RECURSIVE ROBUST PCA OR RECURSIVE SPARSE RECOVERY IN LARGE BUT STRUCTURED NOISE

Chenlu Qiu, Namrata Vaswani, Leslie Hogben

ECE dept, Iowa State University, Ames IA

ABSTRACT

We study the recursive robust principal components' analysis (PCA) problem. Here, "robust" refers to robustness to both independent and correlated sparse outliers. If the outlier is the signal-of-interest, this problem can be interpreted as one of recursively recovering a time sequence of sparse vectors, S_t, in the presence of large but structured noise, L_t: the noise needs to lie in a "slowly changing" low dimensional subspace. We study a novel solution called Recursive Projected CS (ReProCS). Under mild assumptions, we show that, with high probability (w.h.p.), at all times, ReProCS can exactly recover the support set of S_t; and the reconstruction errors of both S_t and L_t are upper bounded by a time-invariant and small value.

***Index Terms*—** robust PCA, compressive sensing

1. INTRODUCTION

This work studies the recursive robust principal components' analysis (PCA) problem. A key application where this occurs is in video analysis where the goal is to separate a slowly changing background from moving foreground objects [1, 2]. If we stack each frame as a column vector, the background is well modeled as lying in a low dimensional subspace that may gradually change over time, while the moving foreground objects constitute the sparse outliers [3, 2] which change in a correlated fashion over time. Other applications include sensor networks based detection and tracking of abnormal events such as forest fires or oil spills; or online detection of brain activation patterns from functional MRI (fMRI) sequences (the "active" part of the brain can be interpreted as a correlated sparse outlier). In many of these applications, an online solution is desirable. In this work, we focus on this case, i.e. on *recursive robust PCA that is robust to both independent and correlated sparse outliers*.

The moving objects or the brain active regions or the oil spill region may be "outliers" for the PCA problem, but in most cases, these are actually the signals-of-interest whereas the background image is the noise. Also, all the above signals-of-interest are sparse vectors that change in a correlated fashion over time. Thus, this problem can also be interpreted as one of recursively recovering a time sequence of correlated sparse signals, S_t, from measurements $M_t := S_t + L_t$ that are corrupted by (potentially) large magnitude but dense and structured noise, L_t. The structure that we require is that L_t be dense and lie in a low dimensional subspace that is either fixed or changes "slowly enough" in the sense quantified in Sec 2.1.

Related Work. There has been a large amount of work on robust PCA, e.g. [1, 2, 4, 5, 6, 7, 8], and recursive robust PCA e.g. [9, 10, 11]. These works either assume that the locations of the missing/corruped data points are assumed known (not a practical assumption); or throw out the entire outlier vector [7, 8] (this is a problem if most frames contain outliers) or cannot detect small magnitude outliers [10, 1, 11], S_t, (this is needed when S_t is the signal of interest). Moreover, except [7, 8], the others do not come with any performance guarantees either. In recent work [2, 4], a new elegant solution to robust PCA called Principal Components' Pursuit (PCP) has been proposed, that removes the above limitations. It redefines batch robust PCA as a problem of separating a low rank matrix, $\mathcal{L}_t := [L_1, \dots, L_t]$, from a sparse matrix, $\mathcal{S}_t := [S_1, \dots, S_t]$, using the measurement matrix, $\mathcal{M}_t := [M_1, \dots, M_t] = \mathcal{L}_t + \mathcal{S}_t$. It was shown in [2] that one can recover $\mathcal{L}_t$ and $\mathcal{S}_t$ exactly by solving $\min_{\mathcal{L},\mathcal{S}} \|\mathcal{L}\|_* + \lambda \|\mathcal{S}\|_1$ subject to $\mathcal{L} + \mathcal{S} = \mathcal{M}_t$ provided that (a) $\mathcal{L}_t$ is dense (its left and right singular vectors satisfy certain conditions); (b) any element of the matrix $\mathcal{S}_t$ is nonzero w.p. ϱ, and zero w.p. $1 - \varrho$, independent of all others (in particular, this means that the support sets of the different S_t's are independent over time); and (c) the rank of $\mathcal{L}_t$ and the support size of $\mathcal{S}_t$ are small enough. Here $\|A\|_*$ is the nuclear norm of A (sum of singular values of A) while $\|A\|_1$ is the ℓ_1 norm of A seen as a long vector.

In many practical applications, e.g. video analysis, it is fair to assume that the background changes are dense (i.e. $\mathcal{L}_t$ is dense). However, the assumption that the foreground support is independent over time is not a valid one. Foreground objects typically move in a correlated fashion, and may even not move for a few frames. This often results in $\mathcal{S}_t$ being sparse as well as low rank. In the case where $\mathcal{L}_t$ is low rank and dense, but $\mathcal{S}_t$ is both sparse and low rank, in general, PCP [2, 4] will not work. Without any extra information, it is also not clear how else to separate $\mathcal{S}_t$ and $\mathcal{L}_t$. But suppose that an initial short sequence of L_t's is available. For example, in the video application, it is often realistic to assume that an initial background-only training sequence is available. The question is, can we use this to do anything better?

Contribution. In [12, 13], we first studied this problem and proposed a novel solution called Recursive Projected CS (ReProCS). In this work we develop a modification of the algorithm of [12] that can be analyzed more easily. The key contribution of this work is as follows. Under mild assumptions, we show that, w.h.p, ReProCS can exactly recover the support set of S_t at all times; and the reconstruction errors of both S_t and L_t are upper bounded by a time invariant and small value at all times. If L_t is the signal of interest, then ReProCS is a solution to recursive robust PCA in the presence of sparse and possibly correlated outliers. To the best of our knowledge, this is the first rigorous analysis of any recursive (online) robust PCA approach and definitely the first to study recursive (online) robust PCA with correlated outliers. Ours is also among the first few works that studies recursive sparse recovery in (potentially) large but structured noise: the noise needs to lie in a "slowly changing" low dimensional subspace as defined in Sec 2.1. Works that study a related problem of sparse signal recovery from large but sparse noise (outlier) include [3, 14, 15]. Since these algorithms are designed for a single

This work was partly supported by NSF grant CCF-1117125. Qiu and Vaswani are with the Electrical and Computer Engineering dept, L. Hogben is with the Mathematics dept at Iowa State University.

signal (without using past or future information), these can also be interpreted as solutions for recursive sparse recovery from large but sparse noise.

Notation. For a set $T \subset \{1, 2, \dots n\}$, we use $|T|$ to denote its cardinality, i.e., the number of elements in T. For a vector v, v_i denotes the ith entry of v and v_T denotes a vector consisting of the entries of v indexed by T. We use $\|v\|_p$ to denote the ℓ_p norm of v.

For a matrix B, B' denotes its transpose, and $B^\dagger$ its Moore-Penroe pseudo-inverse. We use $\|B\|_2 := \max_{x \neq 0} \|Bx\|_2/\|x\|_2$ to denote the induced 2-norm of the matrix. For a Hermitian matrix, B, we use the notation $B \overset{EVD}{=} U \Lambda U'$ to denote the eigenvalue decomposition of B. Here U is an orthonormal matrix and Λ is a diagonal matrix with entries arranged in non-increasing order. We use I to denote an identity matrix. For an index set T and a matrix B, B_T is the sub-matrix of B containing columns with indices in the set T. For a tall matrix P, span(P) denotes the subspace spanned by the column vectors of P. The notation $[.]$ denotes an empty matrix.

Definition 1 *We refer to a matrix P as a* basis matrix *if $P'P = I$.*

The *s-restricted isometry constant (RIC)* [16], δ_s, for an $n \times m$ matrix Ψ is the smallest real number satisfying $(1 - \delta_s)\|x\|_2^2 \le \|\Psi_T x\|_2^2 \le (1+\delta_s)\|x\|_2^2$ for all sets $T \subseteq \{1, 2, \dots n\}$ with $|T| \le s$ and all real vectors x of length $|T|$.

2. PROBLEM FORMULATION

The measurement vector at time t, M_t, is an n dimensional vector which can be decomposed as

$$M_t = L_t + S_t \tag{1}$$

Here S_t is a sparse vector with support set size at most s and minimum magnitude of nonzero values at least $S_{\min}$. L_t is a dense but low dimensional vector that satisfies the model given below. We are given an accurate estimate of the subspace in which the initial t_{train} L_t's lie, i.e. we are given a basis matrix $\hat{P}_0$ so that $\|(I - \hat{P}_0\hat{P}_0')P_0\|_2$ is small. Here P_0 is a basis matrix for span$(\mathcal{L}_{t_{\text{train}}})$, i.e. span$(P_0) =$ span$(\mathcal{L}_{t_{\text{train}}})$. The goal is

1. to estimate both S_t and L_t at each time $t > t_{\text{train}}$, and
2. to estimate span$(\mathcal{L}_t)$ every so often.

Notation for S_t. Let $T_t := \{i : (S_t)_i \neq 0\}$ denote the support of S_t. Define

$$S_{\min} := \min_t \min_{i \in T_t} |(S_t)_i|, \text{ and } s := \max_t |T_t|$$

Model on L_t. The L_t's satisfy the following model.

1. L_t lies in a low dimensional subspace that changes every-so-often. Let t_j denote the change times. Then $L_t = P_{(t)} a_t$ with $P_{(t)} = P_j$ for all $t_j \le t < t_{j+1}$, $j = 0, 1, 2 \cdots J$, i.e. there is a maximum of J subspace change times. We can define $t_{J+1} = \infty$. Here P_j is an $n \times r_j$ basis matrix with $r_j \ll n$ and $r_j \ll (t_{j+1} - t_j)$.
2. At the change times, t_j, P_j changes as $P_j = [P_{j-1} \ P_{j,\text{new}}]$ where $P_{j,\text{new}}$ is a $n \times c_{j,\text{new}}$ basis matrix with $P_{j,\text{new}}' P_{j-1} = 0$. Thus $r_j = r_{j-1} + c_{j,\text{new}}$.
3. There exists a constant c_{mx} such that $0 \le c_{j,\text{new}} \le c_{mx}$.
4. The projection vector, $a_t := {P_{(t)}}' L_t$, is a random variable (r.v.) with the following properties. (a) a_t's are mutually independent over time, t. (b) It is a zero mean bounded r.v., i.e. $\mathbf{E}(a_t) = 0$ and there exists a constant γ_* s.t. $\|a_t\|_\infty \le \gamma_*$ for all t. (c) Its covariance matrix $\Lambda_t :=$ Cov$[a_t] = \mathbf{E}(a_t a_t')$ is diagonal with $\lambda^- := \min_t \lambda_{\min}(\Lambda_t) > 0$ and $\lambda^+ := \max_t \lambda_{\max}(\Lambda_t) < \infty$. Thus the condition number of any Λ_t is bounded by $f := \frac{\lambda^+}{\lambda^-}$.

Moreover, P_j and a_t change slowly as quantified in Sec 2.1. Also, the L_t's, and hence their subspace basis matrices P_j, are dense, i.e. the denseness coefficient $\kappa_s(P_j)$, which is defined in Sec 2.2, is small for all j.

2.1. Slow subspace change

By slow subspace change we mean all the following. First, the delay between consecutive subspace change times, $t_{j+1} - t_j$, is large enough.

Second, the projection of L_t along the newly added directions, $a_{t,\text{new}}$, is initially small, i.e. $\max_{t_j \le t < t_j + \alpha} \|a_{t,\text{new}}\|_\infty \le \gamma_{\text{new}}$, with $\gamma_{\text{new}} \ll \gamma_*$ and $\gamma_{\text{new}} \ll S_{\min}$, but can increase gradually. We model this as follows. Split the interval $[t_j, t_{j+1} - 1]$ into α length periods. We assume that

$$\max_j \max_{t \in [t_j + (k-1)\alpha, t_j + k\alpha - 1]} \|a_{t,\text{new}}\|_\infty \le \gamma_{\text{new},k} := \min(v^{k-1}\gamma_{\text{new}}, \gamma_*)$$

for a $v > 1$ but not too large. This assumption is verified for real video data in [17, Sec X-A].

Third, the number of newly added directions is small, i.e. $c_{j,\text{new}} \le c_{mx} \ll r_0$. This is also verified in [17, Sec X-A].

2.2. Measuring denseness of a matrix and its relation with RIC

For a tall $n \times r$ matrix, B, or for a $n \times 1$ vector, B, we define the the denseness coefficient as follows:

$$\kappa_s(B) := \max_{|T| \le s} \frac{\|{I_T}' B\|_2}{\|B\|_2}.$$

where $\|.\|_2$ is the matrix or vector 2-norm respectively. As we explain in Sec 5, $\kappa_s(B)$ is related to the denseness assumptions required by PCP [2].

The lemma below relates the denseness coefficient of a basis matrix P to the RIC of $I - PP'$. The proof is in [17, Appendix].

Lemma 2 *For an $n \times r$ basis matrix P (i.e P satisfying $P'P = I$),*

$$\delta_s(I - PP') = \kappa_s^2(P).$$

3. RECURSIVE PROJECTED CS (REPROCS)

We summarize the Recursive Projected CS (ReProCS) algorithm in Algorithm 1. It uses the following definition.

Definition 3 *Define the time interval $\mathcal{I}_{j,k} := [t_j + (k-1)\alpha, t_j + k\alpha - 1]$ for $k = 1, \dots K$ and $\mathcal{I}_{j,K+1} := [t_j + K\alpha, t_{j+1} - 1]$. Here, K is the algorithm parameter in Algorithm 1.*

The key idea of ReProCS is as follows. Assume that the current basis matrix $P_{(t)}$ has been accurately predicted using past estimates of L_t, i.e. we have $\hat{P}_{(t-1)}$ with $\|(I - \hat{P}_{(t-1)}\hat{P}_{(t-1)}')P_{(t)}\|_2$ small. We project M_t into the space perpendicular to $\hat{P}_{(t-1)}$ to get the projected measurement vector $y_t := \Phi_{(t)} M_t$ where $\Phi_{(t)} = I - \hat{P}_{(t-1)}\hat{P}_{(t-1)}'$ (step 1a). Since the $n \times n$ projection matrix, $\Phi_{(t)}$ has rank $n - r_*$ where $r_* =$ rank$(\hat{P}_{(t-1)})$, therefore y_t has only $n - r_*$ "effective" measurements[1], even though its length is

[1] i.e. some r_* entries of y_t are linear combinations of the other $n - r_*$ entries

Algorithm 1 Recursive Projected CS (ReProCS)

```
Parameters: algorithm parameters: $\xi$, $\omega$, $\alpha$, $K$, model parameters: $t_j$, $r_0$, $c_{j,\text{new}}$ (set as in Theorem 4 or as in [17, Sec X-B] when the model is not known)
Input: $M_t$, Output: $\hat{S}_t$, $\hat{L}_t$, $\hat{P}_{(t)}$
Initialization: Given training sequence $[L_1, L_2, \cdots, L_{t_{\text{train}}}]$, estimate $\hat{P}_0$ by computing an EVD as $\frac{1}{t_{\text{train}}} \sum_{t=1}^{t_{\text{train}}} L_t {L_t}' \overset{EVD}{=} E \Lambda E'$ and then retaining the eigenvectors with the $r_0$ largest eigenvalues, i.e., $\hat{P}_0 \leftarrow (E)_{\{1,2,\cdots,r_0\}}$.
Let $\hat{P}_{(t)} \leftarrow \hat{P}_0$. Let $j \leftarrow 1$, $k \leftarrow 1$. For $t > t_{\text{train}}$, do the following:
  1. Estimate $T_t$ and $S_t$ via Projected CS:
     (a) Nullify most of $L_t$: compute $\Phi_{(t)} \leftarrow I - \hat{P}_{(t-1)} \hat{P}'_{(t-1)}$, compute $y_t \leftarrow \Phi_{(t)} M_t$
     (b) Sparse Recovery: compute $\hat{S}_{t,\text{cs}}$ as the solution of $\min_x \|x\|_1 \ s.t. \ \|y_t - \Phi_{(t)} x\|_2 \leq \xi$
     (c) Support Estimate: compute $\hat{T}_t = \{i : \ |(\hat{S}_{t,\text{cs}})_i| > \omega\}$
     (d) LS Estimate of $S_t$: compute $(\hat{S}_t)_{\hat{T}_t} = ((\Phi_t)_{\hat{T}_t})^{\dagger} y_t, \ (\hat{S}_t)_{\hat{T}_t^c} = 0$
  2. Estimate $L_t$: $\hat{L}_t = M_t - \hat{S}_t$.
  3. Update $\hat{P}_{(t)}$ by Projection PCA
     (a) If $t = t_j + k\alpha - 1$,
          i. compute $\frac{1}{\alpha} \sum_{t \in \mathcal{I}_{j,k}} (I - \hat{P}_{j-1} \hat{P}'_{j-1}) \hat{L}_t \hat{L}'_t (I - \hat{P}_{j-1} \hat{P}'_{j-1}) \overset{EVD}{=} \begin{bmatrix} \hat{P}_{j,\text{new},k} \ \hat{P}_{j,\text{new},k,\perp} \end{bmatrix} \begin{bmatrix} \Lambda_k & 0 \\ 0 & \Lambda_{k,\perp} \end{bmatrix} \begin{bmatrix} \hat{P}'_{j,\text{new},k} \\ \hat{P}'_{j,\text{new},k,\perp} \end{bmatrix}$
             where $\Lambda_k$ is of size $c_{j,\text{new}} \times c_{j,\text{new}}$.
         ii. set $\hat{P}_{(t)} \leftarrow [\hat{P}_{j-1} \ \hat{P}_{j,\text{new},k}]$; increment $k \leftarrow k+1$.
         Else
          i. set $\hat{P}_{(t)} \leftarrow \hat{P}_{(t-1)}$.
     (b) If $t = t_j + K\alpha - 1$, then set $\hat{P}_j \leftarrow [\hat{P}_{j-1} \ \hat{P}_{j,\text{new},K}]$. Increment $j \leftarrow j+1$. Reset $k \leftarrow 1$.
  4. Increment $t \leftarrow t+1$ and go to step 1.
```

n. Notice that y_t can be rewritten as $y_t = \Phi_{(t)} S_t + \beta_t$ where $\beta_t := \Phi_{(t)} L_t$. Since $\|(I - \hat{P}_{(t-1)} \hat{P}'_{(t-1)}) P_{(t)}\|_2$ is small, the projection nullifies most of the contribution of L_t and so the projected noise β_t is small. Recovering the n dimensional sparse vector S_t from y_t now becomes a traditional sparse recovery or CS problem in small noise [18, 19, 20, 21] (step 1b). If $P_{(t)}$, and hence its estimate, $\hat{P}_{(t-1)}$, is dense enough, then, by Lemma 2, the RIC of $\Phi_{(t)}$ is small enough. By [21, Thm 1], this ensures that S_t can be accurately recovered from y_t. By thresholding on the recovered S_t, one gets an estimate of its support (step 1c). By computing a least squares (LS) estimate of S_t on the estimated support and setting it to zero everywhere else (step 1d), we can get a more accurate final estimate, $\hat{S}_t$, as first suggested in [22]. This $\hat{S}_t$ is used to estimate L_t as $\hat{L}_t = M_t - \hat{S}_t$ (step 2). The sparse recovery error, $e_t := S_t - \hat{S}_t$. Since $\hat{L}_t = M_t - \hat{S}_t$ (step 2), e_t also satisfies $e_t = L_t - \hat{L}_t$. Thus, a small e_t means that L_t is also recovered accurately. The estimated $\hat{L}_t$'s are used to obtain new estimates of $P_{j,\text{new}}$ every α frames for a total of $K\alpha$ frames via a modification of the standard PCA procedure, which we call projection PCA (step 3).

The ReProCS idea is also somewhat related to that of [16, 23, 24] in that all of these also try to cancel the "low rank" part by projecting the original data vector into the perpendicular space of the tall matrix that spans the "low rank" part. However the big difference is that in all these works, this matrix is *known*. In our problem this matrix $P_{(t)}$ is *unknown and can change with time.*

4. PERFORMANCE GUARANTEES

We state the main result here first and then discuss it. For the proof, see [17, Sections V, VI, VII].

Theorem 4 *Consider Algorithm 1. Let $c := c_{mx}$ and $r := r_0 + (J-1)c$. Assume that L_t obeys the model given in Sec. 2 and there are a total of J change times. Assume also that the initial subspace estimate is accurate enough, i.e. $\|(I - \hat{P}_0 \hat{P}'_0) P_0\| \leq r_0 \zeta$, for a ζ that satisfies*

$$\zeta \leq \min\left(\frac{10^{-4}}{r^2}, \frac{1.5 \times 10^{-4}}{r^2 f}, \frac{1}{r^3 \gamma_*^2}\right) \textit{ where } f := \frac{\lambda^+}{\lambda^-}$$

If the following conditions hold:

1. *the algorithm parameters are set as $\xi = \xi_0(\zeta)$, $7\rho\xi \leq \omega \leq S_{\min} - 7\rho\xi$, $K = K(\zeta)$, $\alpha \geq \alpha_{add}(\zeta)$, where $\xi_0(\zeta), \rho, K(\zeta), \alpha_{add}(\zeta)$ are defined in Definition 5.*
2. *$P_{j-1}, P_{j,new}, D_{j,new,k} := (I - \hat{P}_{j-1} \hat{P}'_{j-1} - \hat{P}_{j,new,k} \hat{P}'_{j,new,k}) P_{j,new}$ and $Q_{j,new,k} := (I - P_{j,new} {P_{j,new}}') \hat{P}_{j,new,k}$ have dense enough columns, i.e.*

$$\kappa_{2s}(P_{J-1}) \leq 0.3, \ \max_j \kappa_{2s}(P_{j,new}) \leq 0.15,$$
$$\max_j \max_{0 \leq k \leq K} \kappa_{2s}(D_{j,new,k}) \leq 0.15,$$
$$\max_j \max_{0 \leq k \leq K} \kappa_{2s}(Q_{j,new,k}) \leq 0.15$$

   with $\hat{P}_{j,new,0} = [.]$ (empty matrix).

3. *for a given value of $S_{\min}$, the subspace change is slow enough, i.e.*

$$\max_j (t_{j+1} - t_j) > K\alpha,$$
$$\max_j \max_{t_j + (k-1)\alpha \leq t < t_j + k\alpha} \|a_{t,new}\|_\infty \leq \min(1.2^{k-1} \gamma_{new}, \gamma_*),$$
$$14 \rho \xi_0(\zeta) \leq S_{\min},$$

4. *the condition number of the covariance matrix of $a_{t,new}$ averaged over $t \in \mathcal{I}_{j,k}$, is bounded, i.e.*

$$g_{j,k} \leq \sqrt{2}$$

   where $g_{j,k}$ is defined in Definition 5,

then, with probability at least $(1 - n^{-10})$, the following hold:

1. *at all times, t, $\hat{T}_t = T_t$ and $\|e_t\|_2 = \|L_t - \hat{L}_t\|_2 = \|\hat{S}_t - S_t\|_2 \leq 0.18\sqrt{c}\gamma_{new} + 1.2\sqrt{\zeta}(\sqrt{r} + 0.06\sqrt{c})$.*
2. *the subspace error $SE_{(t)} := \|(I - \hat{P}_{(t)} \hat{P}'_{(t)}) P_{(t)}\|_2$ satisfies*

$$SE_{(t)} \leq \begin{cases} (r_0 + (j-1)c)\zeta + 0.4c\zeta + 0.6^{k-1} & \textit{if } t \in \mathcal{I}_{j,k}, \\ (r_0 + jc)\zeta & \textit{if } t \in \mathcal{I}_{j,K+1} \end{cases}$$
$$\leq \begin{cases} 10^{-2}\sqrt{\zeta} + 0.6^{k-1} & \textit{if } t \in \mathcal{I}_{j,k}, \\ 10^{-2}\sqrt{\zeta} & \textit{if } t \in \mathcal{I}_{j,K+1} \end{cases}$$

3. *e_t follows a trend similar to that of SE$_{(t)}$ at various times (the bounds are available in [17, Theorem ?].*

Proof: See [17, Sections V, VI, VII].

Definition 5 *We define here the parameters used in Theorem 4.*

1. *Define* $K(\zeta) := \left\lceil \frac{\log(0.6c\zeta)}{\log 0.6} \right\rceil$
2. *Define* $\xi_0(\zeta) := \sqrt{c}\gamma_{new} + \sqrt{\zeta}(\sqrt{r} + \sqrt{c})$
3. *Define* $\rho := \max_t\{\kappa_1(\hat{S}_{t,cs} - S_t)\}$. *Notice that* $\rho \leq 1$.
4. *Let* $K = K(\zeta)$. *Define*

$$\begin{aligned}\alpha_{add} = \lceil & \frac{4608(\log 6KJ + 11\log n)}{\zeta^2(\lambda^-)^2} \\ & \max(\min(1.2^{4K}\gamma_{new}^4, \gamma_*^4), \\ & \frac{16}{c^2}, 4(0.186\gamma_{new}^2 + 0.0034\gamma_{new} + 2.3)^2)\rceil.\end{aligned}$$

   In words, α_{add} is the smallest value of the number of data points, α, needed for one projection PCA step to ensure that Theorem 4 holds w.p. at least $(1 - n^{-10})$.

5. *Define the condition number of Cov$(a_{t,new})$ averaged over $t \in \mathcal{I}_{j,k}$ as $g_{j,k} := \frac{\lambda_{j,new,k}{}^+}{\lambda_{j,new,k}{}^-}$ where*
   $\lambda_{j,new,k}{}^+ := \lambda_{\max}(\frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}}(\Lambda_t)_{new})$, and
   $\lambda_{j,new,k}{}^- := \lambda_{\min}(\frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}}(\Lambda_t)_{new})$.

This result says the following. Assume that the initial subspace error is small enough. If (a) the algorithm parameters are set appropriately; (b) the matrices defining the previous subspace, the newly added subspace, and the currently unestimated part of the newly added subspace are dense enough; (c) the subspace change is slow enough; and (d) the condition number of the average covariance matrix of $a_{t,\text{new}}$ is small enough, then, w.h.p., we will get exact support recovery at all times. Moreover, the sparse recovery error will always be bounded by $0.18\sqrt{c}\gamma_{\text{new}}$ plus a constant times $\sqrt{\zeta}$. Since ζ is very small, $\gamma_{\text{new}} \ll S_{\min}$, and c is also small, the normalized reconstruction error for recovering S_t will be small at all times.

In the second conclusion, we bound the subspace estimation error, SE$_{(t)}$. When a subspace change occurs, this error is initially bounded by one. The above result shows that, w.h.p., with each projection PCA step, this error decays exponentially and falls below $0.01\sqrt{\zeta}$ within K projection PCA steps. The third conclusion shows that, with each projection PCA step, w.h.p., the sparse recovery error as well as the error in recovering L_t also decay in a similar fashion.

Notice that $K = K(\zeta)$ is larger if ζ is smaller. Also, α_{add} is inversely proportional to ζ. Thus, if we want to achieve a smaller lowest error level, ζ, we need to compute projection PCA over larger durations α and we need more number of projection PCA steps K.

5. DISCUSSION AND COMPARISON WITH PCP RESULT

We provide a qualitative comparison with [2]. A direct comparison is not possible since the proof techniques used are very different and since we solve a recursive version of the problem where as PCP solves a batch one. Moreover, PCP provides guarantees for exact recovery of $\mathcal{S}_t$ and $\mathcal{L}_t$. In our result, we obtain guarantees for exact support recovery of the S_t's (and hence of $\mathcal{S}_t$) and bounded error recovery of its nonzero values and of $\mathcal{L}_t$. Also, PCP assumes no model knowledge, whereas our algorithm does assume knowledge of model parameters. Of course, in [17, Sec X-B], we explain how to set the parameters when the model is not known.

The first key difference between our result and that of PCP [2] is as follows. The result for PCP [2] assumes that any element of the $n \times t$ matrix $\mathcal{S}_t$ is nonzero w.p. ϱ, and zero w.p. $1 - \varrho$, independent of all others (in particular, this means that the support sets of the different S_t's are independent over time). This ensures that w.h.p. $\mathcal{S}_t$ is sparse but full rank and hence ensures that it can be separated from $\mathcal{L}_t$ which is low rank but dense. As explained earlier, the assumption of independent support sets of S_t is not valid for real video data where the foreground objects usually move in a highly correlated fashion over time. On the other hand, our result for ReProCS does not put any such assumption on the support sets of the S_t's. The reason it can do this is because it assumes accurate knowledge of the subspace spanned by the first few columns of $\mathcal{L}_t$ and it assumes slow subspace change (verified in [17, Sec X-A]), both of which are practically valid assumptions. However, ReProCS does need denseness of $D_{j,\text{new},k}$, whose columns span the currently unestimated part of span$(P_{j,\text{new}})$. In simulations, we observe that this reduces when the support of S_t changes very infrequently.

Next let us compare the denseness assumptions. Let $\mathcal{L}_t = U\Sigma V'$ be its SVD. Then, for $t \in [t_j, t_{j+1} - 1]$, $U = [P_{j-1}, P_{j,\text{new}}]$ and $V = [a_1, a_2 \dots a_t]'\Sigma^{-1}$. PCP [2] assumes denseness of U and of V: it requires $\kappa_1(U) \leq \sqrt{\mu r/n}$ and $\kappa_1(V) \leq \sqrt{\mu r/n}$ for a constant $\mu \geq 1$. Moreover, it also requires $\|UV'\|_{\max} \leq \sqrt{\mu r}/n$. Here $\|B\|_{\max} := \max_{i,j} |(B)_{i,j}|$. On the other hand, our denseness assumptions are on P_{j-1} and $P_{j,\text{new}}$ which are sub-matrices of U. We do not need denseness of V and we do not bound $\|UV'\|_{\max}$.

However, some additional assumptions that we need are (a) denseness of $D_{j,\text{new},k}$ and of $Q_{j,\text{new},k}$; (b) the independence of a_t's over time and (c) condition number of the average covariance matrix of $a_{t,\text{new}}$, is not too large. (c) is an assumption made for simplicity. As explained in [25], this can be removed and replaced if the newly added eigenvalues can be separated into a few clusters, each with small condition number. (b) is assumed so that we can use the matrix Hoeffding inequality [26, Theorem 1.3] to obtain high probability bounds on the terms in the subspace error bound. In experiments, we are able to also deal with correlated a_t's. As explained in [17], it should be possible to replace it by a milder assumption. Consider (a). Our proof only needs $\|I_{T_t}{}'D_{j,\text{new},k}\|_2/\|D_{j,\text{new},k}\|_2$ to be small at every projection PCA time. We attempted to verify this in simulations done with a dense P_j and $P_{j,\text{new}}$. Except for the case of exactly constant support of S_t, in all other cases (including the case of very gradual support change), this ratio was small for most projection PCA times. We also saw that even if at a few projection PCA times, this ratio was close to one, that just meant that, at those times, the subspace error remained roughly equal to that at the previous time. As a result, a larger K was required for the subspace error to become small enough. It did not mean that the algorithm became unstable. It should be possible to use a similar idea to modify our result as well. An analogous discussion applies to $Q_{j,\text{new},k}$.

Extensive experimental comparisons with other works are available at `http://www.ece.iastate.edu/~chenlu/ReProCS/ReProCS.htm` and will be discussed in forthcoming work.

6. CONCLUSIONS AND FUTURE WORK

We studied the recursive (online) robust PCA problem, which can also be interpreted as a problem of recursive sparse recovery in the presence of large but structured noise. Under mild assumptions, we showed that, w.h.p., ReProCS can exactly recover the support set of S_t at all times; and the reconstruction errors of both S_t and L_t are upper bounded by a time-invariant and small value. In ongoing work [25], we are developing and analyzing ReProCS with deletion.

7. REFERENCES

[1] F. De La Torre and M. J. Black, “A framework for robust subspace learning”, *International Journal of Computer Vision*, vol. 54, pp. 117–142, 2003.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?”, *Journal of ACM*, vol. 58, no. 3, 2011.

[3] J. Wright and Y. Ma, “Dense error correction via l1-minimization”, *IEEE Trans. on Info. Th.*, vol. 56, no. 7, pp. 3540–3560, 2010.

[4] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Rank-sparsity incoherence for matrix decomposition”, *SIAM Journal on Optimization*, vol. 21, 2011.

[5] S. Roweis, “Em algorithms for pca and spca”, *Advances in Neural Information Processing Systems*, pp. 626–632, 1998.

[6] T. Zhang and G. Lerman, “A novel m-estimator for robust pca”, *arXiv:1112.4863v1*, 2011.

[7] H. Xu, C. Caramanis, and S. Sanghavi, “Robust pca via outlier pursuit”, *IEEE Tran. on Information Theorey*, vol. 58, no. 5, 2012.

[8] M. McCoy and J. Tropp, “Two proposals for robust pca using semidefinite programming”, *arXiv:1012.1086v3*, 2010.

[9] M. Brand, “Incremental singular value decomposition of uncertain data with missing values”, in *European Conference on Computer Vision*, 2002, pp. 707–720.

[10] D. Skocaj and A. Leonardis, “Weighted and robust incremental method for subspace learning”, in *IEEE Intl. Conf. on Computer Vision (ICCV)*, Oct 2003, vol. 2, pp. 1494 –1501.

[11] Y. Li, L. Xu, J. Morphett, and R. Jacobs, “An integrated algorithm of incremental and robust pca”, in *IEEE Intl. Conf. Image Proc. (ICIP)*, 2003, pp. 245–248.

[12] C. Qiu and N. Vaswani, “Real-time robust principal components’ pursuit”, in *Allerton Conference on Communication, Control, and Computing*, 2010.

[13] C. Qiu and N. Vaswani, “Support-predicted modified-cs for principal components’ pursuit”, in *IEEE Intl. Symp. on Information Theory (ISIT)*, 2011.

[14] J.N. Laska, M.A. Davenport, and R.G. Baraniuk, “Exact signal recovery from sparsely corrupted measurements through the pursuit of justice”, in *Asilomar Conf. on Sig. Sys. Comp.*, Nov 2009, pp. 1556 –1560.

[15] N. H. Nguyen and T. D. Tran, “Robust lasso with missing and grossly corrupted observations”, *To appear in IEEE Transaction on Information Theory*, 2012.

[16] E. Candes and T. Tao, “Decoding by linear programming”, *IEEE Trans. Info. Th.*, vol. 51(12), pp. 4203 – 4215, Dec. 2005.

[17] C. Qiu, N. Vaswani, and L. Hogben, “Recursive robust pca or recursive sparse recovery in large but structured noise”, *arXiv: 1211.3754 [cs.IT]*, 2012.

[18] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit”, *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[19] E. Candes, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information”, *IEEE Trans. Info. Th.*, vol. 52(2), pp. 489–509, February 2006.

[20] D. Donoho, “Compressed sensing”, *IEEE Trans. on Information Theory*, vol. 52(4), pp. 1289–1306, April 2006.

[21] E. Candes, “The restricted isometry property and its implications for compressed sensing”, *Compte Rendus de l’Academie des Sciences, Paris, Serie I*, pp. 589–592, 2008.

[22] E. Candes and T. Tao, “The dantzig selector: statistical estimation when p is much larger than n”, *Annals of Statistics*, 2006.

[23] Y. Jin and B. Rao, “Algorithms for robust linear regression by exploiting the connection to sparse signal recovery”, in *IEEE Intl. Conf. Acoustics, Speech, Sig. Proc. (ICASSP)*, 2010.

[24] K. Mitra, A. Veeraraghavan, and R. Chellappa, “A robust regression using sparse learing for high dimensional parameter estimation problems”, in *IEEE Intl. Conf. Acous. Speech. Sig.Proc.(ICASSP)*, 2010.

[25] C. Qiu and N. Vaswani, “Recursive sparse recovery in large but structured noise – part 2”, *arXiv: 1303.1144 [cs.IT]*, 2013.

[26] J. A. Tropp, “User-friendly tail bounds for sums of random matrices”, *Foundations of Computational Mathematics*, vol. 12, no. 4, 2012.