ROBUST PCA VIA DICTIONARY BASED OUTLIER PURSUIT

Xingguo Li[1], *Jineng Ren*[1], *Sirisha Rambhatla*[1], *Yangyang Xu*[2] *and Jarvis Haupt*[1]

[1]Department of Electrical and Computer Engineering
University of Minnesota Twin Cities
Email: {lixx1661, renxx282, rambh002, jdhaupt}@umn.edu

[2]Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Email: xuy21@rpi.edu

ABSTRACT

In this paper, we examine the problem of locating vector outliers from a large number of inliers, with a particular focus on the case where the outliers are represented in a known basis or dictionary. Using a convex demixing formulation, we provide provable guarantees for exact recovery of the space spanned by the inliers and the supports of the outlier columns, even when the rank of inliers is high and the number of outliers is a constant proportion of total observations. Comprehensive numerical experiments on both synthetic and hyper-spectral imaging real datasets demonstrate the efficiency of our proposed method.

Index Terms— robust PCA, outlier identification , hyper-spectral imaging

1. INTRODUCTION

Suppose we observe a data matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, which we assume admits a decomposition of the form:

$$\mathbf{M} = \mathbf{L} + \mathbf{DC} + \mathbf{N}, \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{n_1 \times d}$ is a known dictionary, $\mathbf{L} \in \mathbb{R}^{n_1 \times n_2}$ is unknown, with $\mathrm{rank}(\mathbf{L}) = r$, $\mathbf{C} \in \mathbb{R}^{d \times n_2}$ is an unknown but column-wise sparse matrix and $\mathbf{N} \in \mathbb{R}^{n_1 \times n_2}$ is an unknown matrix modeling error. We denote the column supports of $\mathbf{C}$ (the set of non-zero columns) as $\mathrm{csupp}(\mathbf{C}) = \mathcal{I}_{\mathbf{C}}$ and assume $|\mathcal{I}_{\mathbf{C}}| = k$. In this model, we refer to the non-zero columns of $\mathbf{DC}$ as *outliers*; they are assumed to not be in the column space of $\mathbf{L}$, denoted by $\mathcal{U} = \mathrm{span}\{\mathrm{col}(\mathbf{L})\}$. Similarly, we call the rest of the columns of $\mathbf{L}$ as *inliers*, indexed by $\mathcal{I}_{\mathbf{L}} = [n_2] \backslash \mathcal{I}_{\mathbf{C}}$ with $|\mathcal{I}_{\mathbf{L}}| = n_2 - k = n_{\mathbf{L}}$ and $[n_2] = \{1, \ldots, n_2\}$. Given the data matrix $\mathbf{M}$ and the dictionary $\mathbf{D}$, our specific goal is to recover the column space $\mathcal{U}$ and the indices of outliers $\mathcal{I}_{\mathbf{C}}$.

Model (1) can be viewed as a generalization of principal component analysis (PCA) [1], where the goal is to estimate a low dimensional embedding of given data, and its robust variants, where the data is contaminated by sparse outliers [2–5]. The investigation of outlier identification is motivated by a number of contemporary "big data" applications, where the outliers themselves may be of interest, such as identifying malicious responses in collaborative filtering applications [6], finding anomalous patterns in network traffic [7] or estimating visually salient regions of images [8–10]. More recently, there has been increasing interest in outlier identification with known bases, motivated by real world applications, e.g., functional magnetic resonance imaging [11], video processing [12], network tracking [13,14], and hyper-spectral (HS) imaging [15,16]. However, no rigorous analyses of identifiability of the model (1) have been provided, motivating our investigation here.

Our Approach: For any pair $(\mathbf{L}_0, \mathbf{C}_0)$, we say $(\mathbf{L}_0, \mathbf{C}_0)$ is in the *oracle model* $\{\mathbf{M}, \mathcal{U}, \mathcal{I}_{\mathbf{C}}\}$, i.e., $(\mathbf{L}_0, \mathbf{C}_0) \in \{\mathbf{M}, \mathcal{U}, \mathcal{I}_{\mathbf{C}}\}$, if $\mathcal{P}_{\mathcal{U}}(\mathbf{L}_0) = \mathbf{L}_0$, $\mathcal{P}_{\mathcal{C}}(\mathbf{DC}_0) = \mathbf{DC}_0$, and $\mathbf{L}_0 + \mathbf{DC}_0 = \mathbf{L} + \mathbf{DC}$ hold simultaneously, where $\mathcal{P}_{\mathcal{U}}$ and $\mathcal{P}_{\mathcal{C}}$ are projections onto the column space $\mathcal{U}$ of $\mathbf{L}$ and column support $\mathcal{I}_{\mathbf{C}}$ of $\mathbf{C}$, respectively. Given the data matrix $\mathbf{M}$ and the dictionary $\mathbf{D}$, we aim to recover $\{\mathcal{U}, \mathcal{I}_{\mathbf{C}}\}$ from noisy observations via the following optimization procedure, which we call *Dictionary based Outlier Pursuit* (DOP),

$$\min_{\mathbf{L},\mathbf{C}} \|\mathbf{L}\|_* + \lambda \|\mathbf{C}\|_{1,2} \ \text{ s.t. } \|\mathbf{M} - \mathbf{L} - \mathbf{DC}\|_{\mathrm{F}} \le \varepsilon_{\mathbf{N}}, \tag{2}$$

where $\|\mathbf{L}\|_*$ is the nuclear norm of $\mathbf{L}$, $\|\mathbf{C}\|_{1,2} = \sum_j \|\mathbf{C}_{:,j}\|_2$, $\mathbf{C}_{:,j}$ is the j-th column of $\mathbf{C}$, and $\lambda \ge 0$ is a regularization parameter. Note that we cannot guarantee recovery of the true parameter pair $(\mathbf{L}, \mathbf{C})$ even when $\mathbf{N} = \mathbf{0}$, since there exists ambiguity that the outlier columns (non-zero columns of $\mathbf{DC}$) may contain "energy" of some inliers. Specifically, $\mathcal{P}_{\mathcal{U}^\perp}(\mathbf{DC}) \neq \mathbf{0}$ in general, where $\mathcal{U}^\perp$ is the orthogonal complement of $\mathcal{U}$ in $\mathbb{R}^{n_1}$ and $\mathcal{P}_{\mathcal{U}^\perp}$ is the projection onto $\mathcal{U}^\perp$. When $\mathcal{P}_{\mathcal{U}^\perp}(\mathbf{DC}) \neq \mathbf{0}$ we can always find $(\mathbf{L}_1, \mathbf{C}_1)$, $(\mathbf{L}_2, \mathbf{C}_2) \in \{\mathbf{M}, \mathcal{U}, \mathcal{I}_{\mathbf{C}}\}$ with $(\mathbf{L}_1, \mathbf{C}_1) \neq (\mathbf{L}_2, \mathbf{C}_2)$.

The main contribution of this paper is that we provide sufficient conditions for the convex optimization (2) to enable the recovery of the column space $\mathcal{U}$ of $\mathbf{L}$, and the identities $\mathcal{I}_{\mathbf{C}}$ of the outliers, even when $\mathrm{rank}(\mathbf{L})$ and k are large. Exact recovery can be guaranteed when $\mathbf{N} = \mathbf{0}$.

Background: A closely related model is studied in [3], which estimates $(\mathcal{U}, \mathcal{I}_{\mathbf{C}})$ in the case $\mathbf{D} = \mathbf{I}$, using a convex formulation termed *Outlier Pursuit* (OP). Simply multiplying the (pseudo) inverse $\mathbf{D}^\dagger$ of $\mathbf{D}$ on both sides of (1) and apply OP do not work here in general. For example, when the subspace spanned by $\mathbf{D}$ do not contain $\mathcal{U}$, such an operation results in the column space of $\mathbf{D}^\dagger\mathbf{L} \subset \mathcal{U}$, so we may not recover

$\mathcal{U}$. We may also not recover $\mathcal{I}_{\mathbf{C}}$ since nonzero columns of $\mathbf{C}$ may be in $\mathcal{U}$. In addition, the prior knowledge on $\mathbf{D}$ enables enhanced performance of recovery, especially when $\mathrm{rank}(\mathbf{L})$ is high. Another related model is studied in [14, 16], which estimates $(\mathcal{U}, \mathcal{I}_{\mathbf{C}})$ of (1) with $\mathbf{C}$ being an entry-wise sparse matrix. However, as indicated by our experiments, when the outlier does have a column-wise structure, our estimator obtained from (2) is more robust than the counterpart of entry-wise sparse $\mathbf{C}$ for large k and r.

Notation. For a low rank matrix $\mathbf{L}$, we denote $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ as the *compact singular value decomposition (SVD)* of $\mathbf{L}$, where columns of $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times r}$ are left and right singular vectors of $\mathbf{L}$ respectively, i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_{r\times r}$, and $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ is a diagonal matrix with $\Sigma_{ii} = \sigma_i$ being the i-th singular value of $\mathbf{L}$, for $i \in [r]$. Given a matrix $\mathbf{X}$, the projection operations are defined as $\mathcal{P}_{\mathcal{U}}(\mathbf{X}) = \mathbf{P}_{\mathbf{U}}\mathbf{X}$ and $\mathcal{P}_{\mathcal{V}}(\mathbf{X}) = \mathbf{X}\mathbf{P}_{\mathbf{V}}$, where $\mathbf{P}_{\mathbf{U}} = \mathbf{U}\mathbf{U}^\top$ and $\mathbf{P}_{\mathbf{V}} = \mathbf{V}\mathbf{V}^\top$ are column and row projection matrices respectively, $\mathcal{P}_{\mathcal{L}}(\mathbf{X}) = (\mathcal{P}_{\mathcal{U}} + \mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\mathcal{U}}\mathcal{P}_{\mathcal{V}})(\mathbf{X}) = \mathbf{P}_{\mathbf{U}}\mathbf{X} + \mathbf{X}\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{U}}\mathbf{X}\mathbf{P}_{\mathbf{V}}$, and $\mathcal{P}_{\mathcal{C}}(\mathbf{X})$ is obtained by keeping the i-th column of $\mathbf{X}$ unchanged for $i \in \mathcal{I}_{\mathbf{C}}$, otherwise setting the i-th column of $\mathbf{X}$ to be zero for $i \notin \mathcal{I}_{\mathbf{C}}$. The complement of the operations are defined correspondingly, i.e., $\mathcal{P}_{\mathcal{U}^\perp}(\mathbf{X}) = (\mathbf{I}-\mathbf{P}_{\mathbf{U}})\mathbf{X}$, $\mathcal{P}_{\mathcal{V}^\perp}(\mathbf{X}) = \mathbf{X}(\mathbf{I}-\mathbf{P}_{\mathbf{V}})$, $\mathcal{P}_{\mathcal{L}^\perp}(\mathbf{X}) = (\mathbf{I}-\mathbf{P}_{\mathbf{U}})\mathbf{X}(\mathbf{I}-\mathbf{P}_{\mathbf{V}})$, and $\mathcal{P}_{\mathcal{C}^\perp}(\mathbf{X})$ is obtained by keeping the i-th column of $\mathbf{X}$ unchanged for $i \notin \mathcal{I}_{\mathbf{C}}$, otherwise setting the i-th column of $\mathbf{X}$ to be zero for $i \in \mathcal{I}_{\mathbf{C}}$.

2. PRELIMINARIES

We define several subspaces, similar to those in [14, 16]:

(s1) $\mathcal{L} = \{\mathbf{X} \in \mathbb{R}^{n_1\times n_2} : \mathbf{X} = \mathbf{U}\mathbf{W}_1^\top + \mathbf{W}_2\mathbf{V}^\top,\ \mathbf{W}_1 \in \mathbb{R}^{n_2\times r},\ \mathbf{W}_2 \in \mathbb{R}^{n_1\times r}\}$, the span of all matrices with the same column or row space of $\mathbf{L}$,

(s2) $\mathcal{C} = \{\mathbf{H} \in \mathbb{R}^{d\times n_2} \text{ for any } d \in \mathbb{N} : \mathrm{csupp}(\mathbf{H}) \subseteq \mathcal{I}_{\mathbf{C}}\}$, the matrices with column support contained in the column support of $\mathbf{C}$ and

(s3) $\mathcal{D} = \{\mathbf{X} \in \mathbb{R}^{n_1\times n_2} : \mathbf{X} = \mathbf{D}\mathbf{H},\ \mathbf{H} \in \mathcal{C}\}$, all matrices with the column space as a column subspace of $\mathbf{D}$.

We denote $(\mathcal{L}, \mathcal{C}, \mathcal{D})$ as the subspaces (s1) $\sim$ (s3).

In the following, we introduce some conditions on the local identifiability used throughout our analysis. We first define the *subspace incoherence property* between two subspaces $\mathcal{L}$ and $\mathcal{D}$ to quantify the degree of their overlap. This is formalized as follows.

Definition 2.1 (Subspace Incoherence Property)**.** *Two subspaces $\mathcal{L}$ and $\mathcal{D}$ are said to satisfy the **subspace incoherence property** with parameter $\mu(\mathcal{L}, \mathcal{D})$ when*

$$\max_{\mathbf{X}\in\mathcal{D}\setminus\{\mathbf{0}\}} \frac{\|\mathcal{P}_{\mathcal{L}}(\mathbf{X})\|_{\mathrm{F}}}{\|\mathbf{X}\|_{\mathrm{F}}} \le \mu(\mathcal{L}, \mathcal{D}). \tag{3}$$

Note that $\mu(\mathcal{L}, \mathcal{D}) \in [0, 1]$, where the upper bound is achieved when $\mathcal{P}_{\mathcal{U}^\perp}(\mathbf{D}\mathbf{C}_{:,j}) = 0$ for some $j \in \mathcal{I}_{\mathbf{C}}$, and the lower bound is achieved when $\mathcal{L}$ and $\mathcal{D}$ are orthogonal. In our problem, we are interested in $\mu(\mathcal{L}, \mathcal{D}) < 1$, which indicates that $\mathcal{P}_{\mathcal{U}^\perp}(\mathbf{D}\mathbf{C}_{:,j}) > 0$ for all $j \in \mathcal{I}_{\mathbf{C}}$. This condition plays an important role in guaranteeing the uniqueness of $\{\mathcal{U}, \mathcal{I}_{\mathbf{C}}\}$ given $\mathbf{M}$ and $\mathbf{D}$.

Another important property is a criterion for the dictionary matrix to preserve the Euclidean norm of any fixed vector in a certain space, which is formalized as follows.

Definition 2.2 (Restricted Frame Property)**.** *An $n_1 \times d$ matrix $\mathbf{D}$ is said to satisfy the **restricted frame property** on $\mathbf{x} \in \mathcal{R}_{\mathbf{C}}$ if for any fixed $\mathbf{x} \in \mathcal{R}_{\mathbf{C}}$,*

$$\alpha_\ell\|\mathbf{x}\|_2^2 \le \|\mathbf{D}\mathbf{x}\|_2^2 \le \alpha_u\|\mathbf{x}\|_2^2, \tag{4}$$

where α_u and α_ℓ are upper and lower bounds respectively with $\alpha_u \ge \alpha_\ell > 0$.

The restricted frame property (RFP) is a fairly generic property that is satisfied by many deterministic and random matrices (e.g., with zero-mean Gaussian or subgaussian entries). We do not restrict $\mathbf{D}$ to be overcomplete or to have orthogonal rows as in [14], which allows for a much broader choices of the dictionary. In fact, $\mathcal{R}_{\mathbf{C}}$ can be simply $\mathbb{R}^d$ when $n_1 \ge d$, thus the frame property is easy to meet when $\mathbf{D}$ is an undercomplete dictionary and better recovery performance can be guaranteed, as shown in experiments. In the case $n_1 < d$, it is equivalent to the popular restricted isometry property (RIP) [17, 18] on sparse input, where given $\varepsilon \in (0, 1)$, we have that for sparse vectors $\mathbf{x}$ with $||\mathbf{x}||_0 \le k$ for some $k \le n_1$,

$$(1-\varepsilon)\|\mathbf{x}\|_2^2 \le \|\mathbf{D}\mathbf{x}\|_2^2 \le (1+\varepsilon)\|\mathbf{x}\|_2^2.$$

Generally, RFP holds for any vector in a subspace that has small enough principal angles with the row space $\mathcal{R}(\mathbf{D})$ of $\mathbf{D}$, for which there exists a constant $\varepsilon \in (0, 1]$ such that

$$\min\left\{\arccos\left(\frac{|\langle\mathbf{u}, \mathbf{v}\rangle|}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}\right) \,\middle|\, \mathbf{u} \in \mathcal{R}_{\mathbf{C}}, \mathbf{v} \in \mathcal{R}(\mathbf{D})\right\} \ge \varepsilon.$$

Note that $\alpha_\ell > 0$ in (4) indicates that for any $\mathbf{H} \in \mathcal{C}\backslash\mathbf{0}_{d\times n_2}$, $\mathbf{D}\mathbf{H} \neq \mathbf{0}_{n_1\times n_2}$. This is another important result for the optimality condition we will address later.

We also define several constants for convenience:

$$\beta_{\mathbf{V}} = \|\mathbf{V}\mathbf{V}^\top\|_{\infty,2},\ \beta_{\mathbf{U},\mathbf{V}} = \|\mathbf{D}^\top\mathbf{U}\mathbf{V}^\top\|_{\infty,2},$$

where $\|\mathbf{A}\|_{\infty,2} = \max_i \|\mathbf{A}_{:,i}\|_2$. Small values of $\beta_{\mathbf{U},\mathbf{V}}$ indicates each column of $\mathbf{L}$ is spanned by "sufficiently many" columns of $\mathbf{D}$. In addition, $\beta_{\mathbf{V}}$ is related to the notion of *column incoherence property* (also called leverage score). Specifically, let $\mathbf{L} \in \mathbb{R}^{n_1\times n_2}$ be a rank r matrix with $n_{\mathbf{L}} \le n_2$ non-zero columns. Given the compact SVD $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, $\mathbf{L}$ is said to satisfy the **column incoherence property** with parameter $\mu_{\mathbf{V}}$ if $\|\mathbf{V}^\top\|_{\infty,2}^2 \le \mu_{\mathbf{V}}\frac{r}{n_{\mathbf{L}}}$. Such quantities have been identified as important for the identifiability of low rank components in previous works [2, 3, 19]. Note that $\mu_{\mathbf{V}} \in [1, \frac{n_{\mathbf{L}}}{r}]$, and a small $\mu_{\mathbf{V}}$ indicates that the vectors comprising columns of $\mathbf{L}$ are "spread out" among the basis vectors spanning the column space of $\mathbf{L}$, or equivalently, $\mathbf{V}$ does not contain sparse rows. Consequently, we have $\beta_{\mathbf{V}}^2 \le \mu_{\mathbf{V}}\frac{r}{n_{\mathbf{L}}}$.

3. THEORETICAL GUARANTEES

We provide sufficient conditions for DOP (2) to guarantee accurate recovery of the column space $\mathcal{U}$ of $\mathbf{L}$ and the identities $\mathcal{I}_{\mathbf{C}}$ of non-zero columns of $\mathbf{C}$. We will first provide the theory and then the corresponding optimality conditions.

3.1. Recovery for DOP

The guarantees for the estimation of $\{\mathcal{U}, \mathcal{I}_{\mathbf{C}}\}$ via DOP are provided as the following theorem.

Theorem 3.1. *Suppose* $\mathbf{M} = \mathbf{L} + \mathbf{DC} + \mathbf{N}$ *with* $(\mathbf{L}, \mathbf{C})$ *belonging to the oracle model* $\{\mathbf{M}, \mathcal{U}, \mathcal{I}_{\mathbf{C}}\}$, $\|\mathbf{N}\|_{\mathrm{F}} \leq \varepsilon_{\mathbf{N}}$, $\mathrm{rank}(\mathbf{L}) = r$, *and* $|\mathcal{I}_{\mathbf{C}}| = k$ *with* k *satisfying* $k \leq \frac{1}{4\beta_{\mathbf{V}}^2}$. *Suppose subspaces* $\mathcal{L}$ *and* $\mathcal{D}$ *satisfy* (3) *with parameter* $\mu(\mathcal{L}, \mathcal{D}) \in [0, 1)$, $\mathbf{D}$ *satisfies* (4) *on* $\mathcal{R}_{\mathbf{C}}$ *with* $\alpha_u \geq \alpha_\ell > 0$, *and* $\mathbf{C}_{:,j} \in \mathcal{R}_{\mathbf{C}}$ *for all* $j \in [n_2]$. *If* λ, r *and* k *satisfy*

$$\frac{\beta_{\mathbf{U},\mathbf{V}} + \sqrt{rk\alpha_u}\mu(\mathcal{L}, \mathcal{D})b_2}{\frac{1}{2} - kb_2} \leq \lambda \leq \frac{\frac{b_1}{2} - \sqrt{r\alpha_u}\mu(\mathcal{L}, \mathcal{D})}{\sqrt{k}}, \quad (5)$$

where b_1 *and* b_2 *are defined as*

$$b_1 = \sqrt{\alpha_\ell}(1 - \mu(\mathcal{L}, \mathcal{D})), \text{ and } b_2 = \max_{\|\mathbf{u}\|=1} \frac{\|(\mathbf{I} - \mathbf{P}_{\mathbf{U}})\mathbf{D}\mathbf{u}\|^2}{\|\mathbf{D}\mathbf{u}\|^2} \frac{\alpha_u \beta_{\mathbf{V}}^2}{\alpha_\ell (1 - \mu(\mathcal{L}, \mathcal{D}))^2},$$

then there exists $(\widetilde{\mathbf{L}}, \widetilde{\mathbf{C}}) \in \{\mathbf{M}, \mathcal{U}, \mathcal{I}_{\mathbf{C}}\}$ *such that the optimal solution* $(\widehat{\mathbf{L}}, \widehat{\mathbf{C}})$ *of NDOP in* (2) *satisfies*

$$\begin{aligned} \|\widehat{\mathbf{L}} - \widetilde{\mathbf{L}}\|_{\mathrm{F}} &\leq \big(8\sqrt{r} + 9\tfrac{\sqrt{r\alpha_u}}{\lambda}\big)\varepsilon_{\mathbf{N}}, \\ \|\widehat{\mathbf{C}} - \widetilde{\mathbf{C}}\|_{\mathrm{F}} &\leq 9\sqrt{r}\big(1 + \tfrac{\sqrt{\alpha_u}}{\lambda}\big)\varepsilon_{\mathbf{N}}. \end{aligned} \quad (6)$$

We interpret the condition (5) for λ, r and k to have better insights of the results. Denote $a \lesssim b$ if $a \leq c \cdot b$ for some constant c and $a \approx b$ if $a \lesssim b$ and $b \lesssim a$ hold simultaneously. Suppose $1 \lesssim \alpha_\ell \leq \alpha_u \lesssim 1$, which can be easily met by a tight frame when $n_1 > d$, or a RIP type condition when $n_1 < d$. Note that $\beta_{\mathbf{V}} \approx \frac{\mu_{\mathbf{V}} r}{n_{\mathbf{L}}}$. Then if $\mu(\mathcal{L}, \mathcal{D}) \lesssim \frac{1}{r}$ and $\beta_{\mathbf{U},\mathbf{V}} \lesssim \frac{1}{r}$ (satisfied when $\mathbf{DC}$ and $\mathbf{L}$ has small coherence), we have $k = \mathcal{O}(\frac{n_{\mathbf{L}}}{r \cdot \mu_{\mathbf{V}}})$ and $\frac{1}{k} \lesssim \lambda \lesssim \frac{1}{\sqrt{k}}$. This is of the same order with the upper bound of k in OP [3], but our experiments in Section 4.1 show that DOP outperforms OP even when the rank of $\mathbf{L}$ is high. Note that when the noise $\mathbf{N} = \mathbf{0}$, (6) implies exact recovery.

3.2. Proof Sketch

The Lagrangian of the problem (2) can be written as

$$\mathcal{F}(\mathbf{L}, \mathbf{C}, \mathbf{U}) = \|\mathbf{L}\|_* + \lambda\|\mathbf{C}\|_{1,2} + \langle \mathbf{A}, \mathbf{M} - \mathbf{L} - \mathbf{DC} \rangle, \quad (7)$$

where $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ is a dual variable. The subdifferentials of (7) with respect to $(\mathbf{L}, \mathbf{C})$ are

$$\begin{aligned} \partial_{\mathbf{L}}\mathcal{F}(\mathbf{L}, \mathbf{C}, \mathbf{U}) &= \{\mathbf{U}\mathbf{V}^\top + \mathbf{W} - \mathbf{A} : \|\mathbf{W}\|_2 \leq 1, \mathcal{P}_{\mathcal{L}}(\mathbf{W}) = \mathbf{0}\}, \\ \partial_{\mathbf{C}}\mathcal{F}(\mathbf{L}, \mathbf{C}, \mathbf{U}) &= \{\lambda\mathbf{H} + \lambda\mathbf{Z} - \mathbf{D}^\top\mathbf{A} : \mathcal{P}_{\mathcal{C}}(\mathbf{H}) = \mathbf{H}, \\ &\qquad \mathbf{H}_{:,j} = \tfrac{\mathbf{C}_{:,j}}{\|\mathbf{C}_{:,j}\|_2}, \mathcal{P}_{\mathcal{C}}(\mathbf{Z}) = \mathbf{0}, \|\mathbf{Z}\|_{\infty,2} \leq 1\}. \end{aligned}$$

We claim that a pair $(\mathbf{L}, \mathbf{C})$ is an optimal point of (2) if and only if the following hold by the optimality conditions:

$$\mathbf{0}_{n_1 \times n_2} \in \partial_{\mathbf{L}}\mathcal{F}(\mathbf{L}, \mathbf{C}, \mathbf{U}), \ \mathbf{0}_{d \times n_2} \in \partial_{\mathbf{C}}\mathcal{F}(\mathbf{L}, \mathbf{C}, \mathbf{U}).$$

We then construct the dual certificate as follows. Let the compact SVD of $\mathbf{L}$ be $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, $\widetilde{\mathbf{D}} = (\mathbf{I} - \mathbf{P}_{\mathbf{V}}) \otimes \mathbf{D}^\top (\mathbf{I} - \mathbf{P}_{\mathbf{U}})$, $\widetilde{\mathbf{D}}_{\mathcal{C}}$ be the row submatrix of $\widetilde{\mathbf{D}}$ corresponding to the non-zero rows of $\mathrm{vec}(\mathbf{C})$, $\mathbf{Y} = \lambda\widetilde{\mathbf{C}} - \mathcal{P}_{\mathcal{C}}(\mathbf{D}^\top\mathbf{U}\mathbf{V}^\top)$, and $\mathbf{Y}_{\mathcal{C}}$ be column submatrix of $\mathbf{Y}$ indexed by $\mathcal{I}_{\mathcal{C}}$. If λ, r, and k satisfy (5), then we construct the dual certificate $\widetilde{\mathbf{Q}}$ as

$$\widetilde{\mathbf{Q}} = \mathbf{U}\mathbf{V}^\top + (\mathbf{I} - \mathbf{P}_{\mathbf{U}})\,\widetilde{\mathbf{X}}\,(\mathbf{I} - \mathbf{P}_{\mathbf{V}}), \quad (8)$$

where $\widetilde{\mathbf{X}}$ is given by $\mathrm{vec}(\widetilde{\mathbf{X}}) = \widetilde{\mathbf{D}}_{\mathcal{C}}^\top(\widetilde{\mathbf{D}}_{\mathcal{C}}\widetilde{\mathbf{D}}_{\mathcal{C}}^\top)^{-1}\mathrm{vec}(\mathbf{Y}_{\mathcal{C}})$. The following lemma states the optimality conditions for the optimal solution pair $(\mathbf{L}, \mathbf{C})$, which proves Theorem 3.1.

Lemma 3.1. *Suppose all conditions in Theorem 3.1 hold. Then the construction of dual certificate* $\widetilde{\mathbf{Q}}$ *in* (8) *satisfies*

*(**q̃1**)* $\mathcal{P}_{\mathcal{L}}(\widetilde{\mathbf{Q}}) = \mathbf{U}\mathbf{V}^\top$, *(**q̃2**)* $\|\mathcal{P}_{\mathcal{L}^\perp}(\widetilde{\mathbf{Q}})\|_2 < \frac{1}{2}$,

*(**q̃3**)* $\mathcal{P}_{\mathcal{C}}(\mathbf{D}^\top\widetilde{\mathbf{Q}}) = \lambda\widetilde{\mathbf{C}}$, *where* $\widetilde{\mathbf{C}}_{:,j} = \frac{\mathbf{C}_{:,j}}{\|\mathbf{C}_{:,j}\|_2}$ *for all* $j \in \mathcal{I}_{\mathbf{C}}$; $\mathbf{0}$ *otherwise,* *(**q4**)* $\|\mathcal{P}_{\mathcal{C}^\perp}(\mathbf{D}^\top\widetilde{\mathbf{Q}})\|_{\infty,2} < \frac{\lambda}{2}$.

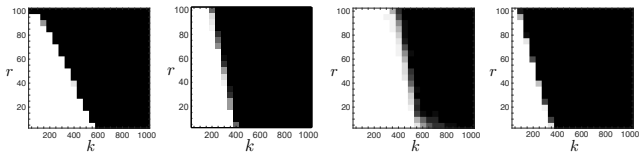
Moreover, if the noise satisfies $\|\mathbf{N}\|_{\mathrm{F}} \leq \varepsilon_{\mathbf{N}}$, *then there exists* $(\widetilde{\mathbf{L}}, \widetilde{\mathbf{C}}) \in \{\mathbf{M}, \mathcal{U}, \mathcal{I}_{\mathbf{C}}\}$ *such that the optimal solution* $(\widehat{\mathbf{L}}, \widehat{\mathbf{C}})$ *of DOP satisfies* (6).

4. NUMERICAL EVALUATION

We provide numerical experiments to study the properties of DOP. We accomplish this by first studying the phase transition in terms of rank and sparsity in Section 4.1, followed by exploration of an application of the proposed approach for target detection in hyper-spectral images in Section 4.2. The competing methods are OP proposed by [3] and a naive procedure multiplying the pseudo inverse of $\mathbf{D}$ on both sides of (1) then applying OP, called Inv+OP when the dictinary is thin. Note that DOP and Inv+OP can be considered as a type of supervised learning method [20]. However, different from most supervised learning methods that need to train implicit model parameters, such as the support vector machines (SVMs) [21], we can detect outliers directly from a given data, i.e., the basis for outliers or a dictionary formed from outliers. Any dictionary learning method or clustering method can be used to form the dictionary. We also test on the popular matched filter [22], which performs uniformly worse than DOP, thus we omit its result here.

4.1. Phase Transition

For ease of exploration, we only discuss the noiseless case, i.e., $\mathbf{N} = \mathbf{0}$. We demonstrate the phase transition with respect to the rank r and the number of outliers k, comparing DOP with OP by setting $\mathbf{M} = \mathbf{L} + \mathbf{C}$ and Inv-OP. Specifically, for DOP, we set $n_1 = 100$, $n_2 = 1000$, $d = 50$ or 150, and



(a) OP (b) Inv+OP (c) DOP (d) DOP

Fig. 1. Phase transitions for (a) OP, (b) Inv+OP, and DOP with (c) $d = 50$; (d) $d = 150$.

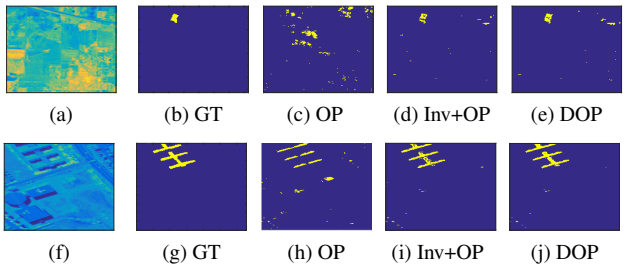
choose $r \in \{5, 10, \ldots, 100\}$ and $k \in \{50, 100, \ldots, 1000\}$ with $\lambda = 0.5$ for $d = 50$ and $\lambda = 1.5$ for $d = 150$. For each pair of r and k, we generate $\mathbf{L} = [\mathbf{U}\mathbf{V}^\top \mathbf{0}_{n_1 \times k}] \in \mathbb{R}^{n_1 \times n_2}$, $\mathbf{C} = [\mathbf{0}_{d \times (n_2 - k)} \mathbf{W}] \in \mathbb{R}^{d \times n_2}$, where $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$, $\mathbf{V} \in \mathbb{R}^{(n_2 - k) \times r}$ and $\mathbf{W} \in \mathbb{R}^{d \times k}$ has i.i.d. $\mathcal{N}(0, 1)$ entries. We also generate $\mathbf{D} \in \mathbb{R}^{n_1 \times d}$ with i.i.d. $\mathcal{N}(0, 1)$ entries, set $\mathbf{M} = \mathbf{L} + \mathbf{D}\mathbf{C}$, and normalized it by column. For OP, we generate $\mathbf{L} \in \mathbb{R}^{n_1 \times n_2}$ and $\mathbf{C} \in \mathbb{R}^{n_1 \times n_2}$ in the same way except that $\mathbf{C} = \mathbf{D}\mathbf{W}$ with $d = 50$ such that columns of $\mathbf{C}$ spans a 50-dimensional subspace of $\mathbb{R}^{100}$.

We demonstrate the average performace over 50 random trials for each of the methods considered in Figure 1. Here, we deem an experiment to be a success (shown in white) if it identifies outlier and inlier locations correctly. We observe that for DOP, even when $\mathbf{L}$ has full row rank, we can recover $\{\mathcal{U}, \mathcal{I}_\mathbf{C}\}$ exactly for a wide range of k. This coincides with both our theory and intuition that the prior knowledge on $\mathbf{D}$ allows for the recovery of wider range of r and k when the problem is properly defined. On the other hand, with OP, the recovery of $\{\mathcal{U}, \mathcal{I}_\mathbf{C}\}$ fails when the rank r is high, even for very small k. This matches with the bound that $k = \mathcal{O}(\frac{n_2}{r \cdot \mu_\mathbf{V}})$. For Inv+OP, it can also recover $\mathcal{I}_\mathbf{C}$ for a range of k when $\mathbf{L}$ has full row rank in this setting. However, when $\mathcal{D}$ does not contain $\mathcal{U}$, i.e., $\|\mathcal{P}_\mathcal{D}(\mathbf{L})\|_\mathrm{F} < \|\mathbf{L}\|_\mathrm{F}$, Inv+OP will lose information of $\mathcal{U}$. In the extreme case, $\mathcal{P}_\mathcal{D}(\mathbf{L}) = \mathbf{0}$ if $\mathcal{D} \perp \mathcal{U}$.

4.2. Target Detection in Hyper-Spectral Imaging Data

We investigate the applicability of our approach for a target detection application in hyper-spectral (HS) imaging. A HS sensor records the response of a scene to different regions of the electromagnetic spectrum. Therefore, the resulting HS image, $\mathcal{Y} \in \mathbb{R}^{s \times m \times w}$ can be viewed as a data cube i.e., a tensor, where each length w voxel corresponds to the spectral response of associated pixel. The aim here is to identify targets in an HS image given the spectral signatures of the targets of interest. Target detection in hyper-spectral images was the topic of one of our previous works [23], where we consider the case of entry-wise sparsity. Here, we analyze the performance of DOP for this application. For our experiments, we consider 2 datasets: (i) Indian Pines collected by AVIRIS sensor [24] with $s = m = 145$ and $w = 200$; and (ii) Pavia University collected by ROSIS sensor [1] with $s = m = 131$ and $w = 201$.

[1]Data is available at http://www.ehu.eus/ccwintco/



(a) (b) GT (c) OP (d) Inv+OP (e) DOP

(f) (g) GT (h) OP (i) Inv+OP (j) DOP

Fig. 2. Demonstration of (a) a slice of Indian Pines HS data array (with $w = 50$) and (f) a slice of Pavia University HS data array (with $w = 100$). We also show (b, g) the ground truth, (c, h) detection results of OP, (d, i) Inv + OP, and (e, j) DOP for Indian Pines and Pavia University.

Table 1. Comparison of the ROC metrics for different methods.

Approach	$d = 4$			$d = 15$		
	TPR	FPR	AUC	TPR	FPR	AUC
DOP	0.989	0.012	0.998	0.989	0.017	0.998
Inv + OP	0.926	0.033	0.980	0.903	0.005	0.946
OP	0.097	0.024	0.095	0.097	0.024	0.095

The data matrix $\mathbf{M} \in \mathbb{R}^{w \times sm}$ is formed by unfolding the tensor data $\mathcal{Y}$ along the third dimension, where each column of $\mathbf{M}$ is the voxel of $\mathcal{Y}$. For our experiments, we form the dictionary $\mathbf{D}$ in two ways – by randomly choosing some voxels from the class of interest, and by learning a dictionary on the class of interest [25] (we skip the description here), which performs better in our experiments.

We demonstrate the results of different approaches, namely DOP, OP, and Inv+OP in Figure 2. The light color here corresponds to the detected targets. For all approaches, we provide the detection result with optimal "visual" detection results compared with ground truth. The corresponding detailed results are shown in Table 1. Here, we report the performance in terms of the ROC metrics, i.e., true positive rate (TPR), false positive rate (FPR), and area under curve (AUC) for each approach for two types of dictionaries – when dictionary is learned (d=4), and when the dictionary is formed from the data voxels directly (d=15). The results show that the proposed method performs better than Inv+OP and OP.

5. DISCUSSION

Further improvement in terms of the sampling and computational efficiency can be achieved via adaptive sensing and sketching [26, 27]. For example, a two-step procedure [28] can be applied to reduce both numbers of the rows and columns in the optimization phase for further speedup. We will leave this investigation to a future effort.

Acknowledgment. The authors acknowledge support from the DARPA Young Faculty Award, Grant N66001-14-1-4047.

6. REFERENCES

[1] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[2] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[3] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," *IEEE Trans. Inform. Theory*, vol. 58, no. 5, pp. 3047–3064, 2012.

[4] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi, "Matrix completion with column manipulation: Near-optimal sample-robustness-rank tradeoffs," *IEEE Trans. on Inform. Theory*, vol. 62, no. 1, pp. 503–526, 2016.

[5] J. Ren, X. Li, and J. Haupt, "Robust PCA via tensor outlier pursuit," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 1744–1749.

[6] B. Mehta and W. Nejdl, "Attack resistant collaborative filtering," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2008, pp. 75–82.

[7] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM Computer Communication Review*. ACM, 2004, vol. 34, pp. 219–230.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency.," in *Advances in Neural Information Processing Systems*, 2006, pp. 545–552.

[10] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *Proc. CVPR*, 2007.

[11] C. Qiu and N. Vaswani, "Recursive sparse recovery in large but correlated noise," in *49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2011, pp. 752–759.

[12] A. Waters, A. Sankaranarayanan, and R. Baraniuk, "Sparcs: Recovering low-rank and sparse matrices from compressive measurements," in *Advances in Neural Information Processing Systems*, 2011, pp. 1089–1097.

[13] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*. USENIX Association, 2005, pp. 30–30.

[14] M. Mardani, G. Mateos, and G. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *IEEE Trans. on Inform. Theory*, vol. 59, no. 8, pp. 5186–5205, 2013.

[15] Chein-I Chang, *Hyperspectral imaging: techniques for spectral detection and classification*, vol. 1, Springer Science & Business Media, 2003.

[16] S. Rambhatla, X. Li, and J. Haupt, "A dictionary based generalization of robust PCA," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2016, pp. 1315–1319.

[17] A. Gilbert, J. Park, and M. Wakin, "Sketched SVD: Recovering spectral features from compressive measurements," *arXiv preprint:1211.0361*, 2012.

[18] D. Achlioptas, "Database-friendly random projections," in *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. ACM, 2001, pp. 274–281.

[19] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[20] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, MIT press, 2012.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] L. Scharf, *Statistical Signal Processing*, vol. 98, Addison-Wesley Reading, MA, 1991.

[23] S. Rambhatla, X. Li, and J. Haupt, "Target-based hyperspectral demixing via generalized robust PCA," in *Fifty First Asilomar Conference on Signals, Systems and Computers*. IEEE, 2017.

[24] M. Baumgardner, L. Biehl, and D. Landgrebe, "220 band AVIRIS hyperspectral image data set: June 12, 1992 indian pine test site 3," Sep 2015.

[25] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. Bach, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems*, 2009, pp. 1033–1040.

[26] E. Arias-Castro, E. Candès, and M. Davenport, "On the fundamental limits of adaptive sensing," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 472–481, 2013.

[27] D. Woodruff, "Sketching as a tool for numerical linear algebra," *Foundations and Trends® in Theoretical Computer Science*, vol. 10, no. 1–2, pp. 1–157, 2014.

[28] X. Li and J. Haupt, "Identifying outliers in large matrices via randomized adaptive compressive sampling," *Trans. Signal Processing*, vol. 63, no. 7, pp. 1792–1807, 2015.