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

Paper ID 1532

Robust PCA via Outlier Pursuit:

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

μ-incoherence condition on matrix $L = U\Sigma V^T$:

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

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

$$||UV^T||_\infty \leq \sqrt{\frac{\mu r}{mn}}. \quad (1c)$$

Ambiguity condition on matrix S:

$$||\mathcal{B}(S)|| \leq \sqrt{\log n}/4. \quad (2)$$

Main Results:

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* (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 $1 - cn^{-10}$, if the column support $\mathcal{I}_0$ of S_0 is uniformly distributed among all sets of cardinality s and*

$$\text{rank}(L_0) \leq \rho_r \frac{n_{(2)}}{\mu \log n} \quad \text{and} \quad s \leq \rho_s n, \quad (3)$$

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 (1a)*, and $S_0 - \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0} H$ 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., Range(L^*) = Range(L_0) and $\{j : S^*_{:j} \notin$ Range$(L^*)\} = \mathcal{I}_0$. Then any solution (L'^*, S'^*) to* (1) *with input $M' = L^* + \mathcal{P}_\mathcal{I} S^*$ succeeds as well, where $\mathcal{I} \subseteq \mathcal{I}^* = \mathcal{I}_0$.*

Table 1: Summary of main notations used in the supplementary material.

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, $\|\|M\|\|_{2,1} = \sqrt{\sum_j \|\|M_{:j}\|\|_2^2}$.
$\|\|\cdot\|\|_{2,\infty}$	$\ell_{2,\infty}$ 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.
L^*, S^*	Optimal solutions to Outlier Pursuit.
L_0, S_0	Ground Truth.
$\hat{U}, \hat{V}$	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^*.
$\hat{\mathcal{T}}$	$\hat{\mathcal{T}} = \{\hat{U}X^T + Y\hat{V}^T, \forall X, Y \in \mathbb{R}^{n\times r}\}$.
$\mathcal{X}^\perp$	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^*.
$\|\mathcal{I}_0\|$	Outliers number of S_0.
$X \in \mathcal{I}$	The column support of X is a subset of $\mathcal{I}$.
$\mathcal{B}(\hat{S})$	$\mathcal{B}(\hat{S}) = \{H : \mathcal{P}_{\hat{\mathcal{I}}^\perp}(H) = 0; H_{:j} = \frac{\hat{S}_{:j}}{\|\|\hat{S}_{:j}\|\|_2}, j \in \hat{\mathcal{I}}\}$.
$\sim$ Ber(p)	Obeys Bernoulli distribution with parameter p.
$\mathcal{N}(a, b^2)$	Gaussian distribution (mean a and variance b^2).

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

$$||L'^*||_* + \lambda||S'^*||_{2,1} \le ||L^*||_* + \lambda||\mathcal{P}_{\mathcal{I}} S^*||_{2,1}.$$

Therefore

$$\begin{aligned} &||L'^*||_* + \lambda||S'^* + \mathcal{P}_{\mathcal{I}^\perp \cap \mathcal{I}_0} S^*||_{2,1} \\ \le\, &||L'^*||_* + \lambda||S'^*||_{2,1} + \lambda||\mathcal{P}_{\mathcal{I}^\perp \cap \mathcal{I}_0} S^*||_{2,1} \\ \le\, &||L^*||_* + \lambda||\mathcal{P}_{\mathcal{I}} S^*||_{2,1} + \lambda||\mathcal{P}_{\mathcal{I}^\perp \cap \mathcal{I}_0} S^*||_{2,1} \\ =\, &||L^*||_* + \lambda||S^*||_{2,1}. \end{aligned}$$

Note that

$$L'^* + S'^* + \mathcal{P}_{\mathcal{I}^\perp \cap \mathcal{I}_0} S^* = M' + \mathcal{P}_{\mathcal{I}^\perp \cap \mathcal{I}_0} S^* = M.$$

Thus $(L'^*, S'^* + \mathcal{P}_{\mathcal{I}^\perp \cap \mathcal{I}_0} S^*)$ is optimal to problem with input M and by assumption we have

$$\text{Range}(L'^*) = \text{Range}(L^*) = \text{Range}(L_0),$$

$$\{j : [S'^* + \mathcal{P}_{\mathcal{I}^\perp \cap \mathcal{I}_0} S^*]_{:j} \notin \text{Range}(L_0)\} = \text{Supp}(S_0).$$

The second equation implies $\mathcal{I} \subseteq \{j : S'^*_{:j} \notin \text{Range}(L_0)\}$. Suppose $\mathcal{I} \neq \{j : S'^*_{:j} \notin \text{Range}(L_0)\}$. Then there exists an index k such that $S'^*_{:k} \notin \text{Range}(L_0)$ and $k \notin \mathcal{I}$, i.e., $M'_{:k} = L^*_{:k} \in \text{Range}(L_0)$. Note that $L'^*_{:j} \in \text{Range}(L_0)$. Thus $S'^*_{:k} \in \text{Range}(L_0)$ and we have a contradiction. Thus $\mathcal{I} = \{j : S'^*_{:j} \notin \text{Range}(L_0)\} = \{j : S'^*_{:j} \notin \text{Range}(L'^*)\}$ and the algorithm succeeds. □

Theorem 2 shows that the success of the algorithm is monotone on $|\mathcal{I}_0|$. 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 (1), $\hat{L} = L_0 + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0} H$ 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* (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 $\mathcal{I}_0 = \{j : S^*_{:j} \notin$ Range$(L^*)\}$), provided that there exists a pair (W, F) obeying*

$$W = \lambda(\mathcal{B}(\hat{S}) + F), \tag{4}$$

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

Proof. We first recall that the subgradients of nuclear norm and $\ell_{2,1}$ norm are as follows:

$$\partial_{\hat{L}}||\hat{L}||_* = \{\hat{U}\hat{V}^T + \hat{Q} : \hat{Q} \in \hat{\mathcal{T}}^\perp, ||\hat{Q}|| \le 1\},$$

$$\partial_{\hat{S}}||\hat{S}||_{2,1} = \{\mathcal{B}(\hat{S}) + \hat{E} : \hat{E} \in \hat{\mathcal{I}}^\perp, ||\hat{E}||_{2,\infty} \le 1\}.$$

Let $H_1 = \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0} H$ and $H_2 = \mathcal{P}_{\mathcal{I}_0^\perp}\mathcal{P}_{\mathcal{U}_0} H + \mathcal{P}_{\mathcal{I}_0^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0^\perp} H$, and note that $\hat{\mathcal{U}} = \mathcal{U}_0$ and $\hat{\mathcal{I}} = \mathcal{I}_0$. By the definition of the subgradient, the inequality follows

$$\begin{aligned} &||L_0 + H||_* + \lambda||S_0 - H||_{2,1} \\ \ge\, &||\hat{L}||_* + \lambda||\hat{S}||_{2,1} + \langle \hat{U}\hat{V}^T + \hat{Q}, H_2\rangle - \lambda\langle \mathcal{B}(\hat{S}) + \hat{E}, H_2\rangle \\ =\, &||\hat{L}||_* + \lambda||\hat{S}||_{2,1} + \langle \hat{U}\hat{V}^T, \mathcal{P}_{\mathcal{I}_0^\perp} H\rangle + \langle \hat{Q}, \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle - \\ &\lambda\langle \mathcal{B}(\hat{S}), \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle - \lambda\langle \hat{E}, \mathcal{P}_{\mathcal{I}_0^\perp} H\rangle \\ \ge\, &||\hat{L}||_* + \lambda||\hat{S}||_{2,1} - \sqrt{\frac{\mu r}{n}}||\mathcal{P}_{\mathcal{I}_0^\perp} H||_{2,1} + \langle \hat{Q}, \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle - \\ &\lambda\langle \mathcal{B}(\hat{S}), \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle - \lambda\langle \hat{E}, \mathcal{P}_{\mathcal{I}_0^\perp} H\rangle. \end{aligned}$$

Now adopt $\hat{Q}$ such that $\langle \hat{Q}, \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle = ||\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_*$ and $\langle \hat{E}, \mathcal{P}_{\mathcal{I}_0^\perp} H\rangle = -||\mathcal{P}_{\mathcal{I}_0^\perp} H||_{2,1}$[1]. We have

$$\begin{aligned} &||L_0 + H||_* + \lambda||S_0 - H||_{2,1} \\ \ge\, &||\hat{L}||_* + \lambda||\hat{S}||_{2,1} - \sqrt{\frac{\mu r}{n}}||\mathcal{P}_{\mathcal{I}_0^\perp} H||_{2,1} + ||\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_* - \\ &\lambda\langle \mathcal{B}(\hat{S}), \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle + \lambda||\mathcal{P}_{\mathcal{I}_0^\perp} H||_{2,1} \\ \ge\, &||\hat{L}||_* + \lambda||\hat{S}||_{2,1} + \left(\frac{\lambda}{4} - \sqrt{\frac{\mu r}{n}}\right)||\mathcal{P}_{\mathcal{I}_0^\perp} H||_{2,1} + \\ &||\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_* - \lambda\langle \mathcal{B}(\hat{S}), \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle + \frac{3\lambda}{4}||\mathcal{P}_{\hat{\mathcal{I}}^\perp} H||_{2,1} \\ \ge\, &||\hat{L}||_* + \lambda||\hat{S}||_{2,1} + \left(\frac{\lambda}{4} - \sqrt{\frac{\mu r}{n}}\right)||\mathcal{P}_{\mathcal{I}_0^\perp} H||_{2,1} + \\ &||\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_* - \lambda\langle \mathcal{B}(\hat{S}), \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle + \frac{3\lambda}{4}||\mathcal{P}_{\hat{\mathcal{I}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_{2,1}. \end{aligned}$$

Notice that

$$\begin{aligned} |\langle -\lambda\mathcal{B}(\hat{S}), \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle| &= |\langle \lambda F - W, \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle| \\ &\le |\langle W, \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle| + \lambda|\langle F, \mathcal{P}_{\mathcal{U}_0^\perp} H\rangle| \\ &\le \frac{1}{2}||\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_* + \frac{\lambda}{2}||\mathcal{P}_{\hat{\mathcal{I}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_{2,1}. \end{aligned}$$

Hence

$$\begin{aligned} &||L_0 + H||_* + \lambda||S_0 - H||_{2,1} \\ \ge\, &||\hat{L}||_* + \lambda||\hat{S}||_{2,1} + \left(\frac{\lambda}{4} - \sqrt{\frac{\mu r}{n}}\right)||\mathcal{P}_{\mathcal{I}_0^\perp} H||_{2,1} + \\ &\frac{1}{2}||\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_* + \frac{\lambda}{4}||\mathcal{P}_{\hat{\mathcal{I}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_{2,1}. \end{aligned}$$

[1]By the duality between the nuclear norm and the operator norm, there exists a Q such that $\langle Q, \mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H\rangle = ||\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp} H||_*$ and $||Q|| \le 1$. Thus we take $\hat{Q} = \mathcal{P}_{\mathcal{U}_0^\perp}\mathcal{P}_{\hat{\mathcal{V}}^\perp} Q \in \hat{\mathcal{T}}^\perp$. It holds similarly for $\hat{E}$.

Since $(L^*, S^*) = (L_0 + H, S_0 - H)$ is optimal, above inequality shows $||\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp}H||_* = ||\mathcal{P}_{\hat{\mathcal{I}}^\perp}\mathcal{P}_{\mathcal{U}_0^\perp}H||_{2,1} = 0$, i.e., $\mathcal{P}_{\mathcal{U}_0^\perp}H \in \hat{\mathcal{I}} \cap \hat{\mathcal{V}}$. Also notice that $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| < 1$ implies $\hat{\mathcal{I}} \cap \hat{\mathcal{V}} = \{\mathbf{0}\}$. We conclude $\mathcal{P}_{\mathcal{U}_0^\perp}H = \mathbf{0}$. Furthermore, $||\mathcal{P}_{\mathcal{I}_0^\perp}H||_{2,1} = 0$ implies $H \in \mathcal{I}_0$. Thus $H \in \mathcal{U}_0 \cap \mathcal{I}_0$, i.e., $\mathcal{U}^* \subseteq \mathcal{U}_0$ and $\mathcal{I}^* \subseteq \mathcal{I}_0$.

We now prove $\mathcal{U}^* = \mathcal{U}_0$. According to the assumption $\text{Range}(L_0) = \text{Range}(P_{I_0^\perp}L_0)$ and $H \in \mathcal{U}_0 \cap \mathcal{I}_0$, $\text{Range}(L^*) = \text{Range}(L_0 + H) = \text{Range}(L_0)$, i.e., $\mathcal{U}^* = \mathcal{U}_0$. We then prove $\mathcal{I}^* = \mathcal{I}_0$. Assume that $\mathcal{I}^* \neq \mathcal{I}_0$, i.e., there exists a $j \in \mathcal{I}_0$ such that $S^*_{:j} = \mathbf{0}$. 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 $\mathcal{U}^* \subseteq \mathcal{U}_0$. So $\mathcal{I}^* = \mathcal{I}_0$. □

Remark 1. *There are two important modifications in our conditions compared with those of (Xu, Caramanis, and Sanghavi 2012): 1. The space $\hat{\mathcal{T}}$ (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 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

$$\begin{cases} W \in \hat{\mathcal{V}}^\perp, \\ ||W|| \leq 1/2, \\ \mathcal{P}_{\hat{\mathcal{I}}}W = \lambda\mathcal{B}(\hat{S}), \\ ||\mathcal{P}_{\hat{\mathcal{I}}^\perp}W||_{2,\infty} \leq \lambda/2. \end{cases} \tag{5}$$

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

$$W = \lambda\mathcal{P}_{\hat{\mathcal{V}}^\perp}\sum_{k\geq 0}(\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}})^k\mathcal{B}(\hat{S}). \tag{6}$$

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 (6) 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 $\sum_{k\geq 0}(\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{T}}}\mathcal{P}_{\hat{\mathcal{I}}})^k$ 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 considers that the elements in the same column of $\hat{S}$ are not independent. A complete proof can be found in Appendices.

Theorem 3. *For any $\mathcal{I} \sim Ber(a)$, with an overwhelming probability*

$$||\mathcal{P}_{\hat{\mathcal{V}}} - a^{-1}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\mathcal{I}}\mathcal{P}_{\hat{\mathcal{V}}}|| < \varepsilon, \tag{7}$$

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

By Theorem 3, our bounds in Theorem 1 guarantee a is 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 3, 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

$$\begin{aligned} ||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}||^2 &= ||\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| \\ &\leq ||\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}} - p\mathcal{P}_{\hat{\mathcal{V}}}|| + ||p\mathcal{P}_{\hat{\mathcal{V}}}|| \\ &\leq (1-p)\varepsilon + p. \end{aligned} \tag{8}$$

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 $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}||^2 \leq (1-p)\varepsilon + p$, provided that $1 - p \geq C_0\varepsilon^{-2}(\mu r \log n)/n$ for some numerical constant $C_0 > 0$.*

Note that $\mathcal{P}_{\hat{\mathcal{I}}}W = \lambda\mathcal{B}(\hat{S})$ and $W \in \hat{\mathcal{V}}^\perp$. So to prove the dual conditions (5), it is sufficient to show that

$$\begin{aligned} &\text{(a)}\ ||W|| \leq 1/2, \\ &\text{(b)}\ ||\mathcal{P}_{\hat{\mathcal{I}}^\perp}W||_{2,\infty} \leq \lambda/2. \end{aligned} \tag{9}$$

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

Proof. Let $\mathcal{R} = \sum_{k\geq 1}(\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}})^k$. Then

$$\begin{aligned} W &= \lambda\mathcal{P}_{\hat{\mathcal{V}}^\perp}\sum_{k\geq 0}(\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_{\hat{\mathcal{I}}})^k\mathcal{B}(\hat{S}) \\ &= \lambda\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{B}(\hat{S}) + \lambda\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{R}(\mathcal{B}(\hat{S})), \end{aligned} \tag{10}$$

Now we check the two conditions in (9).

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

$$\lambda||\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{B}(\hat{S})|| \leq \lambda||\mathcal{B}(\hat{S})|| \leq \frac{1}{4}. \tag{11}$$

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

$$||\mathcal{R}(\mathcal{B}(\hat{S}))|| \leq 4\sup_{x,y\in\mathcal{N}}\langle y, \mathcal{R}(\mathcal{B}(\hat{S}))x\rangle. \tag{12}$$

Note that the operator $\mathcal{R}$ is self-adjoint. Now let

$$\begin{aligned} X(x,y) &= \langle y, \mathcal{R}(\mathcal{B}(\hat{S}))x\rangle \\ &= \langle \mathcal{R}(yx^T), \mathcal{B}(\hat{S})\rangle \\ &= \sum_j \langle [\mathcal{R}(yx^T)]_{:j}, \mathcal{B}(\hat{S})_{:j}\rangle \\ &= \sum_j \langle [\mathcal{R}(yx^T)]_{:j}, \delta_j B_{:j}\rangle \\ &= \sum_j \delta_j B_{:j}^T [\mathcal{R}(yx^T)]_{:j}, \end{aligned}$$

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

$$\delta_j = \begin{cases} 1, & w.p.\ p, \\ 0, & w.p.\ 1-p. \end{cases} \tag{13}$$

Notice that

$$\begin{aligned} |\delta_j B_{:j}^T [\mathcal{R}(yx^T)]_{:j}|^2 &\le ||B_{:j}||_2^2 ||[\mathcal{R}(yx^T)]_{:j}||_2^2 \\ &= ||[\mathcal{R}(yx^T)]_{:j}||_2^2, \end{aligned}$$

and

$$\begin{aligned} \sum_{j=1}^n ||[\mathcal{R}(yx^T)]_{:j}||_2^2 &= ||\mathcal{R}(yx^T)||_F^2 \\ &\le ||\mathcal{R}||^2 ||yx^T||_F^2 \\ &= ||\mathcal{R}||^2 ||y||_2^2 ||x||_2^2 \\ &= ||\mathcal{R}||^2. \end{aligned}$$

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

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

As a result,

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

Namely, by inequality (12),

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

Now suppose $||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| \le \sigma$. Then

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

where γ can be sufficient large, and we have

$$\begin{aligned} &\mathbb{P}(||\mathcal{R}(\mathcal{B}(\hat{S}))|| > t) \\ &\le \mathbb{P}(||\mathcal{R}(\mathcal{B}(\hat{S}))|| > t \mid ||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| \le \sigma) + \mathbb{P}(||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| > \sigma) \\ &\le 2|\mathcal{N}|^2\text{exp}\left(-\frac{t^2}{32||\mathcal{R}||^2}\right) + \mathbb{P}(||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| > \sigma) \\ &\le 2|\mathcal{N}|^2\text{exp}\left(-\frac{\gamma^2 t^2}{32}\right) + \mathbb{P}(||\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}|| > \sigma), \end{aligned}$$

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

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

$$\begin{aligned} \mathcal{P}_{\hat{\mathcal{I}}^\perp} W &= \lambda\mathcal{P}_{\hat{\mathcal{I}}^\perp}\mathcal{P}_{\hat{\mathcal{V}}^\perp}\mathcal{G}(\mathcal{B}(\hat{S})) \\ &= \lambda\mathcal{P}_{\hat{\mathcal{I}}^\perp}\mathcal{G}(\mathcal{B}(\hat{S})) - \lambda\mathcal{P}_{\hat{\mathcal{I}}^\perp}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{G}(\mathcal{B}(\hat{S})) \\ &= -\lambda\mathcal{P}_{\hat{\mathcal{I}}^\perp}\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{G}(\mathcal{B}(\hat{S})). \end{aligned}$$

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

$$\begin{aligned} ||Q_{:j}||^2 &= \sum_i Q_{ij}^2 = \sum_i \langle \mathcal{P}_{\hat{\mathcal{V}}}\mathcal{G}(\mathcal{B}(\hat{S})), e_i e_j^T\rangle^2 \\ &= \sum_i \langle \mathcal{B}(\hat{S}), \mathcal{G}\mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle^2 \\ &= \sum_i \sum_{j_0} \langle [\mathcal{B}(\hat{S})]_{:j_0}, \mathcal{G}\mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T) e_{j_0}\rangle^2 \\ &= \sum_{j_0} \sum_i \langle e_i^T [\mathcal{B}(\hat{S})]_{:j_0}, \mathcal{G}\mathcal{P}_{\hat{\mathcal{V}}}(e_j^T) e_{j_0}\rangle^2 \\ &= \sum_{j_0} \sum_i \langle e_i^T [\mathcal{B}(\hat{S})]_{:j_0}, \mathcal{G}\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}(e_j^T) e_{j_0}\rangle^2 \\ &\le \sum_{j_0} ||\mathcal{G}\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}(e_j^T) e_{j_0}||_2^2 \\ &= ||\mathcal{G}\mathcal{P}_{\hat{\mathcal{I}}}\mathcal{P}_{\hat{\mathcal{V}}}(e_j^T)||_2^2 \\ &\le \left(\frac{\sigma}{1-\sigma^2}\right)^2 \le \frac{1}{4}, \quad \forall j. \end{aligned}$$

Thus $||\mathcal{P}_{\hat{\mathcal{I}}^\perp} W||_{2,\infty} = \lambda||\mathcal{P}_{\hat{\mathcal{I}}^\perp} Q||_{2,\infty} \le \lambda||Q||_{2,\infty} \le \lambda/2$. □

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 $\text{rank}(L_0)$ is tight. McCoy and Tropp (2011) showed that the optimal solution L^* to model (1) satisfies

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

If the order of $\text{rank}(L_0)$ 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 $\text{rank}(L_0)$ should be no larger than $\Theta(n/\log n)$ and the order of our bound is tight. □

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

$$\min_{L,S} ||L||_* + \lambda||S||_{2,1}, \quad \text{s.t.} \quad M = L + S, \tag{15}$$

whose augmented Lagrangian formulation corresponds to

$$\begin{aligned} &\mathcal{L}(L, S, Y, \mu) \\ &= ||L||_* + \lambda||S||_{2,1} + \langle M - L - S, Y\rangle + \frac{\mu}{2}||M - L - S||_F^2. \end{aligned} \tag{16}$$

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

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

and

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

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: $Y_0 = \mathbf{0}$; $L_0 = M$; $S_0 = \mathbf{0}$; $\mu_0 > 0$; $k = 0$.
1: while not converged **do**
2: //Line 3-4 solve $L_{k+1} = \arg\min_L \mathcal{L}(L, S_k, Y_k, \mu_k)$.
3: $(U, S, V) = \text{svd}(M - S_k + \mu_k^{-1} Y_k)$;
4: $L_{k+1} = U\mathcal{S}_{\mu_k^{-1}}(S)V^T$.
5: //Line 6 solves $S_{k+1} = \arg\min_S \mathcal{L}(L_{k+1}, S, Y_k, \mu_k)$.
6: $S_{k+1} = \mathcal{H}_{\lambda\mu_k^{-1}}[M - L_{k+1} + \mu_k^{-1} Y_k]$.
7: $Y_{k+1} = Y_k + \mu_k(M - L_{k+1} - S_{k+1})$.
8: Update μ_k to μ_{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 $\{j : S^*_{:j} \notin \text{Range}(L^*)\} = \mathcal{I}_0$. Notice the fact that

$$\mathbb{P}_{\text{Ber}(p)}(\text{Success}|\ |\mathcal{I}| = k) = \mathbb{P}_{\text{Unif}(k)}(\text{Success}),$$

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

$$\begin{aligned} &\mathbb{P}_{\text{Ber}(p)}(\text{Success}) = \sum_{k=0}^{n} \mathbb{P}_{\text{Ber}(p)}(\text{Success}|\ |\mathcal{I}| = k)\mathbb{P}_{\text{Ber}(p)}(|\mathcal{I}| = k) \\ &\leq \sum_{k=0}^{t-1} \mathbb{P}_{\text{Ber}(p)}(|\mathcal{I}| = k) + \sum_{k=t}^{n} \mathbb{P}_{\text{Ber}(p)}(\text{Success}|\ |\mathcal{I}| = k)\mathbb{P}_{\text{Ber}(p)}(|\mathcal{I}| = k) \\ &\leq \sum_{k=0}^{t-1} \mathbb{P}_{\text{Ber}(p)}(|\mathcal{I}| = k) + \sum_{k=t}^{n} \mathbb{P}_{\text{Unif}(k)}(\text{Success})\mathbb{P}_{\text{Ber}(p)}(|\mathcal{I}| = k) \\ &\leq \mathbb{P}_{\text{Ber}(p)}(|\mathcal{I}| < t) + \mathbb{P}_{\text{Unif}(t)}(\text{Success}). \end{aligned}$$

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

Proof of Theorem 3

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

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

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

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

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

$$Y_\varepsilon - Y'_\varepsilon \triangleq \sum_{ij} \varepsilon_{ij}(\delta_j - \delta'_j) y_{ij} \otimes y_{ij},$$

where ε_{ij}s are i.i.d. Rademacher variables and

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

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

$$\begin{aligned} \mathbb{E}_\delta ||Y|| &= \mathbb{E}_\delta ||Y - \mathbb{E}_{\delta'} Y'|| \\ &= \mathbb{E}_\delta ||\mathbb{E}_{\delta'}(Y - Y')|| \\ &\leq \mathbb{E}_\delta \mathbb{E}_{\delta'} ||Y - Y'|| \\ &= \mathbb{E}||Y_\varepsilon - Y'_\varepsilon|| \\ &\leq \mathbb{E}||Y_\varepsilon|| + \mathbb{E}||Y'_\varepsilon|| \\ &= 2\mathbb{E}||Y_\varepsilon|| \\ &= 2\mathbb{E}\left|\left|\sum_{ij} \varepsilon_{ij}\delta_j y_{ij} \otimes y_{ij}\right|\right|. \end{aligned}$$

According to Rudelson's lemma in (Rudelson 1999), which states that

$$\mathbb{E}_\varepsilon \left|\left|\sum_{ij} \varepsilon_{ij}\delta_j y_{ij} \otimes y_{ij}\right|\right|$$

$$\leq C\sqrt{\log d}\ \max_{ij} ||y_{ij}||\ \left|\left|\sum_{ij} \delta_j y_{ij} \otimes y_{ij}\right|\right|^{\frac{1}{2}},$$

we have

$$\mathbb{E}_\delta \mathbb{E}_\varepsilon \left|\left|\sum_{ij} \varepsilon_{ij}\delta_j y_{ij} \otimes y_{ij}\right|\right|$$

$$\leq C\sqrt{\log d}\ \max_{ij} ||y_{ij}||\ \mathbb{E}_\delta \left|\left|\sum_{ij} \delta_j y_{ij} \otimes y_{ij}\right|\right|^{\frac{1}{2}}.$$

Hence

$$\begin{aligned}
\mathbb{E}||Y|| &\leq 2C\sqrt{\log d}\ \max_{ij} ||y_{ij}||\ \mathbb{E}_\delta \left|\left|\sum_{ij} \delta_j y_{ij} \otimes y_{ij}\right|\right|^{\frac{1}{2}} \\
&\leq 2C\sqrt{\log d}\ \max_{ij} ||y_{ij}|| \sqrt{\mathbb{E}\left|\left|\sum_{ij} \delta_j y_{ij} \otimes y_{ij}\right|\right|} \\
&= 2C\sqrt{\log d}\ \max_{ij} ||y_{ij}|| \sqrt{\mathbb{E}\left|\left|\sum_{j} \delta_j \sum_i y_{ij} \otimes y_{ij}\right|\right|} \\
&= 2C\sqrt{\log d}\ \max_{ij} ||y_{ij}|| \sqrt{\mathbb{E}\left|\left|Y + a\sum_{ij} y_{ij} \otimes y_{ij}\right|\right|} \\
&\leq 2C\sqrt{\log d}\ \max_{ij} ||y_{ij}|| \sqrt{\mathbb{E}||Y|| + a\left|\left|\sum_{ij} y_{ij} \otimes y_{ij}\right|\right|}.
\end{aligned}$$

Thus we have

$$\begin{aligned}
&a^{-1}\mathbb{E}||Y|| \\
&\leq \frac{2C\sqrt{\log d}}{\sqrt{a}} \max_{ij} ||y_{ij}|| \sqrt{a^{-1}\mathbb{E}||Y|| + \left|\left|\sum_{ij} y_{ij} \otimes y_{ij}\right|\right|} \\
&\leq \frac{2C\sqrt{\log d}}{\sqrt{a}} \max_{ij} ||y_{ij}|| \sqrt{a^{-1}\mathbb{E}||Y|| + 1}.
\end{aligned}$$

When $2C\sqrt{\log d} \max_{ij} ||y_{ij}||/\sqrt{a} < 1$, then

$$\begin{aligned}
a^{-1}\mathbb{E}||Y|| &\leq 2\frac{2C\sqrt{\log d}}{\sqrt{a}} \max_{ij} ||y_{ij}|| \\
&\triangleq \widetilde{C}\sqrt{\frac{\log d}{a}} \max_{ij} ||y_{ij}||,
\end{aligned}$$

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 $i = 1, ..., n$ 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 $t \geq 0$,*

$$\mathbb{P}(|Y_* - \mathbb{E}Y_*| > t) \leq 3\exp\left(-\frac{t}{KB}\log\left(1 + \frac{Bt}{\sigma^2 + B\mathbb{E}Y_*}\right)\right),$$

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

$$\mathcal{P}_{\hat{\mathcal{V}}}X = \sum_{ij} \langle \mathcal{P}_{\hat{\mathcal{V}}}X, e_i e_j^T\rangle e_i e_j^T.$$

Thus $\mathcal{P}_\mathcal{I}\mathcal{P}_{\hat{\mathcal{V}}}X = \sum_{ij} \delta_j \langle \mathcal{P}_{\hat{\mathcal{V}}}X, e_i e_j^T\rangle e_i e_j^T$, where where δ_js are i.i.d. Bernoulli variables with parameter a. Then

$$\begin{aligned}
\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_\mathcal{I}\mathcal{P}_{\hat{\mathcal{V}}}X &= \sum_{ij} \delta_j \langle \mathcal{P}_{\hat{\mathcal{V}}}X, e_i e_j^T\rangle \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T) \\
&= \sum_{ij} \delta_j \langle X, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T).
\end{aligned}$$

Namely, $\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_\mathcal{I}\mathcal{P}_{\hat{\mathcal{V}}} = \sum_{ij} \delta_j \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T) \otimes \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)$. Now let

$$\begin{aligned}
Z &= a^{-1}||\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_\mathcal{I}\mathcal{P}_{\hat{\mathcal{V}}} - a\mathcal{P}_{\hat{\mathcal{V}}}|| \\
&= a^{-1}\left|\left|\sum_{ij}(\delta_j - a)\mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T) \otimes \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\right|\right|.
\end{aligned}$$

We first prove the upper bound of $\mathbb{E}Z$. Adopt $y_{ij} = \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)$ in Lemma 3. Since

$$\mathcal{P}_{\hat{\mathcal{V}}} = \sum_{ij} \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T) \otimes \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T),$$

we have

$$\left|\left|\sum_{ij} \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T) \otimes \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\right|\right| = 1.$$

Thus by Lemma 3 and incoherence (1a),

$$\mathbb{E}Z \leq \widetilde{C}\sqrt{\frac{\log n^2}{a}}\sqrt{\frac{\mu r}{n}} \triangleq C\sqrt{\frac{\mu r \log n}{na}}.$$

We then prove the upper bound of Z at an overwhelming probability. Let

$$\mathcal{D}_j = a^{-1}(\delta_j - a)\sum_i \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T) \otimes \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T),$$

and

$$\mathcal{D} = \sum_j \mathcal{D}_j = a^{-1}(\mathcal{P}_{\hat{\mathcal{V}}}\mathcal{P}_\mathcal{I}\mathcal{P}_{\hat{\mathcal{V}}} - a\mathcal{P}_{\hat{\mathcal{V}}}).$$

Notice that the operator $\mathcal{D}$ is self-adjoint. Denote the set $g = \{||X_1||_F \le 1, X_2 = \pm X_1\}$. Then we have

$$
\begin{aligned}
Z &= \sup_g \langle X_1, \mathcal{D}(X_2)\rangle \\
&= \sup_g \sum_j \langle X_1, \mathcal{D}_j(X_2)\rangle \\
&= \sup_g \sum_j a^{-1}(\delta_j - a) \sum_i \langle X_1, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle \langle X_2, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle.
\end{aligned}
$$

Now let

$$
\begin{aligned}
f(\delta_j) &= \langle X_1, \mathcal{D}_j(X_2)\rangle \\
&= a^{-1}(\delta_j - a) \sum_i \langle X_1, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle \langle X_2, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle.
\end{aligned}
$$

To use Talagrand's concentration inequality on Z, we should bound $|f(\delta_j)|$ and $\mathbb{E}f^2(\delta_j)$. Since by assumption, $\hat{L} = L_0 + \mathcal{P}_{\mathcal{I}_0}\mathcal{P}_{\mathcal{U}_0}H$ satisfies incoherence (1a) and

$$
\begin{aligned}
||\mathcal{P}_{\hat{\mathcal{V}}}X_1||_{2,\infty}^2 &= \max_j \sum_i \langle X_1, e_i e_j^T \hat{V}\hat{V}^T\rangle^2 \\
&= \max_j \sum_i \langle e_i^T X_1, e_j^T \hat{V}\hat{V}^T\rangle^2 \\
&\le \max_j \sum_i ||e_i^T X_1||_2^2 ||e_j^T \hat{V}\hat{V}^T||_2^2 \\
&= \max_j ||X_1||_F^2 ||e_j^T \hat{V}\hat{V}^T||_2^2 \\
&\le \frac{\mu r}{n},
\end{aligned}
$$

we have

$$
\begin{aligned}
|f(\delta_j)| &\le a^{-1}|\delta_j - a| \sum_i |\langle X_1, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle| \, |\langle X_2, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle| \\
&= a^{-1}|\delta_j - a| \sum_i \langle X_1, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle^2 \\
&\le a^{-1} \sum_i \langle \mathcal{P}_{\hat{\mathcal{V}}}X_1, e_i e_j^T\rangle^2 \\
&\le a^{-1} ||\mathcal{P}_{\hat{\mathcal{V}}}X_1||_{2,\infty}^2 \\
&\le \frac{\mu r}{na},
\end{aligned}
$$

where the first equality holds since $X_2 = \pm X_1$. Furthermore,

$$
\begin{aligned}
\mathbb{E}f^2(\delta_j) &= a^{-1}(1-a)\left(\sum_i \langle X_1, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle \langle X_2, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle\right)^2 \\
&\le a^{-1}(1-a)\left(\sum_i |\langle X_1, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle \langle X_2, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle|\right)^2 \\
&= a^{-1}(1-a)\left(\sum_i \langle X_1, \mathcal{P}_{\hat{\mathcal{V}}}(e_i e_j^T)\rangle^2\right)^2 \\
&\le a^{-1}\left(\sum_i \langle \mathcal{P}_{\hat{\mathcal{V}}}X_1, e_i e_j^T\rangle^2\right)\left(\sum_i \langle \mathcal{P}_{\hat{\mathcal{V}}}X_1, e_i e_j^T\rangle^2\right) \\
&\le a^{-1}||\mathcal{P}_{\hat{\mathcal{V}}}X_1||_{2,\infty}^2 \sum_i \langle \mathcal{P}_{\hat{\mathcal{V}}}X_1, e_i e_j^T\rangle^2 \\
&\le \frac{\mu r}{na} \sum_i \langle \mathcal{P}_{\hat{\mathcal{V}}}X_1, e_i e_j^T\rangle^2,
\end{aligned}
$$

and

$$
\begin{aligned}
\sigma^2 = \mathbb{E}\sum_j f^2(\delta_j) &\le \frac{\mu r}{na} \sum_{ij} \langle \mathcal{P}_{\hat{\mathcal{V}}}X_1, e_i e_j^T\rangle^2 \\
&= \frac{\mu r}{na} ||\mathcal{P}_{\hat{\mathcal{V}}}X_1||_F^2 \\
&\le \frac{\mu r}{na}.
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{P}(|Z - \mathbb{E}Z| > t) &\le 3\exp\left(-\frac{t}{KB}\log\left(1 + \frac{t}{2}\right)\right) \\
&\le 3\exp\left(-\frac{t\log 2}{KB}\min\left(1, \frac{t}{2}\right)\right),
\end{aligned}
$$

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

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

we have

$$
\begin{aligned}
&\mathbb{P}\left(|Z - \mathbb{E}Z| > \alpha\sqrt{\frac{\mu r \log n}{na}}\right) \\
&\le 3\exp\left(-\gamma_0 \min\left(2\alpha\sqrt{\frac{na\log n}{\mu r}}, \alpha^2 \log n\right)\right) \\
&= 3\exp(-\gamma_0 \alpha^2 \log n),
\end{aligned}
$$

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

$$
\mathbb{P}\left(|Z - \mathbb{E}Z| \le \sqrt{\frac{\beta}{\gamma_0}}\sqrt{\frac{\mu r \log n}{na}}\right) \ge 1 - 3n^{-\beta}.
$$

Note that we have proved $\mathbb{E}Z \leq C\sqrt{\mu r \log n / na}$. We have

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

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

References

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.

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

Lin, Z.; Chen, M.; and Ma, Y. 2009. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report UILU-ENG-09-2215*.

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

Rudelson, M. 1999. Random vectors in the isotropic position. *Journal of Functional Analysis* 164(1):60–72.

Talagrand, M. 1996. New concentration inequalities in product spaces. *Inventiones Mathematicae* 126(3):505–563.

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