A DICTIONARY BASED GENERALIZATION OF ROBUST PCA

Sirisha Rambhatla, Xingguo Li and Jarvis Haupt

Department of Electrical and Computer Engineering,
University of Minnesota-Twin Cities, Minneapolis, MN-55455

{rambh002, lixx1661, jdhaupt}@umn.edu

ABSTRACT

We analyze the decomposition of a data matrix, assumed to be a superposition of a low-rank component and a component which is sparse in a known dictionary, using a convex demixing method. We provide a unified analysis, encompassing both undercomplete and overcomplete dictionary cases, and show that the constituent components can be successfully recovered under some relatively mild assumptions up to a certain *global* sparsity level. Further, we corroborate our theoretical results by presenting empirical evaluations in terms of phase transitions in rank and sparsity for various dictionary sizes.

Index Terms— Low-rank, dictionary sparse, Robust PCA.

1. INTRODUCTION

Exploiting the inherent structure of data for the recovery of relevant information is at the heart of data analysis. In this paper, we analyze a scenario where a data matrix $\mathbf{Y} \in \mathbb{R}^{n \times m}$ arises as a result of a superposition of a rank-r component $\mathbf{X} \in \mathbb{R}^{n \times m}$, and a dictionary sparse component expressed here as $\mathbf{R}\mathbf{A}$. Here, $\mathbf{R} \in \mathbb{R}^{n \times d}$ is an *a priori* known dictionary with normalized columns, and $\mathbf{A} \in \mathbb{R}^{d \times m}$ is the unknown *sparse* coefficient matrix with at most s total non-zeros. Specifically, we will study the following model,

$$\mathbf{Y} = \mathbf{X} + \mathbf{R}\mathbf{A}, \qquad (1)$$

and identify the conditions under which the components $\mathbf{X}$ and $\mathbf{A}$, can be successfully recovered, given $\mathbf{Y}$ and $\mathbf{R}$.

A wide range of problems can be expressed in the form described above. Perhaps the most celebrated of these is principal component analysis (PCA) [1], which can be viewed as a special case of eq.(1), with the matrix $\mathbf{A}$ set to zero. In the absence of the component $\mathbf{X}$, the problem reduces to that of sparse recovery [2–4]; See [5] and references therein for an overview of related works. The popular framework of Robust PCA tackles a case when the dictionary $\mathbf{R}$ is an identity matrix [6,7]; variants include [8–11]. In addition, other variants of Robust PCA, such as Outlier Pursuit [12], where $\mathbf{R} = \mathbf{I}$ and the sparse component is column sparse, and randomized adaptive sensing approaches [13–17], have also been explored.

Our work is most closely related to [18], which explores the application of the model shown in eq.(1) to detect traffic anomalies, and focuses on a case where the dictionary $\mathbf{R}$ is

overcomplete, i.e., *fat*. The model described therein, is applicable to a case where the rows of $\mathbf{R}$ are orthogonal, e.g., $\mathbf{R}\mathbf{R}' = \mathbf{I}$, and the coefficient matrix $\mathbf{A}$, has at most k nonzero elements per row and column. In this paper, we analyze a more general case, where we relax some of the aforementioned assumptions for the *fat* case, and develop an analogous analysis for the *thin* case. Specifically, this paper makes the following contributions towards guaranteeing the recovery of $\mathbf{X}$ and $\mathbf{A}$ in eq.(1). First, we analyze the *thin* case, where we assume $\mathbf{R}$ to be a *frame* [19] with a *global* sparsity of at most s; See [20] for a brief overview of frames. Next, for the *fat* case, we extend the analysis presented in [18], and assume that the dictionary $\mathbf{R}$ satisfies the restricted isometry property (RIP) of order k, with a global sparsity of at most s, and a column sparsity of at most k. Consequently, we eliminate the sparsity constraint on the rows of the coefficient matrix $\mathbf{A}$ and the orthogonality constraint on the rows of the dictionary $\mathbf{R}$ required by [18].

The model shown in eq.(1) is propitious in a number of applications. For example, it can be used for target identification in hyperspectral imaging, and in topic modeling applications to identify documents with certain properties. Further, in source separation tasks, a variant of this model was used in singing voice separation in [21, 22]. Further, we can also envision source separation tasks where $\mathbf{X}$ is not low-rank, but can in turn be modeled as being sparse in a known [23] or unknown [24] dictionary.

The rest of the paper is organized as follows. We formulate the problem, introduce the notation and describe various considerations on the structure of the component matrices in section 2. In section 3, we present our main result and a proof sketch, followed by numerical simulations in section 4. Finally, we conclude in section 5 with some insights on future work.

2. PROBLEM FORMULATION

Our aim is to recover the low-rank component $\mathbf{X}$, and the sparse coefficient matrix $\mathbf{A}$, given the dictionary $\mathbf{R}$, and samples $\mathbf{Y}$ generated according to the model described in eq.(1). Utilizing the assumed structure of the components $\mathbf{X}$ and $\mathbf{A}$, we consider the following convex problem for $\lambda \geq 0$,

$$\underset{\mathbf{X},\mathbf{A}}{\text{minimize}} \ \|\mathbf{X}\|_* + \lambda \|\mathbf{A}\|_1 \ \text{s.t.} \ \mathbf{Y} = \mathbf{X} + \mathbf{R}\mathbf{A}. \qquad (2)$$

where, $\|.\|_*$ denotes the nuclear norm, and $\|.\|_1$ refers to the l_1-norm, which serve as convex relaxations of rank and sparsity (i.e. l_0-norm), respectively. Depending upon the number of dictionary elements, d in $\mathbf{R}$, we analyze the problem described

above for two cases – a) when $d \leq n$, the *thin* case, and b) when $d > n$, the *fat* case. For the *thin* case, we assume that the rows of the dictionary $\mathbf{R}$ to comprise a *frame*, i.e. for any vector $\mathbf{v} \in \mathbb{R}^d$, we have

$$\mathbf{F}_L \|\mathbf{v}\|_2^2 \leq \|\mathbf{R}\mathbf{v}\|_2^2 \leq \mathbf{F}_U \|\mathbf{v}\|_2^2, \tag{3}$$

where $\mathbf{F}_L$ and $\mathbf{F}_U$ are the lower and upper *frame bounds*, respectively, with $0 < \mathbf{F}_L \leq \mathbf{F}_U$. Next, for $d > n$, the *fat* case, we assume that $\mathbf{R}$ obeys the restricted isometry property (RIP) of order k, i.e. for any k-sparse vector $\mathbf{v} \in \mathbb{R}^d$, we have

$$(1 - \delta)\|\mathbf{v}\|_2^2 \leq \|\mathbf{R}\mathbf{v}\|_2^2 \leq (1 + \delta)\|\mathbf{v}\|_2^2, \tag{4}$$

where, $\delta \in [0, 1]$ is the restricted isometry constant (RIC).

The aim of this paper is to answer the following question – Given $\mathbf{R}$, under what conditions can we recover $\mathbf{X}$ and $\mathbf{A}$ from the mixture $\mathbf{Y}$? We observe that there are a few ways we can run into trouble right away, namely– a) the dictionary sparse part $\mathbf{RA}$ is low-rank, and b) the low rank part, $\mathbf{X}$, is sparse in the dictionary, $\mathbf{R}$. Indeed, these relationships take the center stage in our analysis. We begin by defining a few relevant subspaces, similar to those used in [18], which will help us to formalize the said relationships.

Let the pair $\{\mathbf{X_0}, \mathbf{A_0}\}$ be the solution to the problem shown in eq.(2). We define $\mathbf{\Phi}$ as the linear space of matrices spanning the row and column spaces of the low-rank component $\mathbf{X}_0$. Specifically, let $\mathbf{U\Sigma V}'$ denote the singular value decomposition of $\mathbf{X}_0$, then the space $\mathbf{\Phi}$ is defined as

$$\mathbf{\Phi} := \{\mathbf{UW}_1 + \mathbf{W}_2\mathbf{V}', \mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{n \times r}\}.$$

Next, let $\mathbf{\Omega}$ be the space spanned by $d \times m$ matrices that have the same support (location of non-zero elements) as $\mathbf{A}_0$, and let $\mathbf{\Omega_R}$ be defined as

$$\mathbf{\Omega_R} := \{\mathbf{Z} = \mathbf{RH}, \mathbf{H} \in \mathbf{\Omega}\}.$$

In addition, we denote the corresponding complements of the spaces described above by appending '$\perp$'. Next, let the orthogonal projection operator(s) onto the space of matrice(s) defined above be $\mathcal{P}_\Phi(.)$, $\mathcal{P}_\Omega(.)$ and $\mathcal{P}_{\Omega_R}(.)$, respectively. Further, we will use $\mathbf{P}_U$ and $\mathbf{P}_V$ to denote the projection matrices corresponding to the column and row spaces of $\mathbf{X}_0$, respectively, i.e., implying the following for any matrix $\mathbf{X}$,

$$\mathcal{P}_\Phi(\mathbf{X}) = \mathbf{P}_U\mathbf{X} + \mathbf{X}\mathbf{P}_V - \mathbf{P}_U\mathbf{X}\mathbf{P}_V$$
$$\mathcal{P}_{\Phi^\perp}(\mathbf{X}) = (\mathbf{I} - \mathbf{P}_U)\mathbf{X}(\mathbf{I} - \mathbf{P}_V).$$

As alluded to previously, there are indeed situations under which we cannot hope to recover the matrices $\mathbf{X}$ and $\mathbf{A}$. To identify these scenarios, we will employ various notions of incoherence. We define the incoherence between the low-rank part, $\mathbf{X}_0$, and the dictionary sparse part, $\mathbf{RA_0}$ as,

$$\mu := \max_{\mathbf{Z} \in \mathbf{\Omega}_R \setminus \{\mathbf{0}_{d \times m}\}} \frac{\|\mathcal{P}_\Phi(\mathbf{Z})\|_F}{\|\mathbf{Z}\|_F},$$

where $\mu \in [0, 1]$ is small when these components are incoherent (good for recovery). The next two measures of incoherence can be interpreted as a way to avoid the cases where for $\mathbf{X}_0 = \mathbf{U\Sigma V}'$, (a) $\mathbf{U}$ resembles the dictionary $\mathbf{R}$, and (b) $\mathbf{V}$ resembles the sparse coefficient matrix $\mathbf{A}_0$. In this case, the low-rank

part essentially mimics the dictionary sparse component. To this end, similar to [18], we define respectively the following to measure these properties,

$$\gamma_{UR} := \max_i \frac{\|\mathbf{P}_U\mathbf{R}\mathbf{e}_i\|^2}{\|\mathbf{R}\mathbf{e}_i\|^2} \text{ and } \gamma_V := \max_i \|\mathbf{P}_V\mathbf{e}_i\|^2,$$

where $\gamma_V \in [r/m, 1]$. Also, we define $\xi := \|\mathbf{R}'\mathbf{UV}'\|_\infty$.

3. MAIN RESULT

In this section, we present the conditions under which solving the problem stated in eq.(2) will successfully recover the true matrices $\mathbf{X}$ and $\mathbf{A}$. As discussed in the previous section, the structure of the dictionary $\mathbf{R}$ plays a crucial role in the analysis of the two paradigms, i.e. the *thin* case and the *fat* case. Consequently, we provide results corresponding to these cases separately. We begin by introducing a few definitions and assumptions applicable to both cases. To simplify the analysis we assume that $d < m$. Specifically, we will assume that $d \leq \frac{m}{\alpha r}$, where $\alpha > 1$ is a constant. In addition, our analysis is applicable to the case when $s < m$.

Definition D.1. We define
$$c := \begin{cases} c_t, \text{ for } d \leq n \\ c_f, \text{ for } d > n \end{cases},$$

where, c_t and c_f are defined as,
$c_t := \frac{\mathbf{F}_U}{2}[(1 + 2\gamma_{UR})(\min(s, d) + s\gamma_V) + 2s\gamma_V] - \frac{\mathbf{F}_L}{2}[\min(s, d) + s\gamma_V]$ and
$c_f := \frac{1+\delta}{2}[(1 + 2\gamma_{UR})(\min(s, k) + s\gamma_V) + 2s\gamma_V] - \frac{1-\delta}{2}(\min(s, k) + s\gamma_V)$.
Further, we define C and $\lambda_{\min}$ as,

$$C := \begin{cases} \frac{c_t}{\mathbf{F}_L(1-\mu)^2 - c_t}, & \text{for } d \leq n \text{ and } \mathbf{F}_L \leq \frac{1}{(1-\mu)^2} \\ \frac{c_f}{(1-\delta)(1-\mu)^2 - c_f}, & \text{for } d > n \end{cases} \text{ and } \lambda_{\min} := \frac{1+C}{1-C} \xi.$$

Definition D.2.
$$\lambda_{\max} := \begin{cases} \frac{1}{\sqrt{s}}\left(\sqrt{\mathbf{F}_L}(1 - \mu) - \sqrt{r\mathbf{F}_U}\mu\right), & \text{if } d \leq n \\ \frac{1}{\sqrt{s}}\left(\sqrt{(1 - \delta)}(1 - \mu) - \sqrt{r(1 + \delta)}\mu\right), & \text{if } d > n \end{cases}$$

Assumption A.1. $\lambda_{\max} \geq \lambda_{\min}$
Assumption A.2. Let $s_{\max} := \frac{(1-\mu)^2}{2}\frac{m}{r}$, then

$$\gamma_{UR} \leq \begin{cases} \frac{(1-\mu)^2 - 2s\gamma_V}{2s(1+\gamma_V)}, & \text{for } s \leq \min(d, s_{\max}) \\ \frac{(1-\mu)^2 - 2s\gamma_V}{2(d+s\gamma_V)}, & \text{for } d < s \leq s_{\max} \end{cases}.$$

Assumption A.3. For $s_{\max}$ as above,

$$\gamma_{UR} \leq \begin{cases} \frac{(1-\mu)^2 - 2s\gamma_V}{2s(1+\gamma_V)}, & \text{for } s \leq \min(k, s_{\max}) \\ \frac{(1-\mu)^2 - 2s\gamma_V}{2(k+s\gamma_V)}, & \text{for } k < s \leq s_{\max} \end{cases}.$$

Theorem 1. *Consider a superposition of a low-rank matrix* $\mathbf{X}_0 \in \mathbb{R}^{n \times m}$ *of rank* r, *and a dictionary sparse component* $\mathbf{RA_0}$, *wherein the sparse coefficient matrix* $\mathbf{A}_0$ *has at most* s *non-zeros, i.e.,* $\|\mathbf{A}_0\|_0 = s$, *and* $\mathbf{Y} = \mathbf{X}_0 + \mathbf{RA_0}$, *with parameters* γ_{UR}, ξ, $\gamma_V \in [r/m, 1]$ *and* $\mu \in [0, 1]$. *Then, solving the formulation shown in eq.(2) will recover matrices* $\mathbf{X}_0$ *and* $\mathbf{A}_0$ *if the following conditions hold for any* $\lambda \in [\lambda_{min}, \lambda_{max}]$, *as defined in D.1 and D.2, respectively.*

● For $d \leq n$, *the dictionary* $\mathbf{R} \in \mathbb{R}^{n \times d}$ *obeys the frame condition with frame bounds* $[\mathbf{F}_L, \mathbf{F}_U]$ *and assumptions A.1, and A.2 hold.*

● For $d > n > C_1 k \log(d)$, *the dictionary* $\mathbf{R} \in \mathbb{R}^{n \times d}$ *obeys the RIP of order* k *with RIC* δ, *a constant* C_1 *and assumptions A.1 and A.3 hold.*

Thm. 1 establishes the sufficient conditions for the existence of λs to guarantee recovery of $\{\mathbf{X_0}, \mathbf{A_0}\}$ for both the *thin* and the *fat* case. For both cases, we see that the conditions are closely related to the various incoherence measures γ_{UR}, γ_V

and μ between the low-rank part, $\mathbf{X}$, the dictionary, $\mathbf{R}$, and the sparse component $\mathbf{A}$. Further, we observe that the theorem imposes an an upper-bound on the global sparsity, i.e., $s \leq s_{\max}$. This is similar to what was reported in [12], and seems a result of the deterministic analysis presented here. Further, the condition shown in assumption A.1, i.e., $\lambda_{\min} \leq \lambda_{\max}$, translates to a relationship between rank r, and the sparsity s, namely,

$$r \leq \left(\sqrt{\frac{\mathbf{F}_L}{\mathbf{F}_U}} \frac{1-\mu}{\mu} - \frac{\xi}{\sqrt{\mathbf{F}_U}\mu} \frac{1+C}{1-C} \sqrt{s} \right)^2, \quad (5)$$

$\forall s \geq 0$ such that, $\sqrt{s} \leq \frac{\mathbf{F}_L(1-\mu)}{\xi} \frac{1-C}{1+C}$, for the *thin* case, and

$$r \leq \left(\sqrt{\frac{1-\delta}{1+\delta}} \frac{1-\mu}{\mu} - \frac{\xi}{\sqrt{(1+\delta)\mu}} \frac{1+C}{1-C} \sqrt{s} \right)^2, \quad (6)$$

$\forall s \geq 0$ such that, $\sqrt{s} \leq \frac{(1-\delta)(1-\mu)}{\xi} \frac{1-C}{1+C}$, for the *fat* case. These relationships are indeed what we observe in empirical evaluations; this will be revisited in the next section. We now present a brief proof sketch of the results presented in this section.

3.1. Proof Sketch

We use dual certificate construction procedure to prove the main result in Thm. 1 [25]. To this end, we start by constructing a dual certificate for the convex problem shown in eq.(2). In our analysis, we use $\|\mathbf{M}\| := \sigma_{\max}(\mathbf{M})$ for the spectral norm, here $\sigma_{\max}(\mathbf{M})$ denotes the maximum singular value of the matrix $\mathbf{M}$, $\|\mathbf{M}\|_\infty := \max_{\{i,j\}} |M_{ij}|$, and $\|\mathbf{M}\|_{\infty,\infty} := \max_i \|\mathbf{e}_i'\mathbf{M}\|_1$. The following lemma shows the conditions the dual certificate needs to satisfy.

Lemma 2 : (from Lemma 2 in [18] and Thm. 3 in [12]): *If there exists a dual certificate* $\mathbf{\Gamma} \in \mathbb{R}^{n \times m}$ *satisfying*

$C1 : \mathcal{P}_\Phi(\mathbf{\Gamma}) = \mathbf{UV}'$ $C2 : \mathcal{P}_\Omega(\mathbf{R}'\mathbf{\Gamma}) = \lambda\text{sign}(\mathbf{A}_0)$
$C3 : \|\mathcal{P}_{\Phi\perp}(\mathbf{\Gamma})\| < 1$ $C4 : \|\mathcal{P}_{\Omega\perp}(\mathbf{R}'\mathbf{\Gamma})\|_\infty < \lambda$

then the pair $\{\mathbf{X}_0, \mathbf{A}_0\}$ *is the unique solution of eq (2).*

We will now proceed with the construction of the dual certificate which satisfies the conditions outlined by conditions C1-4 by Lemma 2. Using the analysis similar to [18] (section V. B.), we construct the dual certificate as

$$\mathbf{\Gamma} = \mathbf{UV}' + (\mathbf{I} - \mathbf{P}_\mathbf{U})\mathbf{X}(\mathbf{I} - \mathbf{P}_\mathbf{V}),$$

for arbitrary $\mathbf{X} \in \mathbb{R}^{n \times m}$. The condition C2 then translates to

$$\mathcal{P}_\Omega(\mathbf{R}'\mathbf{UV}') + \mathcal{P}_\Omega(\mathbf{R}'(\mathbf{I} - \mathbf{P}_\mathbf{U})\mathbf{X}(\mathbf{I} - \mathbf{P}_\mathbf{V})) = \lambda\,\text{sign}(\mathbf{A}_0)$$

Let $\mathbf{Z} := \mathbf{R}'(\mathbf{I} - \mathbf{P}_\mathbf{U})\mathbf{X}(\mathbf{I} - \mathbf{P}_\mathbf{V})$ and $\mathbf{B}_\Omega := \lambda\text{sign}(\mathbf{A}_0) - \mathcal{P}_\Omega(\mathbf{R}'\mathbf{UV}')$, then we can write the equation above as,

$$\mathcal{P}_\Omega(\mathbf{Z}) = \mathbf{B}_\Omega.$$

Note that $\text{vec}(\mathbf{Z}) = [(\mathbf{I} - \mathbf{P}_\mathbf{V}) \otimes \mathbf{R}'(\mathbf{I} - \mathbf{P}_\mathbf{U})]\text{vec}(\mathbf{X})$. Now, let $\tilde{\mathbf{A}} := (\mathbf{I} - \mathbf{P}_\mathbf{V}) \otimes \mathbf{R}'(\mathbf{I} - \mathbf{P}_\mathbf{U})$, and let $\tilde{\mathbf{A}}_\Omega \in \mathbb{R}^{s \times nm}$ denote the rows of $\tilde{\mathbf{A}}$ that correspond to support of $\mathbf{A}_0$, and $\tilde{\mathbf{A}}_{\Omega\perp}$ correspond to the remaining rows of $\tilde{\mathbf{A}}$. Further, let $\mathbf{b}_\Omega$ be a length s vector containing elements of $\mathbf{B}_\Omega$ corresponding to support of $\mathbf{A}_0$. Using these definitions and results, we conclude that

$$\tilde{\mathbf{A}}_\Omega\text{vec}(\mathbf{X}) = \mathbf{b}_\Omega$$

This implies that $\text{vec}(\mathbf{X}) = \tilde{\mathbf{A}}_\Omega'(\tilde{\mathbf{A}}_\Omega\tilde{\mathbf{A}}_\Omega')^{-1}\mathbf{b}_\Omega$. Now, we look at the condition C3, i.e. $\|\mathcal{P}_{\Phi\perp}(\mathbf{\Gamma})\|$, this is where our analysis departs from [18]; we write

$$\|\mathcal{P}_{\Phi\perp}(\mathbf{\Gamma})\| = \|(\mathbf{I} - \mathbf{P}_\mathbf{U})\mathbf{X}(\mathbf{I} - \mathbf{P}_\mathbf{V})\| \leq \|(\mathbf{I} - \mathbf{P}_\mathbf{U})\|\|\mathbf{X}\|\|(\mathbf{I} - \mathbf{P}_\mathbf{V})\|$$
$$\leq \|\mathbf{X}\| \leq \|\mathbf{X}\|_F \leq \|\tilde{\mathbf{A}}_\Omega'(\tilde{\mathbf{A}}_\Omega\tilde{\mathbf{A}}_\Omega')^{-1}\|\|\mathbf{b}_\Omega\|_2,$$

where we have used the fact that $\|(\mathbf{I} - \mathbf{P}_\mathbf{U})\| \leq 1$ and $\|(\mathbf{I} - \mathbf{P}_\mathbf{V})\| \leq 1$. Now, as $\tilde{\mathbf{A}}_\Omega'(\tilde{\mathbf{A}}_\Omega\tilde{\mathbf{A}}_\Omega')^{-1}$ is the pseudo-inverse of $\tilde{\mathbf{A}}_\Omega$, i.e., $\tilde{\mathbf{A}}_\Omega\tilde{\mathbf{A}}_\Omega'(\tilde{\mathbf{A}}_\Omega\tilde{\mathbf{A}}_\Omega')^{-1} = \mathbf{I}$, we have that $\|\tilde{\mathbf{A}}_\Omega'(\tilde{\mathbf{A}}_\Omega\tilde{\mathbf{A}}_\Omega')^{-1}\| = 1/\sigma_{\min}(\tilde{\mathbf{A}}_\Omega)$, where $\sigma_{\min}(\tilde{\mathbf{A}}_\Omega)$ is the smallest singular value of $\tilde{\mathbf{A}}_\Omega$. Therefore, we have

$$\|\mathcal{P}_{\Phi\perp}(\mathbf{\Gamma})\| \leq \frac{\|\mathbf{b}_\Omega\|_2}{\sigma_{\min}(\tilde{\mathbf{A}}_\Omega)}.$$

To obtain an upper bound on $\|\mathcal{P}_{\Phi\perp}(\mathbf{\Gamma})\|$, we will now present the following lemmata.

Lemma 3 : *The lower bound on* $\sigma_{min}(\tilde{\mathbf{A}}_\Omega)$ *is given by*
$$\sigma_{\min}(\tilde{\mathbf{A}}_\Omega) \geq \begin{cases} \sqrt{\mathbf{F}_L}(1-\mu), \text{for } d \leq n \\ \sqrt{(1-\delta)}(1-\mu), \text{for } d > n \end{cases}.$$

Lemma 4 : *Upper bound on* $\|\mathbf{b}_\Omega\|_2$ *is given by*
$$\|\mathbf{b}_\Omega\|_2 \leq \begin{cases} \lambda\sqrt{s} + \sqrt{r\mathbf{F}_U}\mu, \text{for } d \leq n \\ \lambda\sqrt{s} + \sqrt{r(1+\delta)}\mu, \text{for } d > n \end{cases}.$$

Assembling the results of the lemmata to obtain the upper bound on $\|\mathcal{P}_{\Phi\perp}(\mathbf{\Gamma})\|$ and consequently C3, we arrive at the expression for $\lambda_{\max}$ defined in D.2. Now, we move onto finding conditions under which C4 is satisfied by our dual certificate. For this we will bound $\|\mathcal{P}_{\Omega\perp}(\mathbf{R}'\mathbf{\Gamma})\|_\infty$. From eq.(16) in [18] we have,

$$\|\mathcal{P}_{\Omega\perp}(\mathbf{R}'\mathbf{\Gamma})\|_\infty \leq \|\mathbf{Q}\|_{\infty,\infty}\|\mathbf{b}_\Omega\|_\infty + \|\mathcal{P}_{\Omega\perp}(\mathbf{R}'\mathbf{UV}')\|_\infty, \quad (7)$$

where $\mathbf{Q} := \tilde{\mathbf{A}}_{\Omega\perp}\tilde{\mathbf{A}}_\Omega'(\tilde{\mathbf{A}}_\Omega\tilde{\mathbf{A}}_\Omega')^{-1}$. Our aim now will be to bound $\|\mathbf{Q}\|_{\infty,\infty}$ and $\|\mathbf{b}_\Omega\|_\infty$ for our case. For this, we present the following lemmata.

Lemma 5 (from eq.(17) in [18]): *The upper bound on* $\|\mathbf{b}_\Omega\|_\infty$ *is given by* $\lambda + \|\mathcal{P}_\Omega(\mathbf{R}'\mathbf{UV}')\|_\infty$.

Lemma 6: *The upper bound on* $\|\mathbf{Q}\|_{\infty,\infty}$ *is given by* C, *where* C *is as defined in D.1*

Substituting these in eq.(7) and C4, we have

$$\|\mathcal{P}_{\Omega\perp}(\mathbf{R}'\mathbf{\Gamma})\|_\infty \leq C(\lambda + \|\mathcal{P}_\Omega(\mathbf{R}'\mathbf{UV}')\|_\infty) + \|\mathcal{P}_{\Omega\perp}(\mathbf{R}'\mathbf{UV}')\|_\infty.$$

The expression above is further upper bounded by λ due to C4. Here, C and c are as defined in D.1. Hence, we arrive at the following lower bound for λ,

$$\lambda_{\min} := \frac{1+C}{1-C}\,\xi.$$

Gleaning from the expressions for $\lambda_{\max}$ and $\lambda_{\min}$, we observe that the following conditions need to be satisfied for the existence of λs that can recover the desired matrices– a) $\lambda_{\max} \geq \lambda_{\min} > 0$, and b) $0 < C < 1$. These conditions are satisfied by the assumptions A.1-A.3. This completes the proof.

4. SIMULATIONS

Our analysis in the previous section shows that depending upon the size of the dictionary $\mathbf{R}$, if the conditions of Thm. 1 are met, a convex program which solves eq.(2) will recover the components $\mathbf{X}$ and $\mathbf{A}$. In this section, we empirically evaluate the claims of Thm. 1. To this end, we employ the accelerated proximal gradient algorithm outlined in Algorithm 1 of [18] to analyze the phase transition in rank and sparsity for different sizes of the dictionary $\mathbf{R}$. For our analysis, we consider the case where $n = m = 100$. Here, we generate the low-rank
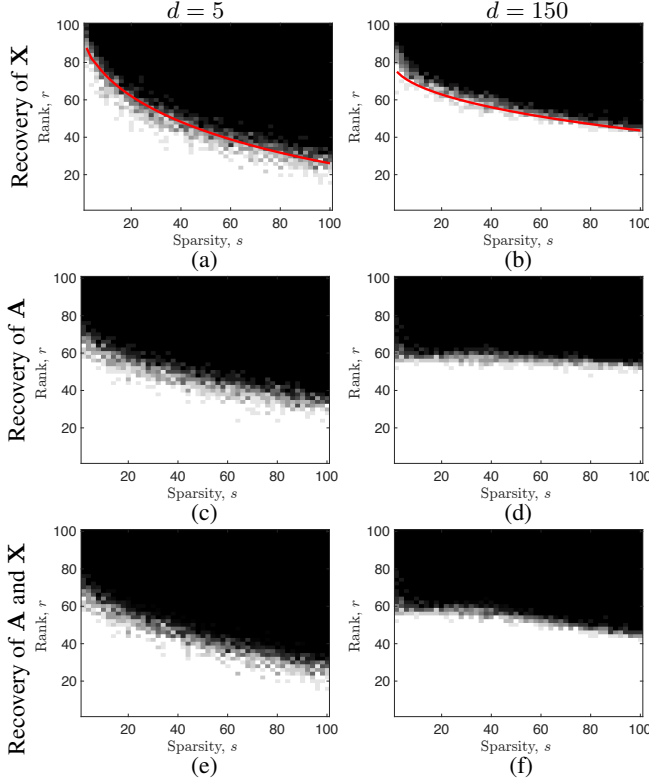
Fig. 1: Recovery for varying rank of $\mathbf{X}$, sparsity of $\mathbf{A}$ and number of dictionary elements in $\mathbf{R}$ as per Thm. 1. Each plot shows average recovery across 10 trials for varying ranks (y-axis) and sparsity (x-axis) up to $s \le m$, white region representing correct recovery, for $n = m = 100$. We decide success if $\|\mathbf{X} - \hat{\mathbf{X}}\|_F / \|\mathbf{X}\|_F \le 0.02$ and $\|\mathbf{A} - \hat{\mathbf{A}}\|_F / \|\mathbf{A}\|_F \le 0.02$, where $\hat{\mathbf{X}}$ and $\hat{\mathbf{A}}$ are the recovered $\mathbf{X}$ and $\mathbf{A}$, respectively. Panels (a)-(b) show the recovery of the low-rank part $\mathbf{X}$ and (c)-(d) show the recovery of the sparse part with varying dictionary sizes $d = 5$ and 150, respectively. Panels (e) -(f) show the region of overlap where both $\mathbf{X}$ and $\mathbf{A}$ are recovered successfully. Also, the predicted trend between rank r and sparsity s as per Thm. 1, eq.(5) and eq.(6) is shown in red in panels (a-b).

part $\mathbf{X}$ by outer product of two random matrices of sizes $n \times r$ and $m \times r$, with entries drawn from the standard normal distribution. In addition, the non-zero entries (s in number), of the sparse component $\mathbf{A}$, are drawn from the Rademacher distribution, also the dictionary $\mathbf{R}$ is drawn from the standard normal distribution, then its columns are normalized. Phase transition in rank and sparsity over 10 trials for dictionaries of sizes $d = 5$ (thin) and $d = 150$ (fat), corresponding to our theoretical results, and for all admissible levels of sparsity are shown in Fig. 1 and Fig. 2, respectively.

Fig. 1 shows the recovery of the low-rank part $\mathbf{X}$ in panels (a-b), while panels (c-d) show the recovery for the sparse component $\mathbf{A}$, for $d = 5$ and 150, respectively. Next, panels (e-f) show the region of overlap between the low-rank recovery and sparse recovery plots, for $d = 5$ and $d = 150$, respectively. This corresponds to the region in which both $\mathbf{X}$ and $\mathbf{A}$ are recovered successfully. Further, the red line in panels (a) and (b) show the trend predicted by our analysis, i.e., eq.(5) and eq.(6), with the parameters hand-tuned for best fit. Indeed, the
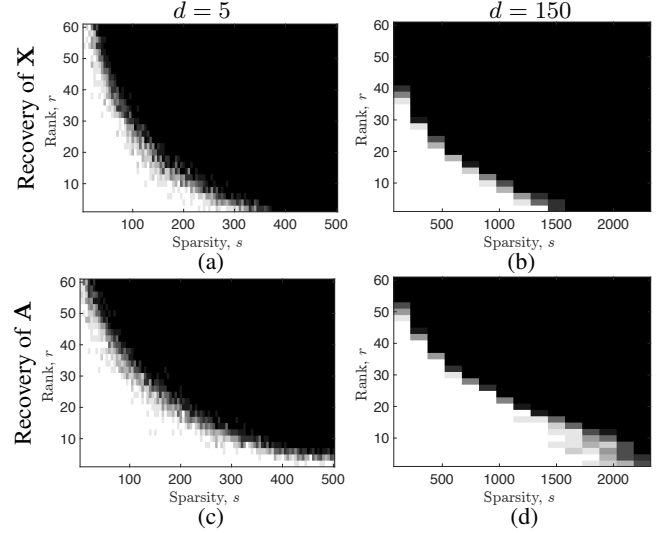
Fig. 2: Recovery for varying rank of $\mathbf{X}$, sparsity of $\mathbf{A}$ and number of dictionary elements in $\mathbf{R}$. Each plot shows average recovery across 10 trials for varying ranks (y-axis) and sparsity (x-axis), white region representing correct recovery, for $n = m = 100$. We decide success if $\|\mathbf{X} - \hat{\mathbf{X}}\|_F / \|\mathbf{X}\|_F \le 0.02$ and $\|\mathbf{A} - \hat{\mathbf{A}}\|_F / \|\mathbf{A}\|_F \le 0.02$, where $\hat{\mathbf{X}}$ and $\hat{\mathbf{A}}$ are the recovered $\mathbf{X}$ and $\mathbf{A}$, respectively. Panels (a)-(b) show the recovery of the low-rank part $\mathbf{X}$ and (c)-(d) show the recovery of the sparse part with varying dictionary sizes $d = 5$ and 150, respectively.

empirical relationship between rank and sparsity has the same trend as predicted by Thm. 1.

Similarly, Fig. 2 shows the recovery of the low-rank part $\mathbf{X}$ in panels (a-b), while panels (c-d) show the recovery for the sparse component $\mathbf{A}$ for $d = 5$ and 150, respectively, for a much wider range of global sparsity s. Indeed, these phase transition plots show that we can successfully recover the components for sparsity levels much greater than those put forth by the theorem. This can be attributed to the deterministic analysis presented here. To this end, we conjecture that a randomized analysis of the problem can potentially improve the results presented here.

5. CONCLUSIONS

We analyze a dictionary based generalization of Robust PCA. Specifically, we extend the theoretical guarantees presented in [18] to a setting wherein the dictionary $\mathbf{R}$ may have arbitrary number of columns, and the coefficient matrix $\mathbf{A}$ has *global* sparsity of s, i.e. $\|\mathbf{A}\|_0 = s \le s_{\max}$. We generalize the results by assuming $\mathbf{R}$ to be a *frame* for the *thin* case, to obey the RIP condition for the *fat* case, and eliminate the orthogonality constraints on the rows of the dictionary $\mathbf{R}$ and the sparsity constraint on the rows of the coefficient matrix $\mathbf{A}$ (as in [18]), rendering the results useful for a potentially wide range of applications. Further, we provide empirical evaluations via phase transition plots in rank and sparsity corresponding to our theoretical results. We draw motivations from the promising phase transition plots, beyond the sparsity level tolerated by our analysis, and propose randomized analysis of the problem to improve the upper-bound on the sparsity as a future work.

6. REFERENCES

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

[2] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM journal on computing*, vol. 24, no. 2, pp. 227–234, 1995.

[3] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[4] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[5] H. Rauhut, "Compressive sensing and structured random matrices," *Theoretical foundations and numerical methods for sparse recovery*, vol. 9, pp. 1–92, 2010.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11, 2011.

[7] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[8] Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma, "Stable principal component pursuit," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1518–1522.

[9] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3419–3430, 2011.

[10] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *Information and Inference*, vol. 2, no. 1, pp. 32–68, 2013.

[11] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix recovery from errors and erasures," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4324–4337, 2013.

[12] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Advances in Neural Information Processing Systems*, 2010, pp. 2496–2504.

[13] 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.

[14] X. Li and J. Haupt, "Locating salient group-structured image features via adaptive compressive sensing," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2015, pp. 393–397.

[15] X. Li and J. Haupt, "Outlier identification via randomized adaptive compressive sampling," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2015, pp. 3302–3306.

[16] M. Rahmani and G. Atia, "Randomized robust subspace recovery for high dimensional data matrices," *arXiv preprint arXiv:1505.05901*, 2015.

[17] X. Li and J. Haupt, "A refined analysis for the sample complexity of adaptive compressive outlier sensing," in *IEEE Statistical Signal Processing Workshop (SSP)*, June 2016, pp. 1–5.

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

[19] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series," *Transactions of the American Mathematical Society*, vol. 72, no. 2, pp. 341–366, 1952.

[20] C. Heil, "What is ... a frame?," *Notices of the American Mathematical Society*, vol. 60, no. 6, June/July 2013.

[21] P. S. Huang, S. D. Chen, P. Smaragdis, and M. J. Hasegawa, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012*, 2012, pp. 57–60.

[22] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low-rank modeling.," in *ISMIR*, 2012, pp. 67–72.

[23] J. L. Starck, Y. Moudden, J. Bobin, M. Elad, and D. L. Donoho, "Morphological component analysis," in *Optics & Photonics 2005*. International Society for Optics and Photonics, 2005, pp. 59140Q–59140Q.

[24] S. Rambhatla and J. Haupt, "Semi-blind source separation via sparse representations and online dictionary learning," in *Signals, Systems and Computers, 2013 Asilomar Conference on*. IEEE, 2013, pp. 1687–1691.

[25] S. Rambhatla, X. Li, and J. Haupt, "A dictionary based generalization of robust PCA with applications," *(In preparation)*, 2016.

[26] G.A. Watson, "Characterization of the subdifferential of some matrix norms," *Linear Algebra and its Applications*, vol. 170, pp. 33 – 45, 1992.