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 ({lixx1661, renxx282, rambh002, jdhaupt}@umn.edu)
[2]Department of Mathematical Sciences, Rensselaer Polytechnic Institute (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 real datasets demonstrate the efficiency of our proposed method.

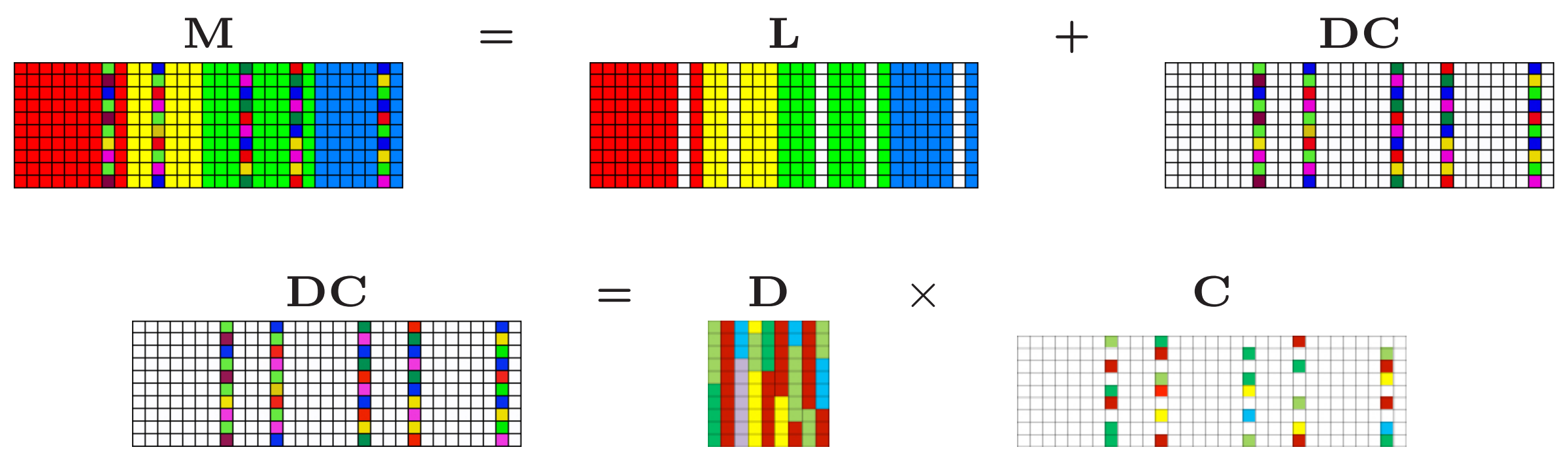
MOTIVATION

Data Model

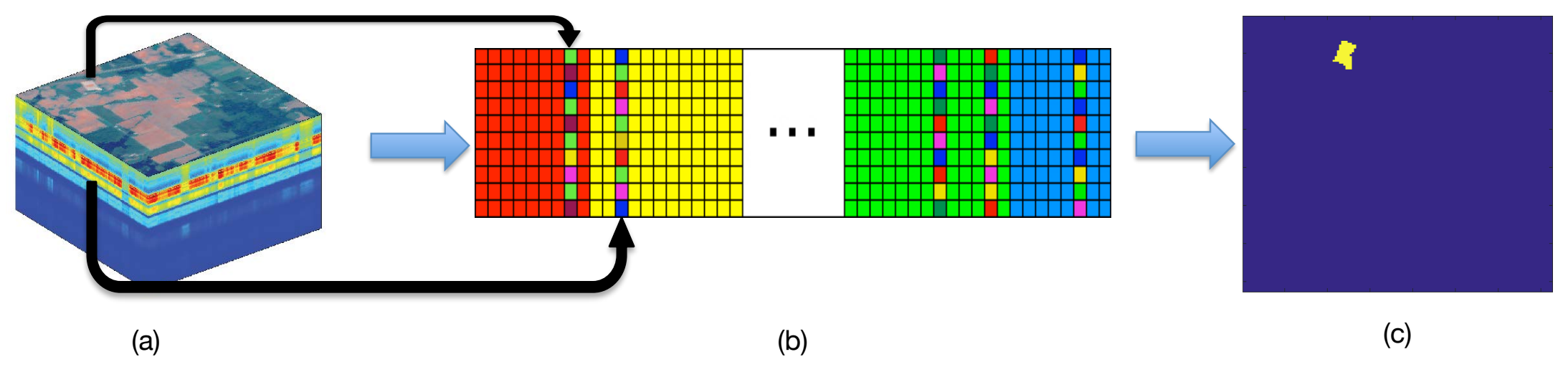
Our particular focus here is on identifying anomalous regions in images. In some applications, information of the anomalous part is known. More specifically, suppose we observe a data matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, which we assume admits a decomposition of the form:

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

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 rank$(\mathbf{L}) = r$, $\mathbf{C} \in \mathbb{R}^{d \times n_2}$ is an unknown but column-wise sparse matrix. We refer to $\mathbf{DC}$ as the *anomalous* part of the data, and our aim is to detect this.



There are indeed some cases where the dictionary/basis of the saliency is known in real-world applications. For example, in hyperspectral imaging data the dictionary can be constructed from the object's class by sampling. The figure below depicts a general example of salient object detection, using hyperspectral imaging data collected by ROSIS sensor from [1].



Given 200 hyperspectral images of size 145×145, which can be regarded as a tensor $\mathcal{Y}$ of size $145 \times 145 \times 200$ as in (a).

(1) Extract voxels of $\mathcal{Y}$, which are column vectors of size 200×1.

(2) Combine column vectors to be a matrix (b) of size 200×145^2, which we call $\mathbf{M}$ matrix

(3) Detect dictionary-based outliers in $\mathbf{M}$, and construct an outlier map (c)

Related Works

Model (1) can be viewed as a generalization of the principal component analysis (PCA) [4], where the goal is to estimate a low dimensional embedding of given data, and its robust variants, where the data matrix is contaminated by sparse outliers [2, 6]. However, existing saliency identification methods (e.g., [3] and many others) only consider the case when there is no outlier information available. A closely related model is studied in [6], which detects the saliency in the case $\mathbf{D} = \mathbf{I}$, using a convex formulation termed *Outlier Pursuit* (OP). When the subspace spanned by $\mathbf{D}$ contains the subspace spanned by $\mathbf{L}$, we can simply multiply the (pseudo) inverse $\mathbf{D}^\dagger$ of $\mathbf{D}$ on both sides of (1) and apply OP. However, in general scenario, such an operation results in the loss of information on $\mathbf{L}$. In addition, the prior knowledge on $\mathbf{D}$ enables enhanced performance of recovery, especially when rank$(\mathbf{L})$ is high.

Idea

Question: How to take advantage of the prior information of the known dictionary?

Answer: In the classical OP procedure, incorporating the decomposition of the outlier $(\mathbf{DC}) \longrightarrow$ DOP (Dictionary-based Outlier Pursuit)

OUR APPROACH

Given the data matrix $\mathbf{M}$ and the dictionary $\mathbf{D}$, we consider to recover the inlier space $\mathcal{U}$ and the support of the ourlier columns $\mathcal{I}_\mathbf{C}$ from a noisy observation via the following optimization procedure, which we call *Dictionary based Outlier Pursuit* (DOP),

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

where $\|\mathbf{L}\|_*$ as 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 $\mathbf{C}_{:,j}$, and $\lambda \geq 0$ is a regularization parameter.

ALGORITHM FOR DOP

We adopt an accelerated proximal gradient descent method to solve the outlier pursuit problem (2), along the lines of the algorithm proposed in [5].

Algorithm 1. APG (Accelerated Proximal Gradient descent) solver for (2)

```
Input: $\mathbf{M}$, $\mathbf{R}$, $\lambda$, $v$, $\nu_0$, $\bar{\nu}$, and $L_f = \lambda_{max}([\mathbf{I}_L\mathbf{R}]'[\mathbf{I}_L\mathbf{R}])$
Initialize: $\mathbf{L}[0] = \mathbf{L}[-1] = \mathbf{0}_{L\times T}$, $\mathbf{C}[0] = \mathbf{C}[-1] = \mathbf{0}_{F\times T}$,
$t[0] = t[-1] = 1$, and set $k = 0$.
    while not converged do
        $\mathbf{T_L}[k] = \mathbf{L}[k] + \frac{t[k-1]-1}{t[k]}(\mathbf{L}[k] - \mathbf{L}[k-1])$
        $\mathbf{T_C}[k] = \mathbf{C}[k] + \frac{t[k-1]-1}{t[k]}(\mathbf{C}[k] - \mathbf{C}[k-1])$
        $\mathbf{G_L}[k] = \mathbf{T_L}[k] + \frac{1}{L_f}(\mathbf{M} - \mathbf{T_L}[k] - \mathbf{RT_C}[k])$
        $\mathbf{G_C}[k] = \mathbf{T_C}[k] + \frac{1}{L_f}\mathbf{R}'(\mathbf{M} - \mathbf{T_L}[k] - \mathbf{RT_C}[k])$
        $\mathbf{U\Sigma V}' = \mathrm{svd}(\mathbf{G_L}[k])$,  $\mathbf{L}[k+1] = \mathbf{U}\mathcal{S}_{\nu[k]/L_f}(\mathbf{\Sigma})\mathbf{V}'$
        $\mathbf{C}[k+1] = \mathcal{S}_{\nu[k]/L_f}(\mathbf{G_C}[k])$
        $t[k+1] = \left[1 + \sqrt{4t^2[k]+1}\right]/2$
        $\nu[k+1] = \max\{v\nu[k], \bar{\nu}\}$
        $k \leftarrow k+1$
    end while
return $\mathbf{L}[k]$, $\mathbf{C}[k]$
```

PRELIMINARIES

Let the compact SVD of $\mathbf{L}$ be $\mathbf{U}\Sigma\mathbf{V}^T$, where rank$(\mathbf{L}) = r$, $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$, $\mathbf{V} \in \mathbb{R}^{n_2 \times r}$.

Given a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$, define:

- $\mathcal{P}_\mathcal{U}(\mathbf{X}) = \mathbf{P_U X}$ and $\mathcal{P}_\mathcal{V}(\mathbf{X}) = \mathbf{XP_V}$, where $\mathbf{P_U} = \mathbf{UU}^\top$ and $\mathbf{P_V} = \mathbf{VV}^\top$
- $\mathcal{P}_\mathcal{L}(\mathbf{X}) = (\mathcal{P}_\mathcal{U} + \mathcal{P}_\mathcal{V} - \mathcal{P}_\mathcal{U}\mathcal{P}_\mathcal{V})(\mathbf{X}) = \mathbf{P_U X} + \mathbf{XP_V} - \mathbf{P_U XP_V}$
- $\mathcal{P}_\mathcal{C}(\mathbf{X})$ 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 $i \notin \mathcal{I}_\mathbf{C}$
- $\mathcal{R}_\mathbf{C}$ is the column space of the dictionary $\mathbf{D}$
- $\beta_\mathbf{V} = \|\mathbf{VV}^\top\|_{\infty,2}$, $\beta_{\mathbf{U},\mathbf{V}} = \|\mathbf{D}^\top\mathbf{UV}^\top\|_{\infty,2}$

DEFINITIONS

We introduce two definitions:

(a1) Two subspaces $\mathcal{L}$ and $\mathcal{D}$ are said to satisfy the **subspace incoherence property** with parameter $\mu(\mathcal{L},\mathcal{D})$ if

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

(a2) 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 $\mathbf{x} \in \mathcal{R}_\mathbf{C}$,

$$\alpha_l\|\mathbf{x}\|_2^2 \leq \|\mathbf{Dx}\|_2^2 \leq \alpha_u\|\mathbf{x}\|_2^2, \tag{4}$$

where α_u and α_l are upper and lower bounds respectively with $\alpha_u \geq \alpha_l > 0$.

MAIN THEOREM

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

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

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 in* (2) *satisfies*

$$\|\widehat{\mathbf{L}} - \widetilde{\mathbf{L}}\|_\mathrm{F} \leq \left(8\sqrt{r} + 9\frac{\sqrt{r\alpha_u}}{\lambda}\right)\varepsilon_\mathbf{N},$$
$$\|\widehat{\mathbf{C}} - \widetilde{\mathbf{C}}\|_\mathrm{F} \leq 9\sqrt{r}\left(1 + \frac{\sqrt{\alpha_u}}{\lambda}\right)\varepsilon_\mathbf{N}.$$

Complexity Analysis

Suppose that

(1) $1 \lesssim \alpha_l \leq \alpha_u \lesssim 1$, which can be easily met by a tight frame when $n_1 > d$, or a RIP type condition when $n_1 < d$,

(2) $\mu(\mathcal{L},\mathcal{D}) \lesssim \frac{1}{r}$ and $\beta_{\mathbf{U},\mathbf{V}} \lesssim \frac{1}{r}$ (satisfied when $\mathbf{DC}$ and $\mathbf{L}$ has small coherence),

then the condition above becomes

$$k = \mathcal{O}\left(\frac{n_\mathbf{L}}{r\cdot\mu_\mathbf{V}}\right) \textbf{ and } \frac{1}{k} \lesssim \lambda \lesssim \frac{1}{\sqrt{k}}.$$

EXPERIMENTS – SYNTHETIC DATA

We examine the performance of our approach first on synthetically generated data, generated as follows:

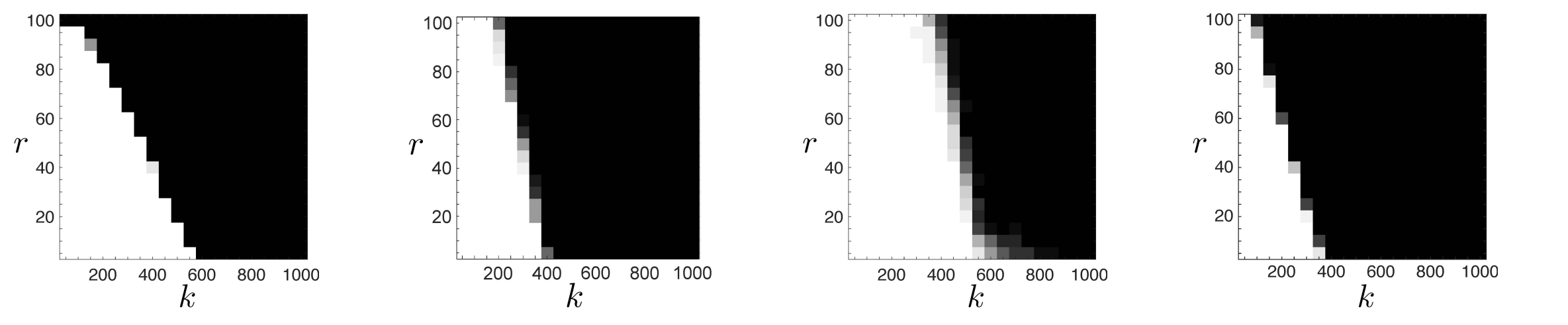
- For DOP and Inv+OP, we set $n_1 = 100$, $n_2 = 1000$, $d = 50$ or 150, and choose $r \in \{5, 10, \dots, 100\}$ and $k \in \{50, 100, \dots, 1000\}$ with $\lambda = 0.5$ for $d = 50$ and $\lambda = 1.5$ for $d = 500$.
- For each pair of r and k, we generate $\mathbf{L} = [\mathbf{UV}^\top\mathbf{0}_{n_1\times k}] \in \mathbb{R}^{n_1\times n_2}$, $\mathbf{C} = [\mathbf{0}_{n_1\times(n_2-k)}\mathbf{W}] \in \mathbb{R}^{d\times n_2}$, where $\mathbf{U} \in \mathbb{R}^{n_1\times r}$, $\mathbf{V} \in \mathbb{R}^{n_2\times r}$ and $\mathbf{W} \in \mathbb{R}^{d\times k}$ has i.i.d. $\mathcal{N}(0,1)$ entries. $\mathbf{D} \in \mathbb{R}^{n_1\times d}$ is generated with i.i.d. $\mathcal{N}(0,1)$ entries and we normalize columns of $\mathbf{M} = \mathbf{L} + \mathbf{DC}$ to be unit vectors.
- 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 $\mathbf{C} = \mathbf{DW}$ with $d = 50$ such that columns of $\mathbf{C}$ spans a 50-dimensional subspace of $\mathbb{R}^{100}$.
- The phase transition results with different r and k for OP, Inv+OP, and DOP when $n_1 = 100$, $n_2 = 1000$, $d = 50$ are shown in **Figure 1** (a), (b), (c) respectively; The phase transition result for DOP when $n_1 = 100$, $n_2 = 1000$, $d = 150$ is shown in **Figure 1** (d). We perform 50 random trials to record the times of successful recovery (from 0 to 50) of $\{\mathcal{U},\mathcal{I}_\mathbf{C}\}$. We also choose different λ's for each case to find the best performing setting. Here white regions correspond to all successes and black regions correspond to all failures.

Competing Algorithms:

(1) **Dictionary based Outlier Pursuit (DOP):** the proposed dictionary based outlier detection approach.

(2) **Outlier Pursuit (OP):** the classical outlier pursuit approach without dictionary information proposed in [6].

(3) **Inverse + Outlier Pursuit (Inv+OP):** multiplying the pseudo inverse of $\mathbf{D}$ on both sides of (1) then applying OP



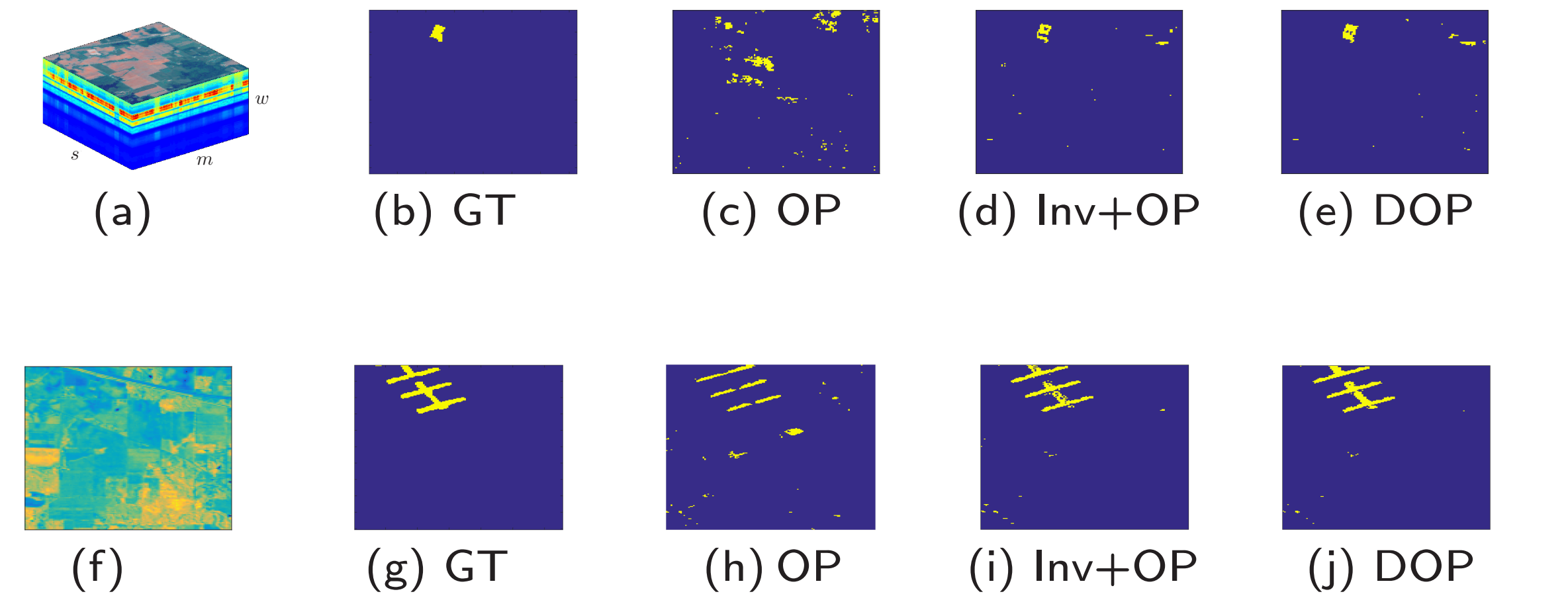
(a) OP (b) Inv+OP (c) DOP (d) DOP

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

EXPERIMENTS – REAL DATA

We also applied our approach to real hyperspectral image data:

- The raw data is a 3-way tensor $\mathcal{Y} \in \mathbb{R}^{s\times m\times w}$, where w is the number of frequency bands, and s and m are the 2-D image dimensions.
- For Indian Pines collected by AVIRIS sensor [1]: $s = m = 145$ and $w = 200$; for Pavia University collected by ROSIS sensor (http://www.ehu.eus/ccwintco/): $s = m = 131$ and $w = 201$.
- 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 $\mathcal{Y}$. This corresponds to, e.g., $n_1 = 200$ and $n_2 = 145^2 = 21,025$ for Indian Pines. The recover results are shown in Figure 2.
- The ROC metrics, i.e., true positive rate (TPR), false positive rate (FPR), and area under curve (AUC), for all approaches are also presented in **Table 1** when we choose different sizes of dictionaries (column numbers d=4, d=15).



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

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

Figure 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 $w = 100$). (b, g) are the ground truth, (c, h) are detection results of OP, (d, i) Inv + OP, and (e, j) DOP for Indian Pines and Pavia University.

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

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

DISCUSSION

Figure 1: For DOP, even when $\mathbf{L}$ has full row rank, we can recover $\mathcal{I}_C$ exactly for a wide range of k (coincides with our theory). For OP, the recovery fails when rank r is high, even for very small k. Inv+OP can recover $\mathcal{I}_\mathbf{C}$ for a smaller range of k when $\mathbf{L}$ has full row rank.

Figure 2: DOP and Inv+OP outperform OP on both real datasets. Moreover, the real detection result of DOP is better than Inv+OP's.

Table 1: DOP achieves better ROC metrics which means that the detection result of DOP is more accurate than the results of Inv+OP and OP.

ACKNOWLEDGMENT

This work was supported by the DARPA Young Faculty Award, Grant N66001-14-1-4047.

SELECTED REFERENCES

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

[2] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011.

[3] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pages 545–552, 2007.

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

[5] 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*, 59(8):5186–5205, 2013.

[6] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *IEEE Trans. Inform. Theory*, 58(5):3047–3064, 2012.