Identifying Outliers in Large Matrices via Randomized Adaptive Compressive Sampling

Xingguo Li and Jarvis Haupt

Abstract—This paper examines the problem of locating outlier columns in a large, otherwise low-rank, matrix. We propose a simple two-step adaptive sensing and inference approach and establish theoretical guarantees for its performance; our results show that accurate outlier identification is achievable using very few linear summaries of the original data matrix – as few as the squared rank of the low-rank component plus the number of outliers, times constant and logarithmic factors. We demonstrate the performance of our approach experimentally in two stylized applications, one motivated by robust collaborative filtering tasks, and the other by saliency map estimation tasks arising in computer vision and automated surveillance.

Index Terms—Adaptive sensing, compressed sensing, robust PCA, sparse inference

I. INTRODUCTION

In this paper we address a matrix outlier identification problem. Suppose $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ is a data matrix that admits a factorization of the form

$$\mathbf{M} = \mathbf{L} + \mathbf{C},$$

where $\mathbf{L}$ is a low-rank matrix, and $\mathbf{C}$ is a matrix of outliers that is nonzero in only a fraction of its columns. We are ultimately interested in identifying the locations of the nonzero columns of $\mathbf{C}$, with a particular focus on settings where “full” data may be very large. The question we address here is, can we accurately (and efficiently) identify the locations of the outliers from a small number of linear measurements of $\mathbf{M}$?

Our investigation is motivated in part by robust collaborative filtering applications, in which the goal may be to identify the locations (or even quantify the number) of corrupted data points or outliers in a large data array. Such tasks may arise in a number of contemporary applications, for example, when identifying malicious responses in survey data or anomalous patterns in network traffic, to name a few. Depending on the nature of the outliers, conventional low-rank approximation approaches based on principal component analysis (PCA) [1]–[3] may be viable options for these tasks, but such approaches become increasingly computationally demanding as the data become very high-dimensional. Here, our aim is to leverage dimensionality reduction ideas along the lines of those utilized in randomized numerical linear algebra, (see, e.g., [4], [5] and the references therein) and compressed sensing (see, e.g., [6]–[8]), in order to reduce the size of the data on which our approach operates. In so doing, we also reduce the computational burden of the inference approach relative to comparable methods that operate on “full data.”

We are also motivated by an image processing task that arises in many computer vision and surveillance applications – that of identifying the “saliency map” [9] of a given image, which (ideally) indicates the regions of the image that tend to attract the attention of a human viewer. Saliency map estimation is a well-studied area, and numerous methods have been proposed for obtaining saliency maps for a given image – see, for example, [10]–[14]. In contrast to these (and other) methods designed to identify saliency map of an image as a “post processing” step, our aim here is to estimate the saliency map directly from compressive samples – i.e., without first performing full image reconstruction as an intermediate step. We address this problem here using a linear subspace-based model of saliency, wherein we interpret an image as a collection of distinct (non-overlapping) patches, so that images may be (equivalently) represented as matrices whose columns are vectorized versions of the patches. Previous efforts have demonstrated that such local patches extracted from natural images may be well approximated as vectors in a union of low-dimensional linear subspaces (see, e.g., [15]). Here, our approach to the saliency map estimation problem is based on an assumption that salient regions in an image may be modeled as outliers from a single common low-dimensional subspace; the efficacy of similar saliency models for visual saliency has been established recently in [16]. Our approach here may find utility in rapid threat detection in security and surveillance applications in high-dimensional imaging tasks where the goal is not to image the entire scene, but rather to merely identify regions in the image space corresponding to anomalous behavior. Successful identification of salient regions could comprise a first step in an active vision task, where subsequent imaging is restricted to the identified regions.

A. Innovations and Our Approach

We propose an approach that employs dimensionality reduction techniques within the context of a two-step adaptive sampling and inference procedure, and our method is based on a few key insights. First, we exploit the fact that the enabling geometry of our problem (to be formalized in the following section) is approximately preserved if we operate not on $\mathbf{M}$ directly, but instead on a “compressed” version $\Phi \mathbf{M}$ of $\mathbf{M}$ that has potentially many fewer rows. Next, we use the fact that we can learn the (ostensibly, low-dimensional) linear subspace spanned by the columns of the low rank component of $\Phi \mathbf{M}$ using a small, randomly selected subset of the columns of $\Phi \mathbf{M}$. Our algorithmic approach for this step utilizes a recently
Algorithm 1 Adaptive Compressive Outlier Sensing (ACOS)

Assume: $M \in \mathbb{R}^{n_1 \times n_2}$

Input: Column sampling Bernoulli parameter $\gamma \in [0,1]$, regularization parameter $\lambda > 0$, Measurement matrices $\Phi \in \mathbb{R}^{m \times n_1}$, $A \in \mathbb{R}^{p \times n_2}$, measurement vector $\phi \in \mathbb{R}^{1 \times m}$

Initialize: Column sampling matrix $S = \mathbb{1}_{sg}$, where $S = \{i : S_i = 1\}$ with $S_i \in \{0,1\}$ i.i.d. Bernoulli($\gamma$)

Step 1
Collect Measurements: $Y(1) = \Phi MS$
Solve: $\hat{L}(1), \hat{C}(1) = \arg\min_{L,C} \|L\|_+ + \lambda\|C\|_{1,2} \quad \text{s.t.} \quad Y(1) = L + C$
Let: $\hat{L}(1)$ be the linear subspace spanned by col’s of $\hat{L}(1)$

Step 2
Compute: $P_{\hat{L}(1)}$, the orthogonal projector onto $\hat{L}(1)$
Set: $P_{\hat{L}(1)} = \mathbb{I} - P_{\hat{L}(1)}$
Collect Measurements: $y(2) = \phi P_{\hat{L}(1)} \Phi MA^T$
Solve: $\hat{c} = \arg\min_{c} \|c\|_1 \quad \text{s.t.} \quad y(2) = cA^T$
Output: $\hat{C} = \{i : \hat{c}_i \neq 0\}$

The proposed method called Outlier Pursuit (OP) [17] that aims to separate a matrix $Y$ into its low-rank and column-sparse components using the convex optimization

$$\arg\min_{L,C} \|L\|_+ + \lambda\|C\|_{1,2} \quad \text{s.t.} \quad Y = L + C$$

where $\|L\|_+$ denotes the nuclear norm of $L$ (the sum of its singular values), $\|C\|_{1,2}$ is the sum of the $\ell_2$ norms of the columns of $C$, and $\lambda > 0$ is a regularization parameter. Finally, we leverage the fact that, contingent on correct identification of the subspace spanned by the low-rank component of $\Phi M$, we may effectivly transform – via linear observations – the overall outlier identification problem into a “conventional” compressed sensing problem, and employ now well-known theoretical results (e.g., [18]) to establish the overall success of our approach. We call our approach Adaptive Compressive Outlier Sensing (ACOS), and summarize it here as Algorithm 1. Our main contributions include a theoretical analysis of this method and experimental evaluation of its performance.

B. Related Work

Our effort here leverages results from Compressive Sensing (CS), where parsimony in the object or signal being acquired, in the form of sparsity, is exploited to devise efficient procedures for acquiring and reconstructing high-dimensional objects [6]–[8], [18]. The sequential and adaptive nature of our proposed approach is inspired by numerous recent works in the burgeoning area of adaptive sensing and adaptive CS (see, for example, [19]–[36] as well as the summary article [37] and the references therein). Our efforts here utilize a generalization of the notion of sparsity, formalized in terms of a low-rank plus outlier matrix model. In this sense, our efforts here are related to earlier work in Robust PCA [38], [39] that seek to identify low-rank matrices in the presence of sparse impulsive outliers, and their extensions to settings where the outliers present as entire columns of an otherwise low-rank matrix [17], [40], [41]. In fact, the computational approach and theoretical analysis of the first step of our approach make direct utilization of the results of [17].

We also note a related work [42], which seeks to decompose matrices exhibiting some simple structure (e.g., low-rank plus sparse, etc.) into their constituent components from compressive observations. Our work differs from that approach in both the measurement model and scope. Namely, our measurements take the form of linear functions of rows and/or columns of the matrix and our overall approach is adaptive in nature, as compared to the non-adaptive “global” compressive measurements acquired in [42], each of which is essentially a linear combination of all of the matrix entries. Further, the goal of [42] was to exactly recover the constituent components, while our aim is only to identify the locations of the outliers. We discuss some further connections with [42] in Section V.

A component of our numerical evaluation here entails assessing the performance of our approach in a stylized image processing task of saliency map estimation. We note that several recent works have utilized techniques from the sparse representation literature in salient region identification, and in compressive imaging scenarios. A seminal effort in this direction was [43], which proposed a model for feature identification via the human visual cortex based on parsimonious (sparse) representations. More recently, [44] applied techniques from dictionary learning [43], [45] and low-rank-plus-sparse matrix decomposition [38], [39] in a procedure to identify salient regions of an image from (uncompressed) measurements. Similar sparse representation techniques for salient feature identification were also examined in [46]. An adaptive compressive imaging procedure driven by a saliency “map” obtained via low-resolution discrete cosine transform (DCT) measurements was demonstrated in [47]. Here, unlike in [44], [46], we consider salient feature identification based on compressive samples, and while our approach is similar in spirit to the problem examined in [47], here we provide theoretical guarantees for the performance of our approach. Finally, we note several recent works [48], [49] that propose methods for identifying salient elements in a data set using compressive samples.

C. Outline

The remainder of the paper is organized as follows. In Section II we formalize our problem, state relevant assumptions, and state our main theoretical result that establishes performance guarantees for the Adaptive Compressive Outlier Sensing (ACOS) approach of Algorithm 1. In Section III we provide a proof of our main result, which entails a straightforward integration of intermediate lemmata that effectively describe conditions under which each step of our approach succeeds. Section IV contains the results of a comprehensive experimental evaluation of our approach on synthetic data, as well as in a stylized image processing application of saliency map estimation. In Section V we provide a brief discussion of the computational complexity of our approach, and discuss a few potential future directions. We relegate proofs and other auxiliary material to the appendix.
D. A Note on Notation

We use bold-face upper-case letters (M, L, C, \Phi, L, C, I etc.) to denote matrices, and use the MATLAB-inspired notation I, S to denote the sub matrix formed by extracting columns of I indexed by i \in S (see, e.g., Algorithm 1). We typically use bold-face lower-case letters (x, v, c, \phi, etc.) to denote vectors, with an exception along the lines of the indexing notation above – i.e., that C_{:,i} denotes the i-th column of C. Note that we employ both “block” and “math” type notation (e.g., L, L), where the latter are used to denote variables in the optimization tasks that arise throughout our exposition. Non-bold letters are used to denote scalar parameters or constants; the usage will be made explicit, or will be clear from context.

The \ell_1 norm of a vector x = [x_1, x_2 \ldots x_n] is \|x\|_1 = \sum_i |x_i| and the \ell_2 norm is \|x\|_2 = (\sum_i |x_i|^2)^{1/2}. We denote the nuclear norm (the sum of singular values) of a matrix L by \|L\|_*, and the 1, 2 norm (the sum of column \ell_2 norms) of a matrix C by \|C\|_{1,2}. We denote the operator norm (the largest singular value) of a matrix L by \|L\|. Superscript asterisks denote complex conjugate transpose.

For positive integers n, we let \lfloor n \rfloor denote the set of positive integers no greater than n; that is, \lfloor n \rfloor = \{1, 2, \ldots, n\}.

II. MAIN RESULTS

A. Problem Statement

Our specific problem of interest here may be formalized as follows. We suppose M \in \mathbb{R}^{n_1 \times n_2} admits a decomposition of the form M = L + C. Here, L is a matrix having rank at most r, and n_L \leq n_2 nonzero columns. The second component C is a column-sparse matrix with k \leq n_2 nonzero “outlier” columns that may occur only at the set of locations where the corresponding column of L is zero. Clearly, n_L \leq n_2 - k. The notion that nonzero columns of C be “outliers” is codified as follows. Let L denote the linear subspace of \mathbb{R}^{n_1} spanned by the columns of L (and having dimension at most r), and denote its orthogonal complement in \mathbb{R}^{n_1} by L^\perp. Let P_L and P_L^\perp denote the orthogonal projection operators onto L and L^\perp, respectively. We assume that the nonzero columns of C occur at the indices \{i : \|P_L C_{:,i}\|_2 > 0\}.

With this setup, our problem of interest here may be stated concisely – our aim is to identify the set I_C.

B. Assumptions

It is well-known in the matrix completion and robust PCA literature that separation of low-rank and sparse matrices from observations of their sum may not be a well-posed task – for example, matrices having only a single nonzero element are simultaneously low rank, sparse, column-sparse, row-sparse, etc. To overcome these identifiability issues, it is common to assume that the linear subspace spanned by the rows and/or columns of the low-rank matrix be “incoherent” with the canonical basis (see, e.g., [17], [38]–[40], [50], among others). Here, we adopt a similar approach, and assume such a condition on the row space of the low-rank component L. We formalize this notion via the following definition from [17].

Definition II.1. (Column Incoherence Property). Let L \in \mathbb{R}^{n_1 \times n_2} be a rank r matrix with at most n_L \leq n_2 nonzero columns, and compact singular value decomposition (SVD) L = U \Sigma V^* where U is n_1 \times r, \Sigma is r \times r, and V is n_2 \times r. The matrix L is said to satisfy the column incoherence property with parameter \mu_L if

\[
\max_i \|V^* e_i\|_2^2 \leq \mu_L \frac{r}{n_L},
\]

where \{e_i\} are basis vectors of the canonical basis for \mathbb{R}^{n_2}.

Note that \mu_L \in [1, n_L/r]; the lower limit is achieved when all elements of V^* have the same amplitude, and the upper limit when any one element of V^* is equal to 1 (i.e., when the row space of L is aligned with the canonical basis).

With this, we may state our structural assumptions concisely, as follows: we assume that the components L and C of the matrix M = L + C satisfy the following structural conditions:

(c1) rank(L) = r,
(c2) L has n_L nonzero columns,
(c3) L satisfies the column incoherence property with parameter \mu_L, and
(c4) |I_C| = k.

C. Recovery Guarantees and Implications

Our main result identifies conditions under which the procedure outlined in Algorithm 1 succeeds. Our particular focus is on measurement matrices satisfying the following property.

Definition II.2. (Distributional Johnson-Lindenstrauss (JL) Property). An m \times n matrix \Phi is said to satisfy the distributional JL property if for any fixed v \in \mathbb{R}^n and any \epsilon \in (0, 1),

\[
\Pr \left( \left| \|\Phi v\|_2^2 - \|v\|_2^2 \right| \geq \epsilon \|v\|_2^2 \right) \leq 2e^{-m f(\epsilon)},
\]

where f(\epsilon) > 0 is a constant depending only on \epsilon that is specific to the distribution of \Phi.

Random matrices satisfying the distributional JL property are those that preserve the length of any fixed vector to within a multiplicative factor of (1 \pm \epsilon) with probability at least 1 - 2e^{-m f(\epsilon)}. By a simple union bounding argument, such matrices can be shown to approximately preserve the lengths of a finite collection of vectors, all vectors in a linear subspace, all vectors in a union of subspaces, etc., provided the number of rows is sufficiently large. As noted in [51], for many randomly constructed and appropriately normalized \Phi, (e.g., such that entries of \Phi are i.i.d. zero-mean Gaussian, or are drawn as an ensemble from any subgaussian distribution), f(\epsilon) is quadratic in \epsilon as \epsilon \to 0. This general framework also allows us to directly utilize other specially constructed fast or sparse JL transforms [52], [53].

With this, we are in position to formulate our main result. We state the result here as a theorem; its proof appears in Section III.

Theorem II.1. Suppose M = L + C, where the components L and C satisfy the structural conditions (c1)-(c4) with

\[
k \leq \frac{1}{40(1 + 121 r \mu_L)} n_2.
\]
For any $\delta \in (0, 1)$, if the column subsampling parameter $\gamma$ satisfies
\[
\gamma \geq \max \left\{ \frac{1}{20}, \frac{200 \log \left( \frac{1}{\delta} \right)}{n_L}, \frac{24 \log \left( \frac{1}{\sqrt{\delta}} \right)}{n_2}, \frac{10r\mu_L \log \left( \frac{1}{\delta} \right)}{n_L} \right\},
\]
the measurement matrices are each drawn from any distribution satisfying (4) with
\[
m \geq \frac{5(r + 1) + \log(k) + \log(2/\delta)}{f(1/2)},
\]
and
\[
p \geq \frac{11k + 2k \log(n_2/k) + \log(2/\delta)}{f(1/4)},
\]
the elements of $\Phi$ are i.i.d. realizations of any continuous random variable, and for any upper bound $k_{ub}$ of $k$ the regularization parameter is set to $\frac{3}{\sqrt{\log(1/k_{ub})}}$ then the following hold simultaneously with probability at least $1 - 3\delta$:
- the ACOS procedure in Algorithm 1 correctly identifies the salient columns of $C$ (i.e., $\tilde{I}_C = I_C$), and
- the total number of measurements collected is no greater than $\left( \frac{\delta}{2} \right) \gamma n_2 + p$.

It is worth noting here that, as with many analyses of this type, the constants arising in the conditions result from conservative bounds throughout the analysis and may thus be (unnecessarily) large. That said, our main result provides an accurate indication of how the performance of our approach scales as a function of the relevant problem parameters. Namely, it follows directly from this result that for appropriate choice of the parameters $\gamma$, $m$, and $p$ the ACOS algorithm correctly identifies the salient columns of $C$ with high probability from relatively few observations, comprising only a fraction of the measurements required by other comparable (non-compressive) procedures [17] that produce the same correct salient support estimate but operate directly on the full $(n_1 \times n_2)$ matrix $M$. Specifically, our analysis shows that the ACOS approach succeeds with high probability with an effective sampling rate of
\[
\frac{\# \text{obs}}{n_1 n_2} = \mathcal{O} \left( \max \left\{ \frac{(r + \log(k/n_2/n_1)) \mu L \log(r)}{n_1}, \frac{(r + \log(k/n_2)) \mu_L}{n_2} \right\} + \frac{k \log(n_2/k)}{n_1} \right),
\]
which may be small when $r$ and $k$ are each small relative to the problem dimensions (and $n_L \sim n_2$ so that $L$ does not have a large number of zero columns outside of $I_C$).

One direct point of comparison for our result is the related work [40], which addresses a different (and in a sense, more difficult) task of identifying both the column space and the set of outlier columns of a matrix $M = L + C$ from observations that take the form of samples of the elements of $M$. There, to deal with the fact that observations take the form of point samples of the matrix (rather than more general linear measurements as here), the authors of [40] assume that $L$ also satisfy a row incoherence property in addition to a column incoherence property, and show that in this setting that the column space of $L$ and set of nonzero columns of $C$ may be recovered from only $\mathcal{O}(n_2 r^2 \mu^2 \log(n_2))$ observations via a convex optimization, where $\mu \in [1, n_1/r]$ is the row incoherence parameter. Normalizing this sample complexity by $n_1 n_2$ facilitates comparison with our result above; we see that the sufficient conditions for the sample complexity of our approach are smaller than for the approach of [40] by a factor of at least $1/r$, and, our approach does not require the row incoherence assumption. We provide some additional, experimental, comparisons between our ACOS method and the RMC method in Section IV.

### III. Proof of Theorem II.1

First, we note that in both of the steps of Algorithm 1 the prescribed observations are functions of $M$ only through $\Phi M$; stated another way, $M$ never appears in the algorithm from isolation from the measurement matrix $\Phi$. Motivated by this, we introduce
\[
\tilde{M} \triangleq \Phi M = \Phi L + \Phi C = \tilde{L} + \tilde{C},
\]
to effectively subsume the action of $\Phi$ into $\tilde{M}$. Now, our proof is a straightforward consequence of assembling three intermediate probabilistic results via a union bounding argument. The first intermediate result establishes that for $M = L + C$ with components $L$ and $C$ satisfying the structural conditions (c1)-(c4), the components $\tilde{L}$ and $\tilde{C}$ of $\tilde{M}$ as defined in (9) satisfy analogous structural conditions provided that $m$, the number of rows of $\Phi$, be sufficiently large. We state this result here as a lemma; its proof appears in Appendix A.

**Lemma III.1.** Suppose $M = L + C$, where $L$ and $C$ satisfy the structural conditions (c1)-(c4). Fix any $\delta \in (0, 1)$, suppose $\Phi$ is an $m \times n_1$ matrix drawn from a distribution satisfying the distributional JL property (4) with $m$ satisfying (7) and let $\tilde{M} = \tilde{L} + \tilde{C}$ be as defined in (9). Then, the components $\tilde{L}$ and $\tilde{C}$ satisfy the following conditions simultaneously with probability at least $1 - \delta$:

1. $\text{rank}(\tilde{L}) = r$,
2. $\tilde{L}$ has $n_1$ nonzero columns,
3. $\tilde{L}$ satisfies the column incoherence property with parameter $\mu_L$, and
4. $I_C = \{ i : \| \Phi_{\tilde{L}_i} C : \| > \delta \} = I_C$, where $\tilde{L}$ is the linear subspace of $\mathbb{R}^m$ spanned by the columns of $\tilde{L}$, and $\Phi_{\tilde{L}_i}$ denotes the orthogonal projection onto the orthogonal complement of $\tilde{L}$ in $\mathbb{R}^m$.

The second intermediate result guarantees two outcomes – first, that Step 1 of Algorithm 1 succeeds in identifying the correct column space of $\tilde{L}$ (i.e., that $\tilde{L}_{(1)} = \tilde{L}$) with high probability provided the components $L$ and $C$ of $\tilde{M}$ as specified in (9) satisfy the structural conditions (c1)-(c4) and the column sampling probability parameter $\gamma$ be sufficiently large, and second, that the number of columns of the randomly generated sampling matrix $S$ be close to $\gamma n_2$. We also provide this result as a lemma; its proof appears in Appendix B.

**Lemma III.2.** Let $\tilde{M} = L + C$ be an $m \times n_2$ matrix, where the components $\tilde{L}$ and $\tilde{C}$ satisfy the conditions (c1)-(c4) with $k$ satisfying (5). Fix $\delta \in (0, 1)$ and suppose the column sampling probability parameter $\gamma$ satisfies (6) when $\lambda = \frac{3}{\sqrt{k_{ub}}}$ for any $k_{ub} \geq \frac{1}{\sqrt{\log(1/\delta)}}$, the following hold simultaneously with probability at least $1 - \delta$: $S$ has $|S| \leq (3/2) \gamma n_2$ columns, and the subspace $\tilde{L}_{(1)}$ resulting from Step 1 of Algorithm 1 satisfies $\tilde{L}_{(1)} = \tilde{L}$. 

\[
\tilde{M} \triangleq \Phi M = \Phi L + \Phi C = \tilde{L} + \tilde{C},
\]
Our third intermediate result shows that the support set of the vector $\hat{c}$ produced in Step 2 of Algorithm 1 is the same as the set of salient columns of $\hat{C}$, provided that $\hat{L}(1) = \hat{L}$ and that $p$, the number of rows of $A$, is sufficiently large. We state this result here as a lemma; its proof appears in Appendix C.

**Lemma III.3.** $\hat{M} = \hat{L} + \hat{C}$ be an $m \times n_2$ matrix, where the components $\hat{L}$ and $\hat{C}$ satisfy the conditions (c1)-(c4) for any $k \leq n_2$, and suppose $\hat{L}(1) = \hat{L}$, the subspace spanned by the columns of $\hat{L}$. Let $\Phi \hat{M} = \hat{M}$ in Step 2 of Algorithm 1. Fix $\delta \in (0, 1)$, suppose $A$ is a $p \times n_2$ matrix drawn from a distribution satisfying the distributional JL property (4) with $p$ satisfying (8), and suppose the elements of $\phi$ are i.i.d. realizations of any continuous random variable. Then with probability at least $1 - \delta$ the support $I_{\hat{C}} = \{ i : \hat{c}_i \neq 0 \}$ of the vector $\hat{c}$ produced by Step 2 of Algorithm 1 satisfies $I_{\hat{C}} = I_{\hat{C}}$.

Our overall result follows from assembling these intermediate results via union bound. In the event that the conclusion of Lemma III.1 holds, then so do the requisite conditions of Lemma III.2. Thus, with probability at least $1 - 2\delta$ the conclusions of Lemmata III.1 and III.2 both hold. This implies that the requisite conditions of Lemma III.3 hold also with probability at least $1 - 3\delta$, and so it follows that the conclusions of all three Lemmata hold with probability at least $1 - 3\delta$.

**IV. Experimental Evaluation**

In this section we provide a comprehensive experimental evaluation of the performance of our approach for both synthetically generated and real data, the latter motivated by a stylized application of saliency map estimation in an image processing task. We compare our method with the Outlier Pursuit (OP) approach of [17] and the Robust Matrix Completion (RMC) approach of [40], each of which employs a convex optimization to identify both the subspace in which the columns of the low rank matrix lie, and the locations of the nonzero columns in the outlier matrix. We implement the OP and RMC methods (as well as the intermediate execution of the OP-like optimization in Step 1 of our approach) using an accelerated approximate alternating direction method of multipliers (ADMM) method inspired by [54] (as well as [17], [55]), and outlined here in Appendix E. We implement the $\ell_1$-regularized estimation in Step 2 of our procedure by casting it as a LASSO problem and also use an accelerated proximal gradient method to obtain its solution. We omit the specific details of that algorithmic approach here, but refer the interested reader to [55].

All experiments were performed with MATLAB R2013a and executed on an iMac with a 3.4 GHz Intel Core i7 processor and 32 GB memory, running OS X 10.8.5.

**A. Synthetic Data**

We experiment on synthetically generated $n_1 \times n_2$ matrices $M$, with $n_1 = 100$ and $n_2 = 1000$, formed as follows. For a specified rank $r$ and number of outliers $k$, we let the number of nonzero columns of $L$ be $n_2 - n_2 - k$, generate two random matrices $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$ with i.i.d. $\mathcal{N}(0, 1)$ entries, and take $L = [UV^T \ 0_{n_1 \times k}]$. We generate the outlier matrix $C = [0_{n_1 \times n_2}]$ where $W \in \mathbb{R}^{n_1 \times k}$ has i.i.d. $\mathcal{N}(0, 1)$ entries (which are also independent of entries of $U$ and $V$). Then, we set $M = L + C$. In all experiments we generate $\phi$, $\Phi$, and $A$ with i.i.d. zero-mean Gaussian entries.

Our first experiment investigates the “phase transition” behavior of our approach; our experimental setting is as follows. First, we set the average sampling rate by fixing the column downsampling fraction $\gamma = 0.2$, and choosing a row sampling parameter $m \in \{0.1n_1, 0.2n_1, 0.3n_1\}$ and column sampling parameter $p \in \{0.1n_2, 0.2n_2, 0.3n_2\}$. Then, for each $(r, k)$ pair with $r \in \{2, 4, 6, \ldots, 40\}$ and $k \in \{5, 10, 15, \ldots, 100\}$ we generate a synthetic matrix $M$ as above, and for each of 3 different values of the regularization parameter $\lambda \in \{0.3, 0.4, 0.5\}$ we perform 30 trials of Algorithm 1 recording in each whether the recovery approach succeeded in identifying the locations of the true outliers for that value of $\lambda$, and associate to each $(r, k)$ pair the (empirical) average success rate. Then, at each $(r, k)$ point examined we identify

\[\begin{align*}
\text{(a) } 2.1\% & \quad \text{(b) } 2.2\% & \quad \text{(c) } 2.3\% \\
\text{(d) } 4.1\% & \quad \text{(e) } 4.2\% & \quad \text{(f) } 4.3\% \\
\text{(g) } 6.1\% & \quad \text{(h) } 6.2\% & \quad \text{(i) } 6.3\%
\end{align*}\]

Fig. 1. Outlier recovery phase transitions plots for ACOS for various sampling parameters (white regions correspond to successful recovery). Each row of the figure corresponds to a different level of compression of columns of $M$, where $m = 0.1n_1, 0.2n_1, 0.3n_1$, respectively, from top to bottom. Each column corresponds to a different level of compression of rows of $M$ in Step 2 of Algorithm 1, with $p = 0.1n_2, 0.2n_2, 0.3n_2$, respectively, from left to right. The fraction of observations obtained (as a percentage, relative to the full dimension) is provided as a caption below each figure. As expected, increasing $m$ (top to bottom) facilitates accurate estimation for increasing rank $r$ of $L$, while increasing $p$ (left to right) allows for recovery of increasing numbers $k$ of outlier columns.

\[\begin{align*}
1\text{We solve the optimization associated with Step 2 of our approach as a LASSO problem, with 10 different choices of regularization parameter } \\
\mu \in (0, 1). \text{ We deem any trial a success if for at least one value of } \mu, \text{ there exists a threshold } \tau > 0 \text{ such that } \\
\min_{\beta \in \mathbb{R}^{n_2}} \|eta(\mu)\|_2 > \tau > \max_{z \in \mathbb{R}^{n_2}} \|z(\mu)\|_2 \text{ for the estimate } \hat{c}(\mu) \text{ produced in Step 2. An analogous threshold-based methodology was employed to assess the outlier detection performance of the Outlier Pursuit approach in [17].}
\end{align*}\]
the point-wise maximum of the average success rates for the 3 different values of $\lambda$; in this way, we assess whether recovery for that $(r, k)$ is achievable by our method for the specified sampling regime for some choice of regularization parameters. The results in Figure 1 depict the outcome of this experiment for the 9 different sampling regimes examined. For easy comparison, we provide the average sampling rate as fraction of observations obtained (relative to the full matrix dimension) in the caption in each figure.

The results of this experiment provide an interesting, and somewhat intuitive, illustration of the efficacy of our approach. Namely, we see that increasing the parameter $m$ of the matrix $\Phi$ in Step 1 of our algorithm while keeping the other sampling parameters fixed (i.e., moving from top to bottom in any one column) facilitates accurate recovery for increasing ranks $r$ of the matrix $L$. Similarly, increasing the parameter $p$ of the matrix $A$ in Step 2 of our algorithm while keeping the other sampling parameters fixed (i.e., moving from left to right in any one row) facilitates accurate recovery for an increasing number $k$ of outlier columns. Overall, our approach can successfully recover the locations of the outliers for non-trivial regimes of $r$ and $k$ using very few measurements – see, for instance, panel (i), where $\sim 30$ outlier columns can be accurately identified in the presence of a rank 30 background using an effective sampling rate of only $\sim 6.3\%$.

We also compute phase transition curves for RMC using a similar methodology\(^2\) to that described above. Those results are provided in Figure 2. We observe\(^3\) that RMC approach is viable for identifying the outliers from subsampled data provided the sampling rate exceeds about 10%, but even then only for small values of the rank $r$. As alluded in the discussion in previous sections, the relative difference in performance is likely due in large part to the difference in the observation models between the two approaches – the RMC approach is inherently operating in the presence of “missing data” (a difficult scenario!) while our approach permits us to observe linear combinations of any row or column of the entire matrix (i.e., we are allowed to “see” each entry of the matrix, albeit not necessarily individually, throughout our approach).

\(^2\)Here, we aggregate the results for three different choices of the regularization parameter $\lambda \in \{0.4, 0.5, 0.6\}$.

\(^3\)Our evaluation of RMC here yields results that agree qualitatively with those in [40], where sampling rates around 10% yielded successful recovery for small $r$.

### B. Real Data

We also evaluate the performance of our proposed method on real data in the context of a stylized image processing task that arises in many computer vision and automated surveillance – that of identifying the “saliency map” of an image. For this, we use images from the MSRA Salient Object Database [13] available online\(^4\).

As discussed above, our approach here is based on representing each test image as a collection of (vectorized) non-overlapping image patches. We transform each (color) test image to gray scale, decompose it into non-overlapping $10 \times 10$-pixel patches, vectorize each patch into a $100 \times 1$ column vector, and assemble the column vectors into an image. Most of the images in the database are of the size $300 \times 400$ (or $400 \times 300$), which here yields matrices of size $100 \times 1200$, corresponding to 1200 patches. Notice that we only used gray scale values of image as the input feature rather than any high-level images feature – this facilitates the use of our approach, which is based on collecting linear measurements of the data (e.g., using a spatial light modulator, or an architecture like the single pixel camera [56]).

Here, our experimental approach is (somewhat necessarily) a bit more heuristic than for the synthetic data experiments above, due in large part to the fact that the data here may not adhere exactly to the low-rank plus outlier model. To compensate for this potential model mismatch, we augment Step 1 of Algorithm 1 with an additional “rank reduction” step, where we further reduce the dimension of the subspace spanned by the columns of the learned $\hat{L}(1)$ by truncating its SVD to retain the smallest number of leading singular values whose sum is at least $0.95 \times \|\hat{L}(1)\|_F$. As above, we implement the estimation procedure of Step 2 via LASSO using several different values of the regularization parameter, yielding outputs of varying sparsity levels. If a column is detected to be salient the corresponding patch in the original image is marked white, otherwise it is marked black. We use a visual heuristic method to determine the “best” LASSO threshold, qualitatively trading off false positives with misses.

We implement our ACOS method using two different sampling regimes, the first corresponding to $\gamma = 0.2$, $m = 0.2n_1$, $p = 0.4n_2$ (an average 4.4% sampling rate) and the other with $\gamma = 0.2$, $m = 0.1n_1$, $p = 0.4n_2$ (an average 2.4% sampling rate). As in the previous section, we generate the $\Phi$ and $A$ matrices for the ACOS approach to have i.i.d. zero-mean Gaussian entries. We compare our approach with the OP approach (which uses the full data) and the RMC approach at sampling rates of 20% and 5%. We again used a visual heuristic to determine the “best” outputs for the OP and RMC methods; here, this corresponds essentially to declaring an image patch to be salient only when its column norm is sufficiently large (instead of strictly nonzero). The results of this experiment are provided in Figure 3.

We note first that the OP approach performs fairly well at identifying the visually salient regions in the image, providing evidence to validate the use of the low-rank plus outlier model

\(^4\)See [55].
V. DISCUSSION AND FUTURE DIRECTIONS

It is illustrative here to note a key difference between our approach and more traditional CS tasks. Namely, the goal of the original CS works \cite{Candes2006a, Donoho2006, Eldar2012} and numerous follow-on efforts was to exactly recover or reconstruct a signal from compressive measurements, whereas the nature of our task here is somewhat simpler, amounting to a kind of multidimensional “support recovery” task (albeit in the presence of a low-rank “background”). Exactly recovering the low-rank and column-sparse components would be sufficient for the outlier identification task we consider here, but as our analysis shows it is not strictly necessary. This is the insight that we exploit when operating on the “compressed” data $\Phi M$ instead of the original data matrix $M$. Ultimately, this allows us to successfully identify the locations of the outliers without first estimating the original (full size) low-rank matrix or the outliers themselves. For some regimes of $\mu_L$, $r$, and $k$, we accomplish the outlier identification task using as few as $O((r + \log k)(\mu_L r \log r) + k \log(n^2/k))$ observations.

Along related lines, it is reasonable to conjecture that any procedure would require at least $r^2 + k$ measurements in order to identify $k$ outliers from an $r$-dimensional linear subspace. Indeed, a necessary condition for the existence of outliers of a rank-$r$ subspace, as we have defined them, is that the number of rows of $M$ be at least $r + 1$. Absent any additional structural conditions on the outliers and the subspace spanned by columns of the low-rank matrix, one would need to identify a collection of $r$ vectors that span the $r$-dimensional subspace containing the column vectors of the low-rank component (requiring specification of some $O(r^2)$ parameters) as well as the locations of the $k$ outliers (which would entail specifying another $k$ parameters). In this sense, our approach may be operating near the sample complexity limit for this problem, at least for some regimes of $\mu_L$, $r$, and $k$.

It would be interesting to see whether the dimensionality reduction insight that we exploit in our approach could be leveraged in the context of the Compressive Principal Component Pursuit (Compressive PCP) of \cite{Rauhut2013} in order to yield a procedure with comparable performance as ours, but which acquires only non-adaptive linear measurements of $M$. Namely, one could envision implementing the Compressive PCP method not on the full data $M$, but on the a priori compressed data $\Phi M$. Our Lemma III.1 establishes that the row compression step preserves rank and column incoherence properties, so it is plausible that the Compressive PCP approach may succeed in recovering the components of the compressed matrix, which would suffice for the outlier identification task. We defer this investigation to a future effort.

**Table I**

| Method | OP Sampling | RMC 20% | RMC 5% | ACOS 4% | ACOS 2.4% |
|--------|-------------|---------|--------|---------|------------|
| Step 1 | 7.8140 (1.3429) | 7.6118 (1.0814) | 7.9139 (1.3106) | 0.1905 (0.0272) | 0.0825 (0.0119) |
| Step 2 | – | – | – | 0.2010 (0.0674) | 0.2014 (0.0692) |

**Discussion and Future Directions**

It is illustrative here to note a key difference between our approach and more traditional CS tasks. Namely, the goal of the original CS works \cite{Candes2006a, Donoho2006, Eldar2012} and numerous follow-on efforts was to exactly recover or reconstruct a signal from compressive measurements, whereas the nature of our task here is somewhat simpler, amounting to a kind of multidimensional “support recovery” task (albeit in the presence of a low-rank “background”). Exactly recovering the low-rank and column-sparse components would be sufficient for the outlier identification task we consider here, but as our analysis shows it is not strictly necessary. This is the insight that we exploit when operating on the “compressed” data $\Phi M$ instead of the original data matrix $M$. Ultimately, this allows us to successfully identify the locations of the outliers without first estimating the original (full size) low-rank matrix or the outliers themselves. For some regimes of $\mu_L$, $r$, and $k$, we accomplish the outlier identification task using as few as $O((r + \log k)(\mu_L r \log r) + k \log(n^2/k))$ observations.

Along related lines, it is reasonable to conjecture that any procedure would require at least $r^2 + k$ measurements in order to identify $k$ outliers from an $r$-dimensional linear subspace. Indeed, a necessary condition for the existence of outliers of a rank-$r$ subspace, as we have defined them, is that the number of rows of $M$ be at least $r + 1$. Absent any additional structural conditions on the outliers and the subspace spanned by columns of the low-rank matrix, one would need to identify a collection of $r$ vectors that span the $r$-dimensional subspace containing the column vectors of the low-rank component (requiring specification of some $O(r^2)$ parameters) as well as the locations of the $k$ outliers (which would entail specifying another $k$ parameters). In this sense, our approach may be operating near the sample complexity limit for this problem, at least for some regimes of $\mu_L$, $r$, and $k$.

It would be interesting to see whether the dimensionality reduction insight that we exploit in our approach could be leveraged in the context of the Compressive Principal Component Pursuit (Compressive PCP) of \cite{Rauhut2013} in order to yield a procedure with comparable performance as ours, but which acquires only non-adaptive linear measurements of $M$. Namely, one could envision implementing the Compressive PCP method not on the full data $M$, but on the a priori compressed data $\Phi M$. Our Lemma III.1 establishes that the row compression step preserves rank and column incoherence properties, so it is plausible that the Compressive PCP approach may succeed in recovering the components of the compressed matrix, which would suffice for the outlier identification task. We defer this investigation to a future effort.

**Fig. 3.** Detection results for the MSRA Salient Object Database for various methods. Our ACOS approach produces results comparable to the “full sampling” OP method using an average sampling rate below 5%. The performance of the RMC approach appears to degrade at low sampling rates. for visual saliency (see also \cite{Soatto2016}). Next, comparing the results of the individual procedures, we see that the OP approach appears to uniformly give the best detection results, which is reasonable since it is using the full data as input. The RMC approach performs well at the 20% sampling rate, but its performance appears to degrade at the 5% sampling rate. The ACOS approach, on the other hand, still produces reasonably accurate results using only 2–4% sampling rate.

We also compare implementation times of the algorithms on this saliency map estimation task. Table I provides the average execution times (and standard deviations) for each approach, evaluated over 1000 images in the MSRA database. Here, we only execute each procedure for one choice of regularization parameter, and we also include the additional “rank reduction” step discussed above for the ACOS method. Overall, we see our approach may be more than 25× faster than the OP and RMC methods. At the expense of increased sample complexity, one could set $A = I$ in Step 2 of the ACOS method and thus eliminate the need to implement an iterative solver there (for orthonormal $A$, the LASSO is solved by coordinate-wise soft thresholding). This could significantly speed up the second step of our method, and could result overall in relative speedups of up to 100×. Overall, our results suggest a significant improvement obtained via ACOS for both detection consistency and timing, which may have a promising impact in a variety of salient signal detection tasks.

**Table I**

**Timing analysis for detection experiments on 1000 images from MRSA database. Each entry is the mean execution time in seconds with the standard deviation in parenthesis.**

| Method | OP Sampling | RMC 20% | RMC 5% | ACOS 4% | ACOS 2.4% |
|--------|-------------|---------|--------|---------|------------|
| Step 1 | 7.8140 (1.3429) | 7.6118 (1.0814) | 7.9139 (1.3106) | 0.1905 (0.0272) | 0.0825 (0.0119) |
| Step 2 | – | – | – | 0.2010 (0.0674) | 0.2014 (0.0692) |
We also comment briefly on the computational complexities of the methods we examined. We consider first the OP and RMC approaches, and assume that the solvers for each utilize an iterative accelerated first-order method (like that outlined in Appendix E). In this case, the computational complexity will be dominated by the SVD step in each iteration. Now, for an \( n_1 \times n_2 \) matrix the computational complexity of the SVD is \( O(n_1 n_2 \cdot \min\{n_1, n_2\}) \); with this, and assuming some IT iterations are used, we have that the complexities of both OP and RMC scale as \( O(\text{IT} \cdot [n_1 n_2 \cdot \min\{n_1, n_2\}] \). By a similar analysis, we can conclude that the complexity of Step 1 of the ACOS method scales like \( O(\text{IT}_1 [m (\gamma n_2) \cdot \min\{m, \gamma n_2\}] + \text{IT}_2 [m n_2]) \).

| Method   | Complexity                                                                 |
|----------|-----------------------------------------------------------------------------|
| OP       | \( O(\text{IT} \cdot [n_1 n_2 \cdot \min\{n_1, n_2\}] \)                  |
| RMC      | \( O(\text{IT} \cdot [n_1 n_2 \cdot \min\{n_1, n_2\}] \)                  |
| ACOS     | \( O(\text{IT}_1 [m (\gamma n_2) \cdot \min\{m, \gamma n_2\}] + \text{IT}_2 [m n_2]) \) |

for any choice of \( \epsilon < 1/2 \). Second, we show that for any \( \delta \in (0, 1) \), generating \( \Phi \) as a random matrix as specified in the lemma ensures it will be a \( \sqrt{2}/4 \)-stable embedding of (11) with probability at least \( 1 - \delta \). The choice of \( \sqrt{2}/4 \) in the last step is somewhat arbitrary – we choose this fixed value for concreteness here, but note that the structural conclusions of the lemma follow using any choice of \( \epsilon < 1/2 \) (albeit with slightly different conditions on \( m \)).

1) Part 1: Throughout this portion of the proof we assume that \( \Phi \) is an \( \epsilon \)-stable embedding of (11) for some \( \epsilon < 1/2 \), and establish each of the four claims in turn. First, to establish that \( \text{rank}(\Phi L) = r = \text{rank}(L) \), we utilize an intermediate result of [51], stated here as a lemma (without proof) and formulated in the language of stable embeddings.

**Lemma A.1** (Adapted from [51], Theorem 1). Let \( L \) be an \( n_1 \times n_2 \) matrix of rank \( r \), and let \( L \) denote the column space of \( L \), which is a \( r \)-dimensional linear subspace of \( \mathbb{R}^{n_1} \). If for some \( \epsilon \in (0, 1) \), \( \Phi \) is an \( \epsilon \)-stable embedding of \( L, \{0\} \) then \( \text{rank}(\Phi L) = r = \text{rank}(L) \).

Here, since \( \Phi \) being an \( \epsilon \)-stable embedding of (11) implies it is also an \( \epsilon \)-stable embedding of \( (L, \{0\}) \), the first claim (of Lemma III.1) follows from Lemma A.1.

Next we show that \( \Phi L \) has \( n_1 \) nonzero columns. Since \( \Phi \) is a stable embedding of \( (L, \{0\}) \), it follows that for each of the \( n_1 \) nonzero columns \( L_{:, i} \) of \( L \) we have \( \|\Phi L_{:, i}\|_2 > (1 - \epsilon)\|L_{:, i}\|_2 > 0 \), while for each of the remaining \( n_2 - n_1 \) columns \( L_{:, j} \) of \( L \) that are identically zero we have \( \|\Phi L_{:, j}\|_2 = 0 \) so that \( L_{:, j} = 0 \).

Continuing, we show next that \( \Phi L \) satisfies the column incoherence property with parameter \( \mu_L \). Recall from above that we write the compact SVD of \( L \) as \( L = U \Sigma V^* \), where \( U \) is \( n_1 \times r \), \( V \) is \( n_2 \times r \), and \( \Sigma \) is an \( r \times r \) nonnegative diagonal matrix of singular values (all of which are strictly positive). The incoherence condition on \( L \) is stated in terms of column norms of the matrix \( V^* \) whose rows form an orthonormal basis for the row space of \( L \). Now, when the rank of \( \Phi L \) is the same as that of \( L \), which is true here on account of Lemma A.1, the row space of \( \Phi L \) is identical to that of \( L \), since each are \( r \)-dimensional subspaces of \( \mathbb{R}^{n_2} \) spanned by linear combinations of the columns of the \( V^* \). Thus since the rank and number of nonzero columns of \( \Phi L \) are the same as for \( L \), the coherence parameter of \( \Phi L \) is just \( \mu_L \), and the third claim is established.

Finally, we establish the last claim, that the set of salient columns of \( \Phi C \) is the same as for \( C \). Recall that the condition that a column \( C_{:, i} \) be salient was equivalent to the condition that \( \|P_{C_{:, i}} C_{:, i}\|_2 > 0 \), where \( P_{C_{:, i}} \) is the orthogonal projection operator onto the orthogonal complement of \( C_{:, i} \) in \( \mathbb{R}^{n_1} \). Here, our aim is to show that an analogous result holds in the “projected” space – that for all \( i \in \mathcal{I}_C \) we have \( \|P_{\Phi C_{:, i}} \Phi C_{:, i}\|_2 > 0 \), where \( \Phi L \) is the linear subspace spanned by the columns of \( \Phi L \). For this we utilize an intermediate result of [57] formulated there in terms of a “compressive interference cancellation” method. We state an adapted version of that result here as a lemma (without proof).

**Lemma A.2** (Adapted from [57], Theorem 5). Let \( V_1 \) be an \( r \)-dimensional linear subspace of \( \mathbb{R}^n \) with \( r < n \), let \( V_2 \) be...
any subset of \(\mathbb{R}^n\), and let
\[
\tilde{V}_2 = \{ P_{\tilde{V}_2}v : v \in V_2 \},
\]
where \(P_{\tilde{V}_2}\) is the orthogonal projection operator onto the orthogonal complement of \(V_1\) in \(\mathbb{R}^n\). If \(\Phi\) is an \(\epsilon\)-stable embedding of \((V_1, \tilde{V}_2 \cup \{0\})\), then for all \(v \in \tilde{V}_2\)
\[
\|P_{\Phi V_1} (\Phi v)\|_2^2 \geq \left(1 - \frac{\epsilon}{1+\epsilon}\right) \|v\|_2^2,
\]
where \(P_{\Phi V_1}\) is the orthogonal projection operator onto the orthogonal complement of the subspace of \(\mathbb{R}^n\) spanned by the elements of \(\Phi V_1 = \{ \Phi v : v \in V_1 \}\).

Before applying this result we first note a useful fact, that \(\Phi\) being an \(\epsilon\)-stable embedding of \((V_1, \tilde{V}_2 \cup \{0\})\) is equivalent to \(\Phi\) being an \(\epsilon\)-stable embedding of \((V_1, \tilde{V}_2 \cup \{0\})\), which follows directly from the definition of stable embeddings and the (easy to verify) fact that \(\{v_1 - \tilde{v}_2 : v_1 \in V_1, \tilde{v}_2 \in \tilde{V}_2 \cup \{0\}\} = \{v_1 - v_2 : v_1 \in V_1, v_2 \in V_2 \cup \{0\}\}\).

To prove Lemma A.2 here, we let \(V_1 = V,\) \(\tilde{V}_2 = V \cap \bigcup_{i \in I_C} \{C_i, i\}\), and \(V_2 = \bigcup_{i \in I_C} \{C_i, i\}\). Since \(\Phi\) is an \(\epsilon\)-stable embedding of (11), we have that for all \(i \in I_C\),
\[
\|P_{(\Phi C) \cup \{\Phi \}} \phi_{C, i, j}\|_2^2 \geq \left(1 - \frac{\epsilon}{1+\epsilon}\right) \|P_{(\Phi C) \cup \{\Phi \}} \phi_{C, i, j}\|_2^2.
\]
Thus, we can conclude that \(I_C = C\).

2) Part 2: Given the structural result established in the previous step, the last part of the proof entails establishing that a random matrix \(\Phi\) generated as specified in the statement of Lemma III.1 is an \(\sqrt{2}/4\)-stable embedding of (11). Our approach here begins with a brief geometric discussion, and a bit of “stable embedding algebra.” Appealing to the definition of stable embeddings, we see that \(\Phi\) being an \(\epsilon\)-stable embedding of (11) is equivalent to \(\Phi\) being such that
\[
(1 - \epsilon)\|v\|_2^2 \leq \|\Phi v\|_2^2 \leq (1 + \epsilon)\|v\|_2^2
\]
holds for all \(v \in L \cup \bigcup_{i \in I_C} L - C_i\), denoting the \(r\)-dimensional affine subspace of \(\mathbb{R}^{n_2}\) comprised of all elements taking the form of a sum between a vector in \(L\) and the fixed vector \(C_i\). Thus, in words, establishing our claim here entails showing that a random \(\Phi\) (generated as specified in the lemma, with appropriate dimensions) approximately preserves the lengths of all vectors in a union of subspaces comprised of one \(r\)-dimensional linear subspace and some \(|I_C| = k\), \(r\)-dimensional affine subspaces.

Stable embeddings of linear subspaces using random matrices is, by now, well-studied (see, e.g., [51], [57], [58], as well as a slightly weaker result [59, Lemma 10.1]), though stable embeddings of affine subspaces has received less attention in the literature. Fortunately, using a straightforward argument we may leverage results for the former in order to establish the latter. Recall the discussion above, and suppose that rather than establishing that (13) holds for all \(v \in L \cup \bigcup_{i \in I_C} L - C_i\), we instead establish a slightly stronger result, that (13) holds for all \(v \in L \cup \bigcup_{i \in I_C} L^i\), where for each \(i \in I_C\), \(L^i\) denotes the \((r+1)\)-dimensional linear subspace of \(\mathbb{R}^{n_2}\) spanned by the columns of the matrix \(L C_{i, i}\). (That the dimension of each \(L^i\) be \(r+1\) follows from the assumption that columns \(C_{i, i}\) for \(i \in I_C\) be outliers.) Clearly, if for some \(i \in I_C\) the condition (13) holds for all \(v \in L^i\), then it holds for all vectors formed as linear combinations of \([L C_{i, i}]\), so it holds in particular for all vectors in the \(r\) dimensional affine subspace denoted by \(L - C_{i, i}\). Further, that (13) holds for any \(i \in I_C\) implies it holds for linear combinations that use a weight of zero on the component \(C_{i, i}\), so in this case (13) holds also for all \(v \in L\).

Based on the above discussion, we see that a sufficient condition to establish that \(\Phi\) be an \(\epsilon\)-stable embedding of (11) is that (13) hold for all \(v \in \bigcup_{i \in I_C} L^i\); in other words, that \(\Phi\) preserve (up to multiplicative \((1 \pm \epsilon)\) factors) the squared lengths of all vectors in a union of \((r+1)\)-dimensional linear subspaces of \(\mathbb{R}^{n_2}\). To this end we make use of another result adapted from [51], and based on the union of subspaces embedding approach utilized in [58].

**Lemma A.3** (Adapted from [51], Lemma 1). Let \(\bigcup_{i=1}^k V^i\) denote a union of \(k\) linear subspaces of \(\mathbb{R}^{n_2}\), each of dimension at most \(d\). For fixed \(\epsilon \in (0, 1)\) and \(\delta \in (0, 1)\), suppose \(\Phi\) is an \(m \times n\) matrix satisfying the distributional JL property with
\[
m \geq \frac{d \log(42/\epsilon) + \log(k) + \log(2/\delta)}{f(\epsilon/\sqrt{2})}
\]
Then (1 - \(\epsilon\))\|v\|_2^2 \leq \|\Phi v\|_2^2 \leq (1 + \epsilon)\|v\|_2^2 holds simultaneously for all \(v \in \bigcup_{i=1}^k V^i\) with probability at least \(1 - \delta\).

Applying this lemma here with \(d = r+1\) and \(\epsilon = 1/4\), and using the fact that \(\log(8\sqrt{2}) < 5\) yields the final result.

**B. Proof of Lemma III.2**

Our approach is comprised of two parts. In the first, we show that the two claims of Lemma III.2 follow directly when the following five conditions are satisfied

(a1) \(S\) has \((1/2)\gamma n_2 \leq |S| \leq (3/2)\gamma n_2\) columns,

(a2) \(LS\) has at most \((3/2)\gamma n_L\) nonzero columns,

(a3) \(CS\) has at most \(k\) nonzero columns,

(a4) \(\sigma_1^2(\hat{V}^*S) \leq (3/2)\gamma^2\) and

(a5) \(\sigma_k^2(\hat{V}^*S) \geq (1/2)\gamma^2\),

where the matrix \(\hat{V}^*\) that arises in (a4)-(a5) is the matrix of right singular vectors from the compact SVD \(L = USV^*\) of \(L\), and \(\sigma_i(\hat{V}^*S)\) denotes the \(i\)-th largest singular value of \(\hat{V}^*S\). Then, in the second part of the proof we show that (a1)-(a5) hold with high probability when \(S\) is a random subsampling matrix generated with parameter \(\gamma\) in the specified range.

We briefly note that parameters \((1/2)\) and \((3/2)\) arising in the conditions (a1)-(a5) are somewhat arbitrary, and are fixed to these values here for ease of exposition. Analogous results to that of Lemma III.2 could be established by replacing \((1/2)\) with any constant in \((0, 1)\) and \((3/2)\) with any constant larger than \(1\), albeit with slightly different conditions on \(\gamma\).

1) Part 1: Throughout this portion of the proof, we assume that conditions (a1)-(a5) hold. Central to our analysis is a main result of [17], which we state as a lemma (without proof).

**Lemma A.4** (Outlier Pursuit, adapted from [17]). Let \(\hat{M} = \hat{L} + \hat{C}\) be an \(n_1 \times n_2\) matrix whose components \(\hat{L}\) and \(\hat{C}\) satisfy the structural conditions
\[
(\epsilon_1) \quad \text{rank}(\hat{L}) = \hat{r},
\]
(e2) \( \hat{L} \) has \( n_L \) nonzero columns,
(e3) \( \hat{L} \) satisfies the column incoherence property with parameter \( \mu_L \), and
(e4) \( |\mathcal{C}| = \{ i : \| \mathbf{P}_{\hat{L}^\perp} \mathbf{C}_i \|_2 > 0 \} = k, \) where \( \hat{L} \) denotes the linear subspace spanned by columns of \( L \) and \( \mathbf{P}_{\hat{L}^\perp} \) is the orthogonal projection operator onto the orthogonal complement of \( \hat{L} \) in \( \mathbb{R}^{m_1} \),

\[
\hat{k} \leq \left( \frac{1}{1 + \frac{121}{9} \frac{9}{\sqrt{n_L}}} \right) \hat{n}_2. \quad (15)
\]

For any upper bound \( \hat{k}_{ab} \geq \hat{k} \) and \( \lambda = \frac{3}{\sqrt{k_{ab}}} \) any solutions of the outlier pursuit procedure

\[
\{ \hat{L}, \hat{C} \} = \min_{L,C} \| L \|_* + \lambda \| C \|_{1,2} \text{ s.t. } \bar{M} = L + C, \quad (16)
\]

are such that the columns of \( \hat{L} \) span the same linear subspace as the columns of \( L \), and the set of nonzero columns of \( \hat{C} \) is the same as the set of locations of the nonzero columns of \( C \).

Introducing the shorthand notation \( \hat{L} = \hat{L}S \), \( \hat{C} = \hat{C}S \), and \( \hat{n}_2 = |S| \), our approach will be to show that conditions (a1)-(a5) along with the assumptions on \( \bar{M} \) ensure that (e1)-(e4) in Lemma A.4 are satisfied for some appropriate parameters \( \hat{r}, n_L, \mu_L, \), and \( \hat{k} \) that depend on analogous parameters of \( \bar{M} \).

First, note that (a5) implies that the matrix \( \hat{V}^* \hat{S} \) has rank \( r \), which in turn implies that \( \hat{L} \) has rank \( r \). Thus, (e1) is satisfied with \( \hat{r} = r \). The condition (e2) is also satisfied here for \( n_L \) no larger than \( \left( \frac{3}{2} \right)^2 \gamma n_L \); this is a restatement of (a2).

We next establish (e3). To this end, note that since \( \hat{L} \) has rank \( r \), it follows that the \( r \)-dimensional linear subspace spanned by the rows of \( \hat{L} = \hat{L}S \) is \( \mathbb{R}^{m_1} \) and that the columns of \( \hat{V} \) span the same linear subspace as the columns of \( \hat{Y} \). Now, let \( S^T \hat{V} \) denote the \( r \)-dimensional linear subspace of \( \mathbb{R}^{n_2} \) spanned by the columns of \( S^T \hat{V} \) and let \( \mathbf{P}_{S^T \hat{V}} \) denote the orthogonal projection operator onto \( S^T \hat{V} \). Then, bounding the column incoherence parameter of \( \hat{L} \) entails establishing an upper bound on

\[
\max_{i \in [n_2]} \| \mathbf{P}_{S^T \hat{V}} e_i \|_2^2 = \max_{i \in [n_2]} \| S^T \hat{V} (\hat{V}^* S S^T \hat{V})^{-1} \hat{V}^* e_i \|_2^2.
\]

where (a) follows from the fact that for any \( i \in [\hat{n}_2] \) the vector \( S e_i \) is either the zero vector or one of the canonical basis vectors for \( \mathbb{R}^{n_2} \), (b) follows from straightforward linear algebraic bounding ideas and the column incoherence assumption on \( L \), and (c) follows from (a4)-(a5). Now, we let \( n_L \) denote the number of nonzero columns of \( L \), and write

\[
\max_{i \in [n_2]} \| \mathbf{P}_{S^T \hat{V}} e_i \|_2^2 \leq \left( \frac{6}{\gamma} \right) \mu_L \frac{r}{n_L} \frac{n_L}{n_L} \leq 9 \mu_L \frac{r}{n_L}, \quad (18)
\]

where the last inequality uses (a2). Thus (e3) holds with

\[
\mu_L = 9 \mu_L. \quad (19)
\]

Next, we establish (e4). Recall from above that \( \hat{L} \) has rank \( r \), and is comprised of columns of \( L \); it follows that the subspace \( \hat{L} \) spanned by columns of \( \hat{L} \) is the same as the subspace \( L \) spanned by columns of \( L \). Thus, \( \| \mathbf{P}_{\hat{L}^\perp} \mathbf{C}_i \|_2 = \| \mathbf{P}_{L^\perp} \mathbf{C}_i \|_2 \), so to obtain an upper bound on \( \hat{k} \) we can simply count the number \( \hat{k} \) of nonzero columns of \( \hat{C} \). By (a3),

\[
\hat{k} \leq \left( \frac{1}{20 \frac{1 + 121}{9} \frac{9}{\sqrt{n_L}}} \right) \left( \frac{1}{2} \right) n_2.
\]

and (b) follows from (a1) and (19) as well as the fact that \( \hat{r} = r \).

Finally, we show that the two claims of Lemma III.2 hold. The first follows directly from (a1). For the second, note that for any \( k_{ab} \geq k \) we have that \( k_{ab} \leq \hat{k}_{ab} \leq k \). Thus, since \( \lambda = \frac{3}{\sqrt{k_{ab}}} = \frac{3}{\sqrt{\hat{k}_{ab}}} \) and (e1)-(e4) hold, it follows from Lemma A.4 that the optimization (16) produces an estimate \( \hat{L} \) whose columns span the same linear subspace as that of \( L \). But, since \( \hat{L} \) has rank \( r \) and its columns are just a subset of columns of the rank-\( r \) matrix \( \hat{L} \), the subspace spanned by the columns of \( \hat{L} \) is the same as that spanned by columns of \( L \).

2) Part 2: The last part of our proof entails showing (a1)-(a5) hold with high probability when \( S \) is randomly generated as specified. Let \( E_1, \ldots, E_5 \) denote the events that conditions (a1)-(a5), respectively, hold. Then \( Pr\left( \left\{ \bigcup_{i=1}^5 E_i \right\} \right) \leq \sum_{i=1}^5 Pr(E_i^c), \) and we consider each term in the sum in turn.

First, since \( |S| \) is a Binomial\((n_2, \gamma)\) random variable, we may bound its tails using [60, Theorem 2.3 (b-c)]. This gives that \( Pr(|S| > 3\gamma n_2/2) \leq \exp(-3\gamma n_2/28) \) and \( Pr(|S| < \gamma n_2/2) \leq \exp(-\gamma n_2/8) \). By union bound, we obtain that \( Pr(E_1^c) \leq \exp(-3\gamma n_2/28) + \exp(-\gamma n_2/8) \).}

Next, observe that conditionally on \( |S| = s \), the number of nonzero columns present in the matrix \( LS \) is a hypergeometric random variable parameterized by a population of size \( n_2 \) with \( n_L \) positive elements and \( s \) draws. Denoting this hypergeometric distribution here by hyp\((n_2, n_L, s)\) and letting \( H_{|S|} \sim hyp(n_2, n_L, |S|) \), we have that \( Pr(E_2^c) = Pr(H_{|S|} > \left( \frac{3}{2} \right) \gamma n_L ) \). Using a simple conditioning argument, \( Pr(E_2^c) \leq \sum_{r=\left( \frac{3}{2} \right) \gamma n_L}^{s} Pr(H_r > \left( \frac{3}{2} \right) \gamma n_L ) Pr(|S| = s) + Pr(|S| > \left( \frac{3}{2} \right) \gamma n_L ) \), and our next step is to simplify the terms in the sum. Note that for any \( s \) in the range of summation, we have \( Pr(H_r > \left( \frac{3}{2} \right) \gamma n_L ) = \exp\left( -\frac{3s(n_L/n_2)}{400} \right) \).

Pr\left( H_s > \left( \frac{3}{2} \right) \gamma n_L \right) \leq \exp\left( -\frac{3s(n_L/n_2)}{400} \right).

(21)
where (a) utilizes the largest value of $s$ to bound the term $\gamma n_2/s$, (b) follows from an application of Lemma A.6 in Appendix D, and (c) results from using the smallest value of $s$ (within the range of summation) to bound the error term. Assembling these results, we have that $\Pr(\mathcal{E}_2^z) \leq \exp(-\gamma n_2/200) + \exp(-\gamma n_2/24) + \exp(-\gamma n_2/18)$, where we use the fact that the probability mass function of $|S|$ sums to one, and another application of [60, Theorem 2.3(b,c)].

Bounding $\Pr(\mathcal{E}_2^z)$ is trivial. Since $\tilde{C}$ itself has $k$ nonzero columns, the subsampled matrix $\tilde{C}_s$ can have at most $k$ nonzero columns too. Thus, $\Pr(\mathcal{E}_2^z) = 0$.

Finally, we can obtain bounds on the largest and smallest singular values of $V^*S$ using the Matrix Chernoff inequalities of [61]. Namely, letting $Z = V^*S$ we note that the matrix $Z Z^*$ may be expressed as a sum of independent positive semidefinite rank-one $r \times r$ Hermitian matrices, as $Z Z^* = V^*S S^*V = \sum_{i=1}^{n_2} S_i (V_i^*)^2 (V_i^*)^*$, where the $\{S_i\}_{i=1}^{n_2}$ are i.i.d. Bernoulli$(\gamma)$ random variables as in the statement of Algorithm 1 (and, $S_i^2 = S_i$). To instantiate the result of [61], we note that $\max_i (S_i (V_i^*)^2 (V_i^*)^*) \leq \|V_i\|_2^2 \leq \mu_i r/n_L \overset{\Delta}{=} R$ almost surely for all $i$. Further, direct calculation yields $\mu_{\min} \overset{\Delta}{=} \min_i (\mathbb{E}[Z Z^*]) = \gamma$ and $\mu_{\max} \overset{\Delta}{=} \max_i (\mathbb{E}[Z Z^*]) = \gamma I$, where the identity matrices in each case are of size $r \times r$. Thus, applying [61, Corollary 5.2] (with $\delta = 1/2$ in that formulation) we obtain that $\Pr(\mathcal{E}_2^z) = \Pr\left(\sigma^2_2 (V^*S) \geq 3\gamma/2\right) \leq r \cdot (9/10)^{n_2/k}$, and $\Pr(\mathcal{E}_2^z) = \Pr\left(2 (V^*S) \leq \gamma/2\right) \leq r \cdot (9/10)^{n_2/k}$.

Putting the results together, and using a further bound on $\Pr(\mathcal{E}_2^z)$, we have $\Pr\left(\left\{\bigcup_{i=1}^{n_2} \mathcal{E}_i^z\right\}\right) \leq \exp\left(-\gamma n_2/200\right) + 2 \exp\left(-\gamma n_2/24\right) + 2 \exp\left(-\gamma n_2/18\right) + r \cdot (9/10)^{n_2/k}$, which is no larger than $\delta$ given that $\gamma$ satisfies (6).

C. Proof of Lemma III.3

First, note that since $\tilde{E}_1 = \tilde{E}_1$, we have that $\|P_{\tilde{E}_1} \tilde{M}_{i,:}\|_2 > 0$ for all $i \in \mathcal{I}_C$, and $\|P_{\tilde{E}_1} \tilde{M}_{i,:}\|_2 = 0$ otherwise. This, along with the fact that the entries of $\phi$ be i.i.d. realizations of a continuous random variable, imply that with probability one the $1 \times n_2$ column vector $\phi \tilde{P}_{\tilde{E}_1} \tilde{M}_{i,:}$ is nonzero at every $i \in \mathcal{I}_C$ and zero otherwise. Indeed, since for each $i \in \mathcal{I}_C$ the distribution of $x_i = \phi \tilde{P}_{\tilde{E}_1} \tilde{M}_{i,:}$ is a continuous random variable with nonzero variance, it takes the value zero with probability zero. On the other hand, for $j \notin \mathcal{I}_C$, $x_j = \phi \tilde{P}_{\tilde{E}_1} \tilde{M}_{i,:} = 0$ with probability one.

With this, we see that exact identification of $\mathcal{I}_C$ can be accomplished if we can identify the support of $x$ from linear measurements of the form $y = (y(2))^T = Ax$.

To proceed, we appeal to (now, well-known) results from the compressive sensing literature. We recall one representative result of [18] that is germane to our effort below. Here, we cast the result in the context of the stable embedding formalism introduced above, and state it as a lemma without proof.

Lemma A.5 (Adapted from Theorem 1.2 of [18]). Let $x \in \mathbb{R}^n$ and $z = Ax$. If $A$ is an $\epsilon$-stable embedding of $\{U_{\epsilon_2k}, \{0\}\}$ for some $\epsilon < \sqrt{2} - 1$ where $U_{\epsilon_2k}$ denotes the union of all $\epsilon_2k$ unique $2k$-dimensional linear subspaces of $\mathbb{R}^n$, and $x$ has at most $s$ nonzero elements, then the solution $\tilde{x}$ of

$$\arg\min_{x} \|x\|_1 \quad s.t. \ z = Ax.$$ (22)

is equal to $x$.

Now, a straightforward application of Lemma A.3 above provides that for any $\delta \in (0, 1)$, if

$$p \geq \frac{2k \log(42/\epsilon) + \log \left(\frac{n}{\epsilon}\right) + \log(2/\delta)}{f(\epsilon/\sqrt{2})}$$ (23)

then the randomly generated $p \times n_2$ matrix $A$ will be an $\epsilon$-stable embedding of $\{U_{\epsilon_2k}, \{0\}\}$ with probability at least $1 - \delta$.

This, along with the well-known bound $\binom{n}{k} \leq \left(\frac{e}{2}\right)^{k}$ and some straightforward simplifications, imply that the condition that $p$ satisfy (8) is sufficient to ensure that with probability at least $1 - \delta$, $A$ is a $(\sqrt{2}/4)$-stable embedding of $\{U_{\epsilon_2k}, \{0\}\}$. Since $\sqrt{2}/4 < \sqrt{2} - 1$, the result follows.

D. An Upper Tail Bound for the Hypergeometric Distribution

Let $\text{hyp}(N,M,n)$ denote the hypergeometric distribution parameterized by a population of size $N$ with $M$ positive elements and $n$ draws, so $H \sim \text{hyp}(N,M,n)$ is a random variable whose value corresponds to the number of positive elements acquired from $n$ draws (without replacement). The probability mass function of $H \sim \text{hyp}(N,M,n)$ is $\Pr(H = k) = \binom{N}{k} \binom{M-N}{n-k}$ for $k \in \{\max\{0, n - M - N\}, \ldots, \min\{M, n\}\}$, and its mean value is $E[H] = nM/N$.

It is well-known that the tails of the hypergeometric distribution are similar to those of the binomial distribution for $n$ trials and success probability $p = M/N$. For example, [62] established that for all $t \geq 0$, $\Pr(H - np \geq nt) \leq e^{-2t^2n}$, a result that follows directly from Hoeffding’s work [63], and exhibits the same tail behavior as predicted by the Hoeffding Inequality for a Binomial$(n,p)$ random variable (see, e.g., [60]). Below we provide a lemma that yields tighter bounds on the upper tail of $H$ when the fraction of positive elements in the population is near 0 or 1. Our result is somewhat analogous to [50, Theorem 2.3(b)] for the Binomial case.

Lemma A.6. Let $H \sim \text{hyp}(N,M,n)$, and set $p = M/N$. For any $\epsilon \geq 0$,

$$\Pr(H \geq (1 + \epsilon)np) \leq e^{-\frac{\epsilon^2np}{M+N}}.$$ (24)

Proof: We begin with an intermediate result of [62], that for any $t \geq 0$ and $h \geq 1$,

$$\Pr(H - np \geq nt) \leq \left(h^{-p+t}(1-p+hp)\right)^n.$$ (25)

Now, for the specific choices $t = cp$ and $h = 1 + \epsilon$ we have

$$\Pr(H - np \geq \epsilon np) \leq \left((1+\epsilon)^{-1+(1+\epsilon)p}(1+cp)\right)^n \leq \left((1+\epsilon)^{-1+(1+\epsilon)e^n}\right)^n \leq e^{-\frac{\epsilon^2np}{M+N}}.$$ (26)
E. Accelerated ADMM for Outlier Pursuit

In this section, we provide a discussion of the algorithmic approach we employ for solving the Outlier Pursuit optimization in Step 1 of Algorithm 1, as well as for implementing the “full data” Outlier Pursuit [17] and Robust Matrix Completion [40] methods in our experimental evaluation. Without loss of generality, we consider an optimization of the form

$$\min_{L,C} \| L \|_* + \lambda \| C \|_{1,2}, \text{ subject to } P_\Omega(M) = P_\Omega(L + C)$$

(27)

where $\lambda > 0$ and $P_\Omega$ is a linear operator that extracts samples at locations indexed by elements of $\Omega$. By the augmented Lagrangian method, the unconstrained form is

$$\min_{L,C} \| L \|_* + \lambda \| C \|_{1,2}$$

$$+ \eta \| U, P_\Omega(M - L - C) \| + \frac{\eta}{2} \| P_\Omega(M - L - C) \| F^2,$$

(28)

where $U$ is the scaled Lagrange multiplier and $\eta > 0$, which is typically set as 1. More details about choices of $\eta$ are discussed in [64]. We propose an accelerated alternating direction method of multipliers (ADMM) approach method inspired by [54], summarized as Algorithm 2.

We discuss our approaches to update primal variables in a bit more detail. If $\Omega = [n_1] \times [n_2]$, the update for $L^{(t)}$ is a proximal operation which has a closed form solution via a soft-thresholding operation of singular values [65]. More specifically, let $L^{(t)} = M + \tilde{U}^{(t)} - \hat{C}^{(t)}$ and $L^{(t)} = \hat{U} \Sigma V^T$ be the compact SVD of $\hat{L}$, then we have

$$L^{(t)} = \hat{U} \hat{D} \{ \Sigma \} V^T,$$

(29)

where $\hat{D}_i(\Sigma) = \max(\max(\sigma_i - \tau, 0))$ for any $\tau > 0$ and $\sigma_i$ is the $i$-th singular value. The update for $C^{(t)}$ also has a closed form solution when $\Omega = [n_1] \times [n_2]$. Let $C^{(t)} = M + \tilde{U}^{(t)} - L^{(t)}$, then for each column $j \in [n_2]$, we have

$$C^{(t)}_{\cdot,j} = G \left( \hat{C}^{(t)}_{\cdot,j} + \lambda \right),$$

(30)

where $G(v, \tau) = \max(v - \tau, 0)$ for a given vector $v \in \mathbb{R}^{n_1}$ and any $\tau > 0$.

If $\Omega \subset [n_1] \times [n_2]$, we linearize the quadratic term and estimate the primal variables approximately. For the first update, define $F(L) = \frac{\lambda}{2} \| P_\Omega(L - \tilde{L}^{(t+1)}) \|^2_F$ and let $Q_L(L) = \frac{\lambda}{2} \| P_\Omega(L^{(t)} - L) \|^2_F + \frac{\lambda}{2} \| L - L^{(t)} \|^2_F + \eta \| P_\Omega(L^{(t)} - L) \|^2_F$, $L - L^{(t)}$ be the quadratic approximation of $F(L)$ at $L^{(t)}$. Here, $\ell_L(L)$ is the Lipschitz constant of $\nabla F(L) = \eta P_\Omega(L^{(t)} - L^{(t)})$, which is no smaller than $\eta$. Then $L^{(t)}$ is approximated by

$$L^{(t)} = \arg\min_{L} \| L \|_*$$

$$\frac{\lambda}{2} \| L - L^{(t)} - \tilde{L}^{(t)} \|^2_F.$$

(31)

Let $L^{(t)} = L^{(t)} - \frac{\lambda}{\ell_L(L^{(t)})} P_\Omega(L^{(t)} - L^{(t)})$ with the compact SVD $L^{(t)} = \hat{U} \Sigma V^T$, then the proximal operation (31) has a closed form solution $L^{(t)} = \hat{U} \hat{D} \{ \Sigma \} V^T$ as in (29).

Algorithm 2 Accelerated ADMM with Restart for Outlier Pursuit

Input: $P_\Omega(M)$, $\lambda$, $\eta > 0$

Initialize: $L^{(-1)} = C^{(0)} = U^{(-1)} = 0^{n_1 \times n_2}$, $\theta^{(0)} = 1$, $\ell^{(0)} = 2\eta$, $t = 1$, $\rho = 0.999$, $c^{(1)} > \varepsilon > 0$

while $c^{(t)} > \varepsilon$ do

$$L^{(t)} = \arg\min_{L} \| L \|_* + \frac{\lambda}{2} \| P_\Omega(L - (M + \tilde{U}^{(t)} - \hat{C}^{(t)})) \|^2_F$$

$$C^{(t)} = \arg\min_{C} \| C \|_{1,2} + \frac{\eta}{2} \| P_\Omega(C - (M + \tilde{U}^{(t)} - \hat{C}^{(t)})) \|^2_F$$

$$U^{(t+1)} = \tilde{U}^{(t)} + P_\Omega(M - L^{(t)} - C^{(t)})$$

if $c^{(t)} < \rho c^{(t-1)}$ then

$$\theta^{(t+1)} = \frac{1 + \sqrt{1 + 4(\theta^{(t)})^2}}{2}$$

$$\hat{C}^{(t+1)} = C^{(t)} + \frac{\theta^{(t+1)} - 1}{\theta^{(t)}} (C^{(t)} - C^{(t-1)})$$

$$\tilde{U}^{(t+1)} = U^{(t)} + \frac{\theta^{(t+1)} - 1}{\theta^{(t)}} (U^{(t)} - U^{(t-1)})$$

else

$$\theta^{(t+1)} = 1$$

$$\tilde{C}^{(t+1)} = \hat{C}^{(t-1)}$$

$$\tilde{U}^{(t+1)} = \hat{U}^{(t-1)}$$

$$c^{(t)} = c^{(t-1)} / \rho$$

end if

end while

Output: $L^{(t)}$ and $C^{(t)}$

In practice, $\ell_L(L^{(t)})$ can be fixed, or obtained via backtracking line-search to further boost the computational efficiency. For the latter, we start with a large $\ell_L(L^{(t)})$ and choose the minimum nonnegative integer $i$ such that $Q_{\alpha^i L^{(t)}}(L^{(t-1)}) > F(L^{(t)})$, where $\alpha \in (0, 1)$ is a shrinkage parameter.

To update $C^{(t)}$ when $\Omega \subset [n_1] \times [n_2]$, we apply an analogous linearization approach as before with either fixed $\ell_C(L^{(t)})$ or backtracking line-search to dynamically update $L^{(t)}$. Then, a closed form solution can be obtained for each column $j \in [n_2]$ of $C^{(t)}$ as in (30), as $C^{(t)}_{\cdot,j} = G \left( \hat{C}^{(t)}_{\cdot,j} + \lambda \right)$.

REFERENCES

[1] K. Pearson, “On lines and planes of closest fit to systems of points in space,” The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, pp. 559–572, 1901.

[2] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” Journal of educational psychology, vol. 24, no. 6, pp. 417, 1933.

[3] I. Jolliffe, Principal component analysis, Wiley Online Library, 2005.

[4] N. Halko, P.-G. Martinsson, and J. A. Tropp, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions,” SIAM review, vol. 53, no. 2, pp. 217–288, 2011.

[5] M. W. Mahoney, “Randomized algorithms for matrices and data,” Foundations and Trends® in Machine Learning, vol. 3, no. 2, pp. 123–224, 2011.

[6] E. J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal recovery from highly incomplete frequency information,” IEEE Trans. Inform. Theory, vol. 52, no. 2, pp. 489–509, 2006.

[7] D. Donoho, “Compressed sensing,” IEEE Trans. Inform. Theory, vol. 52, no. 4, pp. 1289–1306, 2006.

[8] E. J. Candès and T. Tao, “Near optimal signal recovery from random projections: Universal encoding strategies?,” IEEE Trans. Inform. Theory, vol. 52, no. 12, pp. 5406–5425, 2006.

[9] C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the underlying neural circuity,” in Matters of Intelligence, pp. 115–141. Springer, 1987.

[10] J. Totsios, S. Calhane, W. Wai, Y. Lai, N. Davis, and F. Nufio, “Modeling visual attention via selective tuning,” Artificial Intelligence, vol. 78, no. 1-2, 1995.
[10] P. Indyk, E. Price, and D. P. Woodruff, “On the power of adaptivity in sparse recovery,” in Proc. IEEE Foundations of Computer Science, 2011, pp. 285–294.
[11] M. Iwen and A. Tewﬁk, “Adaptive group testing strategies for target detection and localization in noisy environments,” IEEE Trans. Signal Proc., vol. 60, no. 5, pp. 2344–2353, 2012.
[12] M. L. Malloy and R. Nowak, “Sequential testing for sparse recovery,” arXiv preprint arXiv:1212.1801, 2012.
[13] S. Balakrishnan, M. Kolar, A. Rinaldo, and A. Singh, “Recovering block-structured activities using compressive measurements,” arXiv preprint arXiv:1206.3431, 2012.
[14] R. M. Castro, “Adaptive sensing performance lower bounds for sparse signal detection and support estimation,” arXiv preprint arXiv:1206.0468, 2012.
[15] E. Price and D. P. Woodruff, “Lower bounds for adaptive sparse recovery,” arXiv preprint arXiv:1205.3518, 2012.
[16] M. Malloy and R. Nowak, “Near-optimal adaptive compressive sensing,” in Proc. Asilomar Conf. on Signals, Systems, and Computers, 2012.
[17] M. A. Davenport and E. Arias-Castro, “Compressive binary search,” in Proc. IEEE Intl. Symp on Information Theory, 2012, pp. 1827–1831.
[18] A. Krishnamurthy, J. Sharpnack, and A. Singh, “Recovering graph-structured activities using adaptive compressive measurements,” arXiv preprint arXiv:1305.0213, 2013.
[19] T. Hastie, R. Tibshirani, and J. E. J. Candès, and M. A. Davenport, “On the fundamental limits of adaptive sensing,” IEEE Transactions on Information Theory, vol. 59, no. 1, pp. 472–481, 2013.
[20] D. Wei and A. O. Hero, “Multistage adaptive estimation of sparse signals,” IEEE Journal of Selected Topics in Signal Processing, vol. 7, no. 5, pp. 783–796, 2013.
[21] A. Krishnamurthy and A. Singh, “Low-rank matrix and tensor completion via adaptive sensing,” arXiv preprint arXiv:1304.4672, 2013.
[22] A. Soni and J. Haupt, “On the fundamental limits of recovering tree sparse vectors from noisy linear measurements,” IEEE Trans. Information Theory, vol. 60, no. 1, pp. 133–149, 2014.
[23] J. Haupt and R. Nowak, “Adaptive sensing for sparse recovery,” in Compressed Sensing: Theory and applications, Y. Eldar and G. Kutyniok, Eds. Cambridge University Press, 2011.
[24] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, “Rank-sparse incoherence for matrix decomposition,” SIAM J. Optimization, vol. 21, no. 2, 2011.