr>
<td>(B(\hat{S}))</td>
<td>(B(\hat{S}) = {H : P_{\hat{X}}(H) = 0; H_j = \hat{S}<em>{j,.}/|\hat{S}</em>{j,.}|_2, j \in I_0}).</td>
</tr>
<tr>
<td>(\sim \text{Ber}(p))</td>
<td>Obey Bernoulli distribution with parameter (p).</td>
</tr>
<tr>
<td>(|N(\alpha, \beta^2)|_2)</td>
<td>Gaussian distribution (mean (\alpha) and variance (\beta^2)).</td>
</tr>
</tbody>
</table>
Proof. Let $(L^*, S^*)$ be the solution of (1) with input matrix $M'$ and $(L^*, S^*)$ be the solution of (1) with input matrix $M$. Then we have
\[ ||L^*|| + \lambda ||S^*|| \leq ||L|| + \lambda ||P^* S^*||. \]
Therefore
\[ ||L^*|| + \lambda ||S^* + P^* S^*|| \leq ||L|| + \lambda ||P^* S^*||. \]
Thus $(L^*, S^* + P^* S^*)$ is optimal to problem with input $M'$ and by assumption we have
\[ \text{Range}(L^*) = \text{Range}(L), \{ j : |S^*| \notin \text{Range}(L) \} = \text{Supp}(S_0). \]
The second equation implies $I \subseteq \{ j : |S^*| \notin \text{Range}(L) \}$. Suppose $I \neq \{ j : |S^*| \notin \text{Range}(L) \}$. Then there exists an index $k$ such that $S^* \notin \text{Range}(L)$ and $k \notin I$, i.e., $M_k = L_k \in \text{Range}(L)$. Note that $L_k \notin \text{Range}(L)$. Thus $S_k \notin \text{Range}(L)$ and we have a contradiction. Then $I = \{ j : |S^*| \notin \text{Range}(L) \}$ and the algorithm succeeds.

Theorem 2 shows that the success of the algorithm is monotone on $|I_0|$. 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 equivalent holds for the uniform distribution. For completeness, we give the details in the appendix. In the following, we will assume $I_0 \sim \text{Ber}(p)$.

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 (1). $L = L_0 + P_{L_0} P_{L_0} H$ and $S = S_0 - P_{L_0} P_{L_0} H$, where $\text{Range}(L_0) = \text{Range}(P_{L_0} L_0)$ and $|S_0| \notin \text{Range}(L_0)$ for all $j \in I_0$. Assume that $||P^* P|| < 1$, $\lambda > \sqrt{n} / \sqrt{m}$, and $L^*$ obeys incoherence (1a). 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 $I_0 = \{ j : S^*_j \notin \text{Range}(L^*) \}$), provided that there exists a pair $(W, F)$ obeying
\[ W = \lambda(B(S) + F), \]
with $P_Y W = 0$, $||W|| \leq 1 / 2$, $P_Z F = 0$ and $||F||_\infty \leq 1 / 2$.

Proof. We first recall that the subgradients of nuclear norm and $\ell_{2,1}$ norm are as follows:
\[ \partial ||L|| = \{ U \tilde{V}^T + Q : \tilde{Q} \in \tilde{T}, ||\tilde{Q}|| \leq 1 \}, \]
\[ \partial ||S|| = \{ B(S) + \tilde{E} : \tilde{E} \in \tilde{T}, ||\tilde{E}||_\infty \leq 1 \}. \]
Let $H_1 = P_{I_0} P_{I_0} H$ and $H_2 = P_{I_0} P_{I_0} H + P_{I_0} P_{I_0} H + P_{I_0} P_{I_0} H$, and note that $U = U_0$ and $\tilde{I} = I_0$. By the definition of the subgradient, the inequality follows
\[ ||L_0 + H|| + \lambda ||S_0 - H|| \leq ||L|| + \lambda ||S + P^* S|| \]
\[ \geq ||L|| + \lambda ||S + P^* S|| + (\hat{U}^T \hat{Q}) \hat{S} - \lambda (B(S) + \hat{E}) \hat{S} \]
\[ = ||L|| + \lambda ||S + P_{I_0} H|| + \langle \hat{Q}, P_{I_0} H \rangle - \lambda (B(S), P_{I_0} H) - \lambda (B(S), P_{I_0} H)^{-1} - \lambda (\hat{E}, P_{I_0} H). \]
Now adopt $\hat{Q}$ such that $\langle \hat{Q}, P_{I_0} H \rangle = ||P_{Y} P_{I_0} H||$ and $\langle \hat{E}, P_{I_0} H \rangle = -||P_{Y} P_{I_0} H||$. We have
\[ ||L_0 + H|| + \lambda ||S_0 - H|| \]
\[ \geq ||L|| + \lambda ||S|| + \frac{\lambda}{4} \frac{\mu_r}{n} ||P_{I_0} H|| + \frac{3 \lambda}{4} ||P_{X_0} H||, \]
\[ \geq ||L|| + \lambda ||S|| + \frac{\lambda}{4} \frac{\mu_r}{n} ||P_{I_0} H|| + \frac{3 \lambda}{4} ||P_{X_0} H||, \]
Notice that
\[ ||\langle -\lambda B(S), P_{I_0} H \rangle || = ||\langle \lambda F - W, P_{I_0} H \rangle || \]
\[ \leq ||\langle W, P_{I_0} H \rangle || + ||\langle F, P_{I_0} H \rangle || \]
\[ \leq \frac{1}{2} ||P_{Y} P_{I_0} H|| + \lambda ||P_{Y} P_{I_0} H||, \]
\[ \frac{1}{2} ||P_{Y} P_{I_0} H|| + \frac{\lambda}{4} ||P_{X_0} H|| + \frac{3 \lambda}{4} ||P_{X_0} H||. \]

\[ \frac{1}{2} ||P_{Y} P_{I_0} H|| + \frac{\lambda}{4} ||P_{X_0} H|| + \frac{3 \lambda}{4} ||P_{X_0} H||. \]

\[ \text{By the duality between the nuclear norm and the operator norm, there exists a Q such that } (Q, P_{Y} P_{I_0} H) = ||P_{Y} P_{I_0} H||, \]
and $||Q|| \leq 1$. Thus we take $\hat{Q} = P_{I_0} H$ and $Q \in \tilde{T}$. It holds similarly for $\hat{E}$. 

\[ \frac{1}{2} ||P_{Y} P_{I_0} H|| + \frac{\lambda}{4} ||P_{X_0} H|| + \frac{3 \lambda}{4} ||P_{X_0} H||. \]
Since \((L^*, S^*) = (L_0 + H, S_0 - H)\) is optimal, above inequality shows \(||P_{\hat{Y}^\perp} P_{\hat{U}_0^\perp} H||_\ast = ||P_{\hat{Y}^\perp} P_{\hat{U}_1^\perp} H||_2.1 = 0\), i.e., \(P_{\hat{U}_0^\perp} H \in \hat{I} \cap \hat{V}\). Also notice that \(||P_{\hat{Y}^\perp} P_{\hat{Y}}|| < 1\) implies \(\hat{I} \cap \hat{V} = \{0\}\). We conclude \(P_{\hat{U}_0^\perp} H = 0\). Furthermore, \(||P_{\hat{I}_0^\perp} H||_2.1 = 0\) implies \(H \in \mathcal{I}_0\). Thus \(H \in \mathcal{U}_0 \cap \mathcal{I}_0\), i.e., \(U^* \subseteq \mathcal{U}_0\) and \(I^* \subseteq \mathcal{I}_0\).

We now prove \(U^* = \mathcal{U}_0\). According to the assumption \(\text{Range}(L_0) = \text{Range}(P_{\hat{U}_0^\perp} L_0)\) and \(I \in \mathcal{U}_0 \cap \mathcal{I}_0\), Range \((I^*) = \text{Range}(L_0 + H) = \text{Range}(L_0)\), i.e., \(U^* = \mathcal{U}_0\). We then prove \(I^* = \mathcal{I}_0\). Assume that \(I^* \neq \mathcal{I}_0\), i.e., there exists a \(j \in \mathcal{I}_0\) such that \(S^*_{j} = 0\). Note that \([S_0]_{j} \notin \text{Range}(L_0)\). Thus \(M_{j} = [L_0]_{j} + [S_0]_{j} = L^*_{j} \notin \mathcal{U}_0\), which contradicts \(U^* \subseteq \mathcal{U}_0\). So \(I^* = \mathcal{I}_0\). 

**Remark 1.** There are two important modifications in our conditions compared with those of (Xu, Caramanis, and Sanghavi 2012): 1. The space \(\hat{I}\) (see Table 1) is not involved in our conclusion. Instead, we restrict \(W\) in the complementary space of \(\hat{Y}\). The subsequent proofs benefit from such a modification. 2. Our conditions slightly simplify the constraint \(\hat{U} \hat{V}^T + W = \lambda B(\hat{S}) + F\) in (Xu, Caramanis, and Sanghavi 2012), where \(\hat{U}\) is another dual certificate which need 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

\[
\begin{aligned}
W &\in \hat{V}^\perp, \\
\|W\| &\leq 1/2, \\
P_{\hat{Z}} W &\in \lambda B(\hat{S}), \\
\|P_{\hat{Z}^\perp} W\|_{2.\infty} &\leq \lambda/2.
\end{aligned}
\]

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 (5). Note that \(\hat{I} = \mathcal{I}_0 \sim \text{Ber}(p)\). To construct \(W\), we consider the method of least squares, which is

\[
W = \lambda P_{\hat{V}^\perp} \sum_{k \geq 0} (P_{\hat{Z}} P_{\hat{Y}} P_{\hat{Z}})^k B(\hat{S}).
\]

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

### Proofs of Dual Conditions

Since we have constructed the dual certificates \(W\), the remainder is to prove that such a construction satisfies our dual conditions (9), as shown in the following lemma.

**Lemma 2.** Assume that \(\hat{I} \sim \text{Ber}(p)\). Then under the other assumptions of Theorem 1, W given by (6) obeys the dual conditions (9).

**Proof.** Let \(R = \sum_{k \geq 1} (P_{\hat{Z}} P_{\hat{Y}} P_{\hat{Z}})^k\). Then

\[
W = \lambda P_{\hat{V}^\perp} \sum_{k \geq 0} (P_{\hat{Z}} P_{\hat{Y}} P_{\hat{Z}})^k B(\hat{S}) = \lambda P_{\hat{V}^\perp} B(\hat{S}) + \lambda P_{\hat{V}^\perp} R(B(\hat{S})).
\]

Now we check the two conditions in (9).

(a) By the assumption, we have \(\|B(\hat{S})\| \leq \sqrt{\log n}/4\). Thus the first term in (10) obeys

\[
\lambda \|P_{\hat{V}^\perp} B(\hat{S})\| \leq \lambda \|B(\hat{S})\| \leq \frac{1}{4}.
\]

Now we focus on the second term in (10). Let \(N\) represent the 1/2-net of the unit ball \(S^*\), whose cardinality \(|N|\) is at most \(6^n\) (see Eldar and Kutyniok 2012). Then a standard argument in (Eldar and Kutyniok 2012) showed that

\[
\|R(B(\hat{S}))\| \leq 4 \sup_{x, y \in N} \langle y, R(B(\hat{S}))x \rangle.
\]
Note that the operator $\mathcal{R}$ is self-adjoint. Now let

$$X(x, y) = \langle y, \mathcal{R}(B(\hat{S}))x \rangle = \langle \mathcal{R}(yx^T), B(\hat{S}) \rangle$$

$$= \sum_j \langle [\mathcal{R}(yx^T)]_{:, j}, B(\hat{S})_{:, j} \rangle = \sum_j \langle \mathcal{R}(yx^T)_{:, j}, \delta_j B_{:, j} \rangle = \sum_j \delta_j B^T_{:, j} [\mathcal{R}(yx^T)]_{:, j},$$

where $B$ is a matrix such that $\|B_{:, j}\|_2 = 1$ and $B(\hat{S})_{:, j} = \delta_j B_{:, j}$, and $\delta_j$ is a random variable such that

$$\delta_j = \begin{cases} 1, & w.p. p, \\ 0, & w.p. 1 - p. \end{cases}$$

(13)

Notice that

$$|\delta_j B^T_{:, j} [\mathcal{R}(yx^T)]_{:, j}|^2 \leq \|B_{:, j}\|^2 \|\mathcal{R}(yx^T)_{:, j}\|^2_2 = \||\mathcal{R}(yx^T)\||^2_{\mathcal{F}},$$

and

$$\sum_{j=1}^n \||\mathcal{R}(yx^T)\||^2_{\mathcal{F}} = \||\mathcal{R}(yx^T)\||^2_{\mathcal{F}} \leq \||\mathcal{R}\||^2 \|yx^T\|^2_2 = \||\mathcal{R}\||^2 \|y\|^2_2 \|x\|^2_2 = \||\mathcal{R}\||^2.$$

Also note that $\mathbb{E} X(x, y) = 0$. Thus, by the Hoeffding inequality, we have

$$\mathbb{P}(\|X(x, y)\| > t\hat{Z}) \leq 2\exp \left(-\frac{t^2}{2\||\mathcal{R}\||^2}\right).$$

As a result,

$$\mathbb{P}(\sup_{x, y \in \mathcal{N}} |X(x, y)| > t\hat{Z}) \leq 2|\mathcal{N}|^2 \exp \left(-\frac{t^2}{2\||\mathcal{R}\||^2}\right).$$

Namely, by inequality (12),

$$\mathbb{P}(\|\mathcal{R}(B(\hat{S}))\| > t\hat{Z}) \leq 2|\mathcal{N}|^2 \exp \left(-\frac{t^2}{32\||\mathcal{R}\||^2}\right).$$

Now suppose $\|P_\perp \mathcal{P}_\mathcal{V}\| \leq \sigma$. Then

$$\||\mathcal{R}\|| \leq \sum_{k=1}^{\gamma} \sigma^{2k} = \sigma^2 \frac{\gamma^2}{1 - \sigma^2} = \frac{1}{\gamma},$$

where $\gamma$ can be sufficient large, and we have

$$\mathbb{P}(\|\mathcal{R}(B(\hat{S}))\| > t) \leq \mathbb{P}(\|\mathcal{R}(B(\hat{S}))\| > t \mid \|P_\perp \mathcal{P}_\mathcal{V}\| \leq \sigma) + \mathbb{P}(\|P_\perp \mathcal{P}_\mathcal{V}\| > \sigma) \leq 2|\mathcal{N}|^2 \exp \left(-\frac{t^2}{32\||\mathcal{R}\||^2}\right) + \mathbb{P}(\|P_\perp \mathcal{P}_\mathcal{V}\| > \sigma) \leq 2|\mathcal{N}|^2 \exp \left(-\frac{\gamma^2 t^2}{32}\right) + \mathbb{P}(\|P_\perp \mathcal{P}_\mathcal{V}\| > \sigma),$$

where $\mathbb{P}(\|P_\perp \mathcal{P}_\mathcal{V}\| > \sigma)$ is tiny. Adopt $t = 1/(4\lambda) = \sqrt{\log n}/4$. Then $\lambda\|\mathcal{R}(B(\hat{S}))\| \leq 1/4$ holds with an overwhelming probability. This together with (10) and (11) proves $\|W\| \leq 1/2$.

(b) Let $\mathcal{G}$ stand for $\mathcal{G} = \sum_{k \geq 0} (P_\perp \mathcal{P}_\mathcal{V} P_\perp)^k$. Then $W = \lambda \mathcal{P}_\perp \mathcal{G}(B(\hat{S}))$. Notice that $\mathcal{G}(B(\hat{S})) \in \mathcal{I}$. Thus

$$P_\perp W = \lambda P_\perp \mathcal{P}_\perp \mathcal{G}(B(\hat{S})) = \lambda P_\perp \mathcal{P}_\perp \mathcal{G}(B(\hat{S})) - \lambda P_\perp \mathcal{P}_\perp \mathcal{G}(B(\hat{S})) = -\lambda P_\perp \mathcal{P}_\perp \mathcal{G}(B(\hat{S})).$$

Now denote $Q \triangleq \mathcal{P}_\perp \mathcal{G}(B(\hat{S}))$. Note that $\mathcal{G}$ is an operator functioning at the right hand side of a matrix and

$$\|Q_{ij}\|^2 = \sum_i Q_{ij}^2 = \sum_i \langle \mathcal{P}_\perp \mathcal{G}(B(\hat{S})), e_i e_j^T \rangle^2$$

$$= \sum_i \sum_{j_0} \langle B(\hat{S})_{:, j_0}, \mathcal{P}_\perp \mathcal{G}(e_i e_j^T) e_{j_0} \rangle^2$$

$$= \sum_{j_0} \sum_i \langle e_i^T B(\hat{S})_{:, j_0}, \mathcal{P}_\perp \mathcal{G}(e_j^T) e_{j_0} \rangle^2$$

$$= \sum_{j_0} \sum_i \langle \mathcal{P}_\perp \mathcal{G}(e_j^T) \rangle_{ij}^2$$

$$\leq \sum_{j_0} \|\mathcal{P}_\perp \mathcal{G}(e_j^T)\|_{\perp}^2$$

$$= \|\mathcal{P}_\perp \mathcal{G}(e_j^T)\|_{\mathcal{F}}$$

$$\leq \left(\frac{\sigma^2}{1 - \sigma^2}\right)^2 \leq \frac{1}{4}, \ \forall j.$$

Thus $\|P_\perp W\|_{2, \infty} = \lambda \|P_\perp Q\|_{2, \infty} \leq \lambda\|Q\|_{2, \infty} \leq \lambda/2$. \hfill $\Box$

Now we have proved that $W$ satisfies the dual conditions (9). So our proofs finish.

### Tightness of Bounds

The following theorem shows a good property of our bounds in inequalities (3).

**Theorem 4.** The orders of the upper bounds given by inequalities (3) are tight.

**Proof.** Since $O(n)$ is the highest order for the possible number of corruptions, the order of our bound for the corruption cardinality $s$ is tight.

We then demonstrate that our bound for rank($L_0$) is tight. McCoy and Tropp (2011) showed that the optimal solution $L^*$ to model (1) satisfies

$$\text{rank}(L^*) \leq n/\log n. \ \ (14)$$

If the order of rank($L_0$) is strictly higher than $\Theta(n/\log n)$, then according to (14) it is impossible for $L^*$ to exactly recover the column space of $L_0$ due to their different ranks. So rank($L_0$) should be no larger than $\Theta(n/\log n)$ and the order of our bound is tight. \hfill $\Box$
**Algorithm**

In this section, we give the algorithm for Robust PCA (R-PCA) via Outlier Pursuit. To solve the model, we apply the alternating direction method (ADM) (Lin, Chen, and Ma 2009), which is probably the most widely used method for solving nuclear norm minimization problems.

Given the Outlier Pursuit model

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

whose augmented Lagrangian formulation corresponds to

$$
\mathcal{L}(L, S, Y, \mu) = \|L\|_* + \lambda \|S\|_{2,1} + \langle M - L - S, Y \rangle + \frac{\mu}{2} \|M - L - S\|^2_F,
$$

(15)

ADM solves model (15) by updating one argument in (16) and fixing others in each step. For any matrix $X$, denote $S_{\varepsilon}$ and $H_{\varepsilon}$ the soft-thresholding operators on $X$ such that

$$
[S_{\varepsilon}(X)]_{ij} = \begin{cases} 
X_{ij} - \varepsilon, & \text{if } X_{ij} > \varepsilon; \\
X_{ij} + \varepsilon, & \text{if } X_{ij} < -\varepsilon; \\
0, & \text{otherwise.}
\end{cases}
$$

and

$$
[H_{\varepsilon}(X)]_{ij} = \begin{cases} 
\frac{|X_{ij}| - \varepsilon}{\|X_{ij}\|_2} X_{ij}, & \text{if } \|X_{ij}\|_2 > \varepsilon; \\
0, & \text{otherwise.}
\end{cases}
$$

The detailed procedures of the ADM are listed in the following algorithm:

**Algorithm 1** The ADM for R-PCA via Outlier Pursuit

**Input:** Observation matrix $M \in \mathbb{R}^{m \times n}$, $\lambda = 1/\sqrt{\log n}$.

**Initialize:** $Y_0 = 0$; $L_0 = M$; $S_0 = 0$; $\mu_0 > 0$; $k = 0$.

1: while not converged do
2: //Line 3-4 solve $L_{k+1} = \text{arg min}_L \mathcal{L}(L, S_k, Y_k, \mu_k)$.
3: $(U, S, V) = \text{svd}(M - S_k + \mu_k^{-1} Y_k)$.
4: $L_{k+1} = U \mu_k^{-1} (S) V^T$.
5: //Line 6 solves $S_{k+1} = \text{arg min}_S \mathcal{L}(L_{k+1}, S, Y_k, \mu_k)$.
6: $S_{k+1} = H_{\mu_k^{-1}}[M - L_{k+1} + \mu_k^{-1} Y_k]$.
7: $Y_{k+1} = Y_k + \mu_k(M - L_{k+1} - S_{k+1})$.
8: Update $\mu_k$ to $\mu_{k+1}$.
9: $k \leftarrow k + 1$.
10: end while

**Output:** $(L^*, S^*)$.

**Appendices**

**Equivalence of Probabilistic Models**

We show that the exact recovery result proved for the Bernoulli distribution holds for the uniform distribution as well. Let “success” be the event that the algorithm succeeds, i.e., $\text{Range}(L_0) = \text{Range}(L^*)$ and $\{j : S^*_j \not\in \text{Range}(L^*)\} = Z_0$. Notice the fact that $\mathbb{P}_{\text{Berm}(p)}(\text{Success} \mid |Z| = k) = \mathbb{P}_{\text{Unif}(k)}(\text{Success})$, and Theorem 2 implies that for $k \geq t$,

$$
\mathbb{P}_{\text{Unif}(k)}(\text{Success}) \leq \mathbb{P}_{\text{Unif}(t)}(\text{Success}).
$$

Thus we have

$$
\mathbb{P}_{\text{Berm}(p)}(\text{Success}) = \sum_{k=0}^{n} \mathbb{P}_{\text{Berm}(p)}(\text{Success} \mid |Z| = k) \mathbb{P}_{\text{Berm}(p)}(|Z| = k)
\leq \sum_{k=0}^{n} \mathbb{P}_{\text{Berm}(p)}(|Z| = k) + \sum_{k=t}^{n} \mathbb{P}_{\text{Berm}(p)}(\text{Success} \mid |Z| = k) \mathbb{P}_{\text{Berm}(p)}(|Z| = k)
\leq \sum_{k=0}^{n} \mathbb{P}_{\text{Berm}(p)}(|Z| = k) + \sum_{k=t}^{n} \mathbb{P}_{\text{Unif}(k)}(\text{Success}) \mathbb{P}_{\text{Berm}(p)}(|Z| = k)
\leq \mathbb{P}_{\text{Berm}(p)}(|Z| < t) + \mathbb{P}_{\text{Unif}(t)}(\text{Success}).
$$

Taking $p = t/n + \varepsilon$ gives $\mathbb{P}_{\text{Berm}(p)}(|Z| < t) \leq \exp(-\frac{\varepsilon^2 n}{2p})$, which completes the proof.

**Proof of Theorem 3**

We proceed to prove Theorem 3. The following lemma is critical.

**Lemma 3.** Assume $\|\sum_i y_{ij} \otimes y_{ij}\| \leq 1$ for $y_{ij} \in \mathbb{R}^d$ and $\delta_j$'s are i.i.d. Bernoulli variables with $\mathbb{P}(\delta_j = 1) = a$. Then

$$
\mathbb{E} \left[ a^{-1} \left\| \sum_j (\delta_j - a) \sum_i y_{ij} \otimes y_{ij} \right\| \right] \leq C \sqrt{\frac{\log d}{a}} \max_{ij} ||y_{ij}||,
$$

provided that $C \sqrt{\log d/a} \max_{ij} ||y_{ij}|| < 1$.

**Proof.** Let

$$
Y = \sum_j (\delta_j - a) \sum_i y_{ij} \otimes y_{ij},
$$

and let $Y' = \sum_j (\delta'_j - a) \sum_i y_{ij} \otimes y_{ij}$ be an independent copy of $Y$. Since $\delta_j - \delta'_j$ is symmetric, $Y - Y'$ has the same distribution as

$$
Y_k - Y'_k \triangleq \sum_{ij} \varepsilon_{ij} (\delta_j - \delta'_j) y_{ij} \otimes y_{ij},
$$

where $\varepsilon_{ij}$'s are i.i.d. Rademacher variables and

$$
Y_k = \sum_{ij} \varepsilon_{ij} \delta_j y_{ij} \otimes y_{ij}.
$$

Notice that $|| \cdot ||$ is a convex function and $\mathbb{E}_{\delta'} Y' = 0$. Thus by Jensen’s inequality, we have

$$
\mathbb{E}_{\delta} ||Y|| = \mathbb{E}_{\delta} ||Y - \mathbb{E}_{\delta'} Y'||
\leq \mathbb{E}_{\delta} ||\mathbb{E}_{\delta'} (Y - Y')||
\leq \mathbb{E}_{\delta} ||Y - Y'||
\leq ||Y - Y'||
\leq ||Y - Y'||
= 2\mathbb{E} ||Y_k||
\leq 2 \mathbb{E} \left[ \sum_{ij} \varepsilon_{ij} \delta_j y_{ij} \otimes y_{ij} \right].
$$
According to Rudelson’s lemma in (Rudelson 1999), which states that
\[
\mathbb{E}_e \left| \sum_{ij} \varepsilon_{ij} \delta_{ij} y_{ij} \otimes y_{ij} \right| \leq C \sqrt{\log d} \max_{ij} \|y_{ij}\| \left( \sum_{ij} \delta_{ij} y_{ij} \otimes y_{ij} \right)^{\frac{1}{2}},
\]
we have
\[
\mathbb{E}_d \mathbb{E}_e \left| \sum_{ij} \varepsilon_{ij} \delta_{ij} y_{ij} \otimes y_{ij} \right| \leq C \sqrt{\log d} \max_{ij} \|y_{ij}\| \mathbb{E}_e \left( \sum_{ij} \delta_{ij} y_{ij} \otimes y_{ij} \right)^{\frac{1}{2}}.
\]
Hence
\[
\mathbb{E}\|Y\| \leq 2C \sqrt{\log d} \max_{ij} \|y_{ij}\| \mathbb{E}_d \left( \sum_{ij} \delta_{ij} y_{ij} \otimes y_{ij} \right)^{\frac{1}{2}}.
\]
Thus we have
\[
a^{-1} \mathbb{E}\|Y\| \leq 2C \log d \max_{ij} \|y_{ij}\| \sqrt{a^{-1} \mathbb{E}\|Y\| + \sum_{ij} \|y_{ij}\|}. \]
When \(2C \sqrt{\log d} \max_{ij} \|y_{ij}\|/\sqrt{a} < 1\), then
\[
a^{-1} \mathbb{E}\|Y\| \leq 2 \frac{C \log d}{\sqrt{a}} \max_{ij} \|y_{ij}\| \leq \tilde{C} \sqrt{\frac{\log d}{a}} \max_{ij} \|y_{ij}\|.
\]
and the proof finishes.

The following concentration inequality is also important to our proof of Theorem 3.

**Theorem 5** (Talagrand (1996)). Assume that \(|f| \leq B\) and \(\mathbb{E}f(Y_i) = 0\) for every \(f \in \mathcal{F}\), where \(i = 1, \ldots, n\) and \(\mathcal{F}\) is a countable family of functions such that if \(f \in \mathcal{F}\) then \(-f \in \mathcal{F}\). Let \(Y_* = \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(Y_i)\). Then for any \(t \geq 0\),
\[
\mathbb{P}(|Y_* - \mathbb{E}Y_*| > t) \leq 3 \exp \left( -\frac{t^2}{KB} \log \left( 1 + \frac{Bt}{\sigma^2 + B \mathbb{E}Y_*} \right) \right),
\]
where \(\sigma^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^n \mathbb{E}f^2(Y_i)\), and \(K\) is a constant.

Now we are ready to prove Theorem 3.

**Proof.** For any matrix \(X\), we have
\[
\mathcal{P}_\psi X = \sum_{ij} (\mathcal{P}_\psi X, e_i e_j^T) e_i e_j^T.
\]
Thus \(\mathcal{P}_\psi \mathcal{P}_\psi X = \sum_{ij} \delta_{ij} (\mathcal{P}_\psi X, e_i e_j^T) e_i e_j^T\), where \(\delta_{ij}\) are i.i.d. Bernoulli variables with parameter \(a\). Then
\[
\mathcal{P}_\psi \mathcal{P}_\psi \mathcal{P}_\psi X = \sum_{ij} \delta_{ij} (\mathcal{P}_\psi X, e_i e_j^T) \mathcal{P}_\psi (e_i e_j^T).
\]
Namely, \(\mathcal{P}_\psi \mathcal{P}_\psi \mathcal{P}_\psi = \sum_{ij} \delta_{ij} (\mathcal{P}_\psi (e_i e_j^T) \otimes \mathcal{P}_\psi (e_i e_j^T))\). Now let
\[
Z = a^{-1} \left( \mathcal{P}_\psi \mathcal{P}_\psi \mathcal{P}_\psi - a \mathcal{P}_\psi \right) = a^{-1} \left( \sum_{ij} (\delta_{ij} - a) \mathcal{P}_\psi (e_i e_j^T) \otimes \mathcal{P}_\psi (e_i e_j^T) \right).
\]
We first prove the upper bound of \(\mathbb{E}Z\). Adopt \(y_{ij} = \sum_{ij} \mathcal{P}_\psi (e_i e_j^T)\) in Lemma 3. Since
\[
\mathcal{P}_\psi = \sum_{ij} \mathcal{P}_\psi (e_i e_j^T) \otimes \mathcal{P}_\psi (e_i e_j^T),
\]
we have
\[
\left( \sum_{ij} \mathcal{P}_\psi (e_i e_j^T) \otimes \mathcal{P}_\psi (e_i e_j^T) \right) = 1.
\]
Thus by Lemma 3 and incoherence (1a),
\[
\mathbb{E}Z \leq \tilde{C} \sqrt{\frac{\log n^2}{a}} \sqrt{\frac{\mu r}{n}} \leq C \sqrt{\frac{\mu r \log n}{na}}.
\]

We then prove the upper bound of \(Z\) at an overwhelming probability. Let
\[
D_j = a^{-1} (\delta_{ij} - a) \sum \mathcal{P}_\psi (e_i e_j^T) \otimes \mathcal{P}_\psi (e_i e_j^T),
\]
and
\[
D = \sum_j D_j = a^{-1} (\mathcal{P}_\psi \mathcal{P}_\psi \mathcal{P}_\psi - a \mathcal{P}_\psi).
\]
Notice that the operator $D$ is self-adjoint. Denote the set $g = \{||X_1||_F \leq 1, X_2 = \pm X_1\}$. Then we have

$$Z = \sup_g \{X_1, D_j(X_2)\}$$

$$= \sup_g \sum_j \{X_1, D_j(X_2)\}$$

$$= \sup_g \sum_j a^{-1}(\delta_j - a) \sum_i \langle X_1, P_{\psi}(e_i e_j^T) \rangle \langle X_2, P_{\psi}(e_i e_j^T) \rangle.$$

Now let

$$f(\delta_j) = \langle X_1, D_j(X_2) \rangle$$

$$= a^{-1}(\delta_j - a) \sum_i \langle X_1, P_{\psi}(e_i e_j^T) \rangle \langle X_2, P_{\psi}(e_i e_j^T) \rangle.$$

To use Talagrand’s concentration inequality on $Z$, we should bound $|f(\delta_j)|$ and $\mathbb{E}f^2(\delta_j)$. Since by assumption, $L = L_0 + \hat{P}_{\eta_0}.P_{\eta_0}, H$ satisfies incoherence $(1a)$ and

$$||P_{\psi}X_1||^2_{2,\infty} = \max_j \sum_i ||X_1, e_i e_j^T \hat{V} \hat{V}^T||^2$$

$$= \max_j \sum_i ||e_i^T X_1, e_j^T \hat{V} \hat{V}^T||^2$$

$$\leq \max_j \sum_i ||e_i^T X_1||^2_2 ||\hat{V} \hat{V}^T||^2$$

$$= \max_j ||X_1||^2_2 ||\hat{V} \hat{V}^T||^2$$

$$\leq \frac{\mu_r}{n},$$

we have

$$|f(\delta_j)| \leq a^{-1}|\delta_j - a| \sum_i ||X_1, P_{\psi}(e_i e_j^T)|| \langle X_2, P_{\psi}(e_i e_j^T) \rangle$$

$$= a^{-1}|\delta_j - a| \sum_i \langle X_1, P_{\psi}(e_i e_j^T) \rangle^2$$

$$\leq a^{-1} \sum_i \langle P_{\psi}X_1, e_i e_j^T \rangle^2$$

$$\leq a^{-1} ||P_{\psi}X_1||^2_{2,\infty}$$

$$\leq \frac{\mu_r}{na},$$

where the first equality holds since $X_2 = \pm X_1$. Further-

more,

$$\mathbb{E}f^2(\delta_j) = a^{-1}(1 - a) \left( \sum_i \langle X_1, P_{\psi}(e_i e_j^T) \rangle \langle X_2, P_{\psi}(e_i e_j^T) \rangle \right)^2$$

$$\leq a^{-1}(1 - a) \left( \sum_i ||X_1, P_{\psi}(e_i e_j^T)|| \langle X_2, P_{\psi}(e_i e_j^T) \rangle \right)^2$$

$$= a^{-1}(1 - a) \left( \sum_i \langle X_1, P_{\psi}(e_i e_j^T) \rangle^2 \right)^2$$

$$\leq a^{-1} \left( \sum_i \langle P_{\psi}X_1, e_i e_j^T \rangle^2 \right) \left( \sum_i \langle P_{\psi}X_1, e_i e_j^T \rangle^2 \right)$$

$$\leq a^{-1} \langle P_{\psi}X_1||^2_2, \sum_i \langle P_{\psi}X_1, e_i e_j^T \rangle^2$$

$$\leq \frac{\mu_r}{na} \sum_i \langle P_{\psi}X_1, e_i e_j^T \rangle^2,$$

and

$$\sigma^2 = \mathbb{E} \sum_j f^2(\delta_j) \leq \frac{\mu_r}{na} \sum_{ij} \langle P_{\psi}X_1, e_i e_j^T \rangle^2$$

$$= \frac{\mu_r}{na} ||P_{\psi}X_1||^2_F$$

$$\leq \frac{\mu_r}{na}.$$

Since we have proved $\mathbb{E}Z \leq 1$ in the first part of the proof, by Theorem 5,

$$\mathbb{P}(|Z - \mathbb{E}Z| > t) \leq 3 \exp \left(-\frac{t}{KB} \log \left(1 + \frac{t}{K} \right)\right)$$

$$\leq 3 \exp \left(-\frac{t \log 2}{KB} \min \left(1, \frac{t}{2}\right)\right),$$

where the second inequality holds since $\log(1 + u) \geq \log 2 \min(1, u)$ for any $u \geq 0$. Set

$$B = \frac{\mu_r}{na} \quad \text{and} \quad t = \alpha \sqrt{\frac{\mu_r \log n}{na}},$$

we have

$$\mathbb{P} \left(|Z - \mathbb{E}Z| > \alpha \sqrt{\frac{\mu_r \log n}{na}}\right)$$

$$\leq 3 \exp \left( -\gamma_0 \min \left(2\alpha \sqrt{\frac{na \log n}{\mu_r}}, \alpha^2 \log n\right)\right)$$

$$= 3 \exp(-\gamma_0 \alpha^2 \log n),$$

where $\gamma_0 = \log 2/(2K)$ is a numerical constant. We now adopt $\alpha = \sqrt{\beta}/\gamma_0$. Thus

$$\mathbb{P} \left(|Z - \mathbb{E}Z| \leq \sqrt{\frac{\beta}{\gamma_0} \sqrt{\frac{\mu_r \log n}{na}}}\right) \geq 1 - 3n^{-\beta}.$$
Note that we have proved $E Z \leq C \sqrt{\mu r \log n / na}$. We have

$$
P(Z \leq \varepsilon) \geq P\left(Z \leq \sqrt{C_0} \sqrt{\frac{\mu r \log n}{na}}\right)
= P\left(Z \leq \left(C + \sqrt{\frac{\beta \gamma}{\gamma_0}}\right) \sqrt{\frac{\mu r \log n}{na}}\right)
\geq P\left(|Z - E Z| \leq \sqrt{\frac{\beta \gamma}{\gamma_0}} \sqrt{\frac{\mu r \log n}{na}}\right)
\geq 1 - 3n^{-\beta},
$$

where $C_0 \equiv (C + \sqrt{\beta / \gamma_0})^2$ and the first inequality holds since $a \geq C_0 \varepsilon^{-2}(\mu r \log n)/n$ by assumption. Thus the proof completes.

References


