Recursive Robust PCA or Recursive Sparse Recovery in Large but Structured Noise

Chenlu Qiu, *Member, IEEE*, Namrata Vaswani, *Senior Member, IEEE*, Brian Lois, *Graduate Student Member, IEEE*, and Leslie Hogben

***Abstract*— This paper studies the recursive robust principal components analysis problem. 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 structure that we assume on L_t is that L_t is dense and lies in a low-dimensional subspace that is either fixed or changes slowly enough. A key application where this problem occurs is in video surveillance where the goal is to separate a slowly changing background (L_t) from moving foreground objects (S_t) on-the-fly. To solve the above problem, in recent work, we introduced a novel solution called recursive projected CS (ReProCS). In this paper, we develop a simple modification of the original ReProCS idea and analyze it. This modification assumes knowledge of a subspace change model on the L_t's. Under mild assumptions and a denseness assumption on the unestimated part of the subspace of L_t at various times, we show that, with high probability, the proposed approach 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 simulation experiments, we observe that the last assumption holds as long as there is some support change of S_t every few frames.**

***Index Terms*— Robust PCA, sparse recovery, compressive sensing, robust matrix completion.**

I. Introduction

PRINCIPAL Components Analysis (PCA) is a widely used dimension reduction technique that finds a small number of orthogonal basis vectors, called principal components (PCs), along which most of the variability of the dataset lies. It is well known that PCA is sensitive to outliers. Accurately computing the PCs in the presence of outliers is called robust PCA [4]–[7]. Often, for time series data, the PCs space changes gradually over time. Updating it on-the-fly (recursively) in the presence of outliers, as more data comes in is referred to as online or recursive robust PCA [8]–[10]. "Outlier" is a loosely defined term that refers to any corruption that is not small compared to the true data vector and that occurs occasionally. As suggested in [6] and [11], an outlier can be nicely modeled as a sparse vector whose nonzero values can have any magnitude.

A key application where the robust PCA problem occurs is in video analysis where the goal is to separate a slowly changing background from moving foreground objects [5], [6]. If we stack each frame as a column vector, the background is well modeled as being dense and lying in a low dimensional subspace that may gradually change over time, while the moving foreground objects constitute the sparse outliers [6], [11]. Other applications include detection of brain activation patterns from functional MRI (fMRI) sequences (the "active" part of the brain can be interpreted as a sparse outlier), detection of anomalous behavior in dynamic social networks and sensor networks based detection and tracking of abnormal events such as forest fires or oil spills. Clearly, in all these applications, an online solution is desirable.

The moving objects or the active regions of the brain 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. Thus, this problem can also be interpreted as one of recursively recovering a time sequence of 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 III-B.

A. Related Work

There has been a large amount of work on robust PCA, see [4]–[7], [12]–[14], and recursive robust PCA see [8]–[10]. In most of these works, either the locations of the missing/corruped data points are assumed known [8] (not a practical assumption); or they first detect the corrupted data points and then replace their values using nearby values [9]; or weight each data point in proportion to its reliability (thus soft-detecting and down-weighting the likely outliers) [5], [10]; or just remove the entire outlier vector [13], [14]. Detecting or soft-detecting outliers (S_t) as in [5], [9], [10] is easy when the outlier magnitude is large, but not otherwise. When the signal

Manuscript received March 4, 2013; revised December 17, 2013; accepted May 11, 2014. Date of publication June 17, 2014, date of current version July 10, 2014. This work was supported by the National Science Foundation under Grant CCF-0917015 and Grant CCF-1117125. This paper was presented at the 2010 Allerton Conference on Communication, Control, and Computing, 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing, and 2013 International Symposium on Information Theory.

C. Qiu and N. Vaswani are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: chenlu@iastate.edu; namrata@iastate.edu).

B. Lois and L. Hogben are with the Department of Mathematics, Iowa State University, Ames, IA 50011 USA (e-mail: blois@iastate.edu; lhogben@iastate.edu).

Communicated by Y. Ma, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2014.2331344

0018-9448 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

of interest is S_t, the most difficult situation is when nonzero elements of S_t have small magnitude compared to those of L_t and in this case, these approaches do not work.

In recent works [6] and [7], a new and elegant solution to robust PCA called Principal Components' Pursuit (PCP) has been proposed, that does not require a two step outlier location detection/correction process and also does not throw out the entire vector. 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, $\mathcal{M}_t := [M_1, \dots, M_t] = \mathcal{L}_t + \mathcal{S}_t$. Other recent works that also study batch algorithms for recovering a sparse $\mathcal{S}_t$ and a low-rank $\mathcal{L}_t$ from $\mathcal{M}_t := \mathcal{L}_t + \mathcal{S}_t$ or from undersampled measurements include [15]–[24].

Let $\|A\|_*$ be 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. It was shown in [6] that, with high probability (w.h.p.), one can recover $\mathcal{L}_t$ and $\mathcal{S}_t$ exactly by solving PCP:

$$\min_{\mathcal{L},\mathcal{S}} \|\mathcal{L}\|_* + \lambda \|\mathcal{S}\|_1 \text{ subject to } \mathcal{L} + \mathcal{S} = \mathcal{M}_t \tag{1}$$

provided that (a) the left and right singular vectors of $\mathcal{L}_t$ are dense; (b) any element of the matrix $\mathcal{S}_t$ is nonzero w.p. ϱ, and zero w.p. $1 - \varrho$, independent of all others; and (c) the rank of $\mathcal{L}_t$ is bounded by a small enough value.

As described earlier, many applications where robust PCA is required, such as video surveillance, require an online (recursive) solution. Even for offline applications, a recursive solution is typically faster than a batch one. In recent work [1], [25], and [26], we introduced a novel solution approach, called Recursive Projected Compressive Sensing (ReProCS), that recursively recovered S_t and L_t at each time t. In simulation and real data experiments (see [26] and http://www.ece.iastate.edu/~chenlu/ReProCS/ReProCS_main.htm), it was faster than batch methods such as PCP and also significantly outperformed them in situations where the support changes were correlated over time (as long as there was some support change every few frames) or when the background subspace dimension was large (for a given support size). In this work we develop a simple modification of the original ReProCS idea and analyze it. This modification assumes knowledge of the subspace change model on the L_t's.

B. Our Contributions

We show that (i) if an estimate of the subspace of L_t at the initial time is available; (ii) if L_t, lies in a slowly changing low dimensional subspace as defined in Sec III-B, (iii) if this subspace is dense, if (iv) the unestimated part of the changed subspace is dense at all times, and (v) if the subspace change model is known to the algorithm, then, 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. Moreover, after every subspace change time, w.h.p., the subspace error decays to a small enough value within a finite delay. Because (iv) depends on an algorithm estimate, our result, in its current form, cannot be interpreted as a correctness result but only a useful step towards it. From simulation experiments, we have observed that (iv) holds for correlated support changes as long as the support changes every few frames. This connection is being quantified in ongoing work. Assumption (v) is also restrictive and we explain in Sec IV-D how it can possibly be removed in future work.

We also develop and analyze a generalization of ReProCS called ReProCS with cluster-PCA (ReProCS-cPCA) that is designed for a more general subspace change model, and that needs an extra clustering assumption. Its main advantage is that it does not require a bound on the number of subspace changes, J, as long as the separation between the change points is allowed to grow logarithmically with J. Equivalently, it does not need a bound on the rank of $\mathcal{L}_t$.

If L_t is the signal of interest, then ReProCS is a solution to recursive robust PCA in the presence of sparse outliers. To the best of our knowledge, this is the first analysis of any recursive (online) robust PCA approach. If S_t is the signal of interest, then ReProCS is a solution to recursive sparse recovery in large but low-dimensional noise. To our knowledge, this work is also the first to analyze any recursive (online) sparse plus low-rank recovery algorithm. Another online algorithm that addresses this problem is given in [27], however, it does not contain any performance analysis. Our results directly apply to the recursive version of the matrix completion problem [28], [29] as well since it is a simpler special case of the current problem (the support set of S_t is the set of indices of the missing entries and is thus known) [6].

The proof techniques used in our work are very different from those used to analyze other recent batch robust PCA works [6], [7], [12]–[16], [20]–[24]. The works of [13] and [14] also study a different case: that where an entire vector is either an outlier or an inlier. Our proof utilizes (a) sparse recovery results [30]; (b) results from matrix perturbation theory that bound the estimation error in computing the eigenvectors of a perturbed Hermitian matrix with respect to eigenvectors of the original Hermitian matrix (the famous sin θ theorem of Davis and Kahan [31]) and (c) high probability bounds on eigenvalues of sums of independent random matrices (matrix Hoeffding inequality [32]).

A key difference of our approach to analyzing the subspace estimation step compared with most existing work analyzing finite sample PCA, see [33] and references therein, is that it needs to provably work in the presence of error/noise that is correlated with L_t. Most existing works, including [33] and the references it discusses, assume that the noise is independent of (or at least uncorrelated with) the data. However, in our case, because of how the estimate $\hat{L}_t$ is computed, the error $e_t := L_t - \hat{L}_t$ is correlated with L_t. As a result, the tools developed in these earlier works cannot be used for our problem. This is also the reason why simple PCA cannot be used and we need to develop and analyze projection-PCA based approaches for subspace estimation (see Appendix B for details).

The ReProCS approach is related to that of [34]–[36] in that all of these first try to nullify the low dimensional signal by projecting the measurement vector into a subspace perpendicular to that of the low dimensional signal, and then solve for the sparse "error" vector (outlier). However, the big difference is that in all of these works the basis for the subspace of the low dimensional signal is *perfectly known.* Our work studies *the case where the subspace is not known*. We have an initial approximate estimate of the subspace, but over time it can change significantly. In this work, to keep things simple, we use ℓ_1 minimization done separately for each time instant (also referred to as basis pursuit denoising (BPDN)) [30], [37]. However, this can be replaced by any other sparse recovery algorithm, either recursive or batch, as long as the batch algorithm is applied to α frames at a time, e.g. one can replace BPDN by modified-CS or support-predicted modified-CS [38].

C. Paper Organization

The rest of the paper is organized as follows. We give the notation and background required for the rest of the paper in Sec II. The problem definition and the model assumptions are given in Sec III. We explain the ReProCS algorithm and give its performance guarantees (Theorem 4.2) in Sec IV. The terms used in the proof are defined in Sec V. The proof is given in Sec VI. A more general subspace change model and ReProCS-cPCA which is designed to handle this model are given in Sec. VII. We also give the main result for ReProCS-cPCA in this section and discuss it. A discussion with respect to the result for PCP [6] is also provided here. Section VIII contains the proof of this theorem. In Sec IX-A, we show that our slow subspace change model indeed holds for real videos. In Sec IX-B, we show numerical experiments demonstrating Theorem 4.2, as well as comparisons of ReProCS with PCP. Conclusions and future work are given in Sec X.

II. Notation and Background

A. 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. We use T^c to denote its complement w.r.t. $\{1, 2, \dots n\}$, i.e. $T^c := \{i \in \{1, 2, \dots n\} : i \notin T\}$.

We use the interval notation, $[t_1, t_2]$, to denote the set of all integers between and including t_1 to t_2, i.e. $[t_1, t_2] := \{t_1, t_1 + 1, \dots, t_2\}$. 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. The support of v, supp(v), is the set of indices at which v is nonzero, $\text{supp}(v) := \{i : v_i \neq 0\}$. We say that v is s-sparse if $|\text{supp}(v)| \le s$.

For a matrix B, B' denotes its transpose, and $B^\dagger$ its pseudo-inverse. For a matrix with linearly independent columns, $B^\dagger = (B'B)^{-1}B'$. We use $\|B\|_2 := \max_{x \neq 0} \|Bx\|_2/\|x\|_2$ to denote the induced 2-norm of the matrix. Also, $\|B\|_*$ is the nuclear norm (sum of singular values) and $\|B\|_{\max}$ denotes the maximum over the absolute values of all its entries. We let $\sigma_i(B)$ denotes the ith largest singular value of B. 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 decreasing order. Also, we use $\lambda_i(B)$ to denote the ith largest eigenvalue of a Hermitian matrix B and we use $\lambda_{\max}(B)$ and $\lambda_{\min}(B)$ denote its maximum and minimum eigenvalues. If B is Hermitian positive semi-definite (p.s.d.), then $\lambda_i(B) = \sigma_i(B)$. For Hermitian matrices B_1 and B_2, the notation $B_1 \preceq B_2$ means that $B_2 - B_1$ is p.s.d. Similarly, $B_1 \succeq B_2$ means that $B_1 - B_2$ is p.s.d.

For a Hermitian matrix B, $\|B\|_2 = \sqrt{\max(\lambda_{\max}^2(B), \lambda_{\min}^2(B))}$ and thus, $\|B\|_2 \le b$ implies that $-b \le \lambda_{\min}(B) \le \lambda_{\max}(B) \le b$.

We use I to denote an identity matrix of appropriate size. 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. Notice that $B_T = BI_T$. Given a matrix B of size $m \times n$ and B_2 of size $m \times n_2$, $[B \; B_2]$ constructs a new matrix by concatenating matrices B and B_2 in the horizontal direction. Let B_{rem} be a matrix containing some columns of B. Then $B \setminus B_{\text{rem}}$ is the matrix B with columns in B_{rem} removed.

For a tall matrix P, span(P) denotes the subspace spanned by the column vectors of P.

The notation [.] denotes an empty matrix.

Definition 2.1: We refer to a tall matrix P as a basis matrix *if it satisfies $P'P = I$.*

Definition 2.2: We use the notation $Q = \text{basis}(B)$ to mean that Q is a basis matrix and $\text{span}(Q) = \text{span}(B)$. In other words, the columns of Q form an orthonormal basis for the range of B.

Definition 2.3: The *s-restricted isometry constant (RIC)* [34], *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 $|T|$.*

It is easy to see that $\max_{T:|T|\le s} \|({\Psi_T}'\Psi_T)^{-1}\|_2 \le \frac{1}{1-\delta_s(\Psi)}$ [34].

Definition 2.4: We give some notation for random variables in this definition.

1) *We let $\mathbf{E}[Z]$ denote the expectation of a random variable (r.v.) Z and $\mathbf{E}[Z|X]$ denote its conditional expectation given another r.v. X.*
2) *Let $\mathcal{B}$ be a set of values that a r.v. Z can take. We use $\mathcal{B}^e$ to denote the* event $Z \in \mathcal{B}$, *i.e. $\mathcal{B}^e := \{Z \in \mathcal{B}\}$.*
3) *The probability of any event $\mathcal{B}^e$ can be expressed as [39],*
$$\mathbf{P}(\mathcal{B}^e) := \mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)].$$
where
$$\mathbb{I}_{\mathcal{B}}(Z) := \begin{cases} 1 & \text{if } Z \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$
is the indicator function on the set $\mathcal{B}$.
4) *For two events $\mathcal{B}^e, \tilde{\mathcal{B}}^e$, $\mathbf{P}(\mathcal{B}^e|\tilde{\mathcal{B}}^e)$ refers to the conditional probability of $\mathcal{B}^e$ given $\tilde{\mathcal{B}}^e$, i.e. $\mathbf{P}(\mathcal{B}^e|\tilde{\mathcal{B}}^e) := \mathbf{P}(\mathcal{B}^e, \tilde{\mathcal{B}}^e)/\mathbf{P}(\tilde{\mathcal{B}}^e)$.*
5) *For a r.v. X, and a set $\mathcal{B}$ of values that the r.v. Z can take, the notation $\mathbf{P}(\mathcal{B}^e|X)$ is defined as*
$$\mathbf{P}(\mathcal{B}^e|X) := \mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)|X].$$
Notice that $\mathbf{P}(\mathcal{B}^e|X)$ is a r.v. (it is a function of the r.v. X) that always lies between zero and one.

Finally, RHS refers to the right hand side of an equation or inequality; w.p. means "with probability"; and w.h.p. means "with high probability".

B. Compressive Sensing Result

The error bound for noisy compressive sensing (CS) based on the RIC is as follows [30].

Theorem 2.5 ([30]): Suppose we observe

$$y := \Psi x + z$$

where z is the noise. Let $\hat{x}$ be the solution to following problem

$$\min_x \|x\|_1 \text{ subject to } \|y - \Psi x\|_2 \le \xi \qquad (2)$$

Assume that x is s-sparse, $\|z\|_2 \le \xi$, and $\delta_{2s}(\Psi) < b(\sqrt{2}-1)$ for some $0 \le b < 1$. Then the solution of (2) obeys

$$\|\hat{x} - x\|_2 \le C_1 \xi$$

with $C_1 = \dfrac{4\sqrt{1+\delta_{2s}(\Psi)}}{1-(\sqrt{2}+1)\delta_{2s}(\Psi)} \le \dfrac{4\sqrt{1+b(\sqrt{2}-1)}}{1-b}$.

Remark 2.6: Notice that if b is small enough, C_1 is a small constant but $C_1 > 1$. For example, if $\delta_{2s}(\Psi) \le 0.15$, then $C_1 \le 7$. If $C_1 \xi > \|x\|_2$, the normalized reconstruction error bound would be greater than 1, making the result useless. Hence, (2) gives a small reconstruction error bound only for the small noise case, i.e., the case where $\|z\|_2 \le \xi \ll \|x\|_2$.

C. Results From Linear Algebra

Davis and Kahan's $sin\theta$ Theorem [31] studies the rotation of eigenvectors by perturbation.

Theorem 2.7 ($\sin\theta$ theorem [31]): Given two Hermitian matrices $\mathcal{A}$ and $\mathcal{H}$ satisfying

$$\mathcal{A} = \begin{bmatrix} E & E_\perp \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & A_\perp \end{bmatrix} \begin{bmatrix} E' \\ E_\perp{}' \end{bmatrix},$$
$$\mathcal{H} = \begin{bmatrix} E & E_\perp \end{bmatrix} \begin{bmatrix} H & B' \\ B & H_\perp \end{bmatrix} \begin{bmatrix} E' \\ E_\perp{}' \end{bmatrix} \qquad (3)$$

where $[E \ E_\perp]$ is an orthonormal matrix. The two ways of representing $\mathcal{A} + \mathcal{H}$ are

$$\mathcal{A} + \mathcal{H} = \begin{bmatrix} E & E_\perp \end{bmatrix} \begin{bmatrix} A+H & B' \\ B & A_\perp + H_\perp \end{bmatrix} \begin{bmatrix} E' \\ E_\perp{}' \end{bmatrix}$$
$$= \begin{bmatrix} F & F_\perp \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda_\perp \end{bmatrix} \begin{bmatrix} F' \\ F_\perp{}' \end{bmatrix}$$

where $[F \ F_\perp]$ is another orthonormal matrix. Let $\mathcal{R} := (\mathcal{A}+\mathcal{H})E - \mathcal{A}E = \mathcal{H}E$. If $\lambda_{\min}(A) > \lambda_{\max}(\Lambda_\perp)$, then

$$\|(I - FF')E\|_2 \le \frac{\|\mathcal{R}\|_2}{\lambda_{\min}(A) - \lambda_{\max}(\Lambda_\perp)}$$

The above result bounds the amount by which the two subspaces $\text{span}(E)$ and $\text{span}(F)$ differ as a function of the norm of the perturbation $\|\mathcal{R}\|_2$ and of the gap between the minimum eigenvalue of A and the maximum eigenvalue of $\Lambda_\perp$. Next, we state Weyl's theorem which bounds the eigenvalues of a perturbed Hermitian matrix, followed by Ostrowski's theorem.

Theorem 2.8 (Weyl [40]): Let $\mathcal{A}$ and $\mathcal{H}$ be two $n \times n$ Hermitian matrices. For each $i = 1, 2, \ldots, n$ we have

$$\lambda_i(\mathcal{A}) + \lambda_{\min}(\mathcal{H}) \le \lambda_i(\mathcal{A} + \mathcal{H}) \le \lambda_i(\mathcal{A}) + \lambda_{\max}(\mathcal{H})$$

Theorem 2.9 (Ostrowski [40]): Let H and W be $n \times n$ matrices, with H Hermitian and W nonsingular. For each $i = 1, 2 \ldots n$, there exists a positive real number θ_i such that $\lambda_{\min}(WW') \le \theta_i \le \lambda_{\max}(WW')$ and $\lambda_i(WHW') = \theta_i \lambda_i(H)$. Therefore,

$$\lambda_{\min}(WHW') \ge \lambda_{\min}(WW')\lambda_{\min}(H)$$

The following lemma proves some simple linear algebra facts.

Lemma 2.10: Suppose that P, $\hat{P}$ and Q are three basis matrices. Also, P and $\hat{P}$ are of the same size, $Q'P = 0$ and $\|(I - \hat{P}\hat{P}')P\|_2 = \zeta_*$. Then,

1) $\|(I - \hat{P}\hat{P}')PP'\|_2 = \|(I - PP')\hat{P}\hat{P}'\|_2 = \|(I - PP')\hat{P}\|_2 = \|(I - \hat{P}\hat{P}')P\|_2 = \zeta_*$
2) $\|PP' - \hat{P}\hat{P}'\|_2 \le 2\|(I - \hat{P}\hat{P}')P\|_2 = 2\zeta_*$
3) $\|\hat{P}'Q\|_2 \le \zeta_*$
4) $\sqrt{1-\zeta_*^2} \le \sigma_i((I - \hat{P}\hat{P}')Q) \le 1$

Further, if P is an $n \times r_1$ basis matrix and $\hat{P}$ is an $n \times r_2$ basis matrix with $r_2 \ge r_1$, then $\|(I - \hat{P}\hat{P}')PP'\|_2 \le \|(I - PP')\hat{P}\hat{P}'\|_2$

The proof is in the Appendix.

D. High Probability Tail Bounds for Sums of Independent Random Matrices

The following lemma follows easily using Definition 2.4. We will use this at various places in the paper.

Lemma 2.11: Suppose that $\mathcal{B}$ is the set of values that the r.v.s X, Y can take. Suppose that $\mathcal{C}$ is a set of values that the r.v. X can take. For a $0 \le p \le 1$, if $\mathbf{P}(\mathcal{B}^e|X) \ge p$ for all $X \in \mathcal{C}$, then $\mathbf{P}(\mathcal{B}^e|\mathcal{C}^e) \ge p$ as long as $\mathbf{P}(\mathcal{C}^e) > 0$.

The proof is in the Appendix.

The following lemma is an easy consequence of the chain rule of probability applied to a contracting sequence of events.

Lemma 2.12: For a sequence of events $E_0^e, E_1^e, \ldots E_m^e$ that satisfy $E_0^e \supseteq E_1^e \supseteq E_2^e \cdots \supseteq E_m^e$, the following holds

$$\mathbf{P}(E_m^e|E_0^e) = \prod_{k=1}^{m} \mathbf{P}(E_k^e|E_{k-1}^e).$$

Proof: $\mathbf{P}(E_m^e|E_0^e) = \mathbf{P}(E_m^e, E_{m-1}^e, \ldots E_0^e|E_0^e) = \prod_{k=1}^m \mathbf{P}(E_k^e|E_{k-1}^e, E_{k-2}^e, \ldots E_0^e) = \prod_{k=1}^m \mathbf{P}(E_k^e|E_{k-1}^e)$. ■

Next, we state the matrix Hoeffding inequality [32, Th. 1.3] which gives tail bounds for sums of independent random matrices.

Theorem 2.13 (Matrix Hoeffding for a Zero Mean Hermitian Matrix [32]): Consider a finite sequence $\{Z_t\}$ of independent, random, Hermitian matrices of size $n \times n$, and let $\{A_t\}$ be a sequence of fixed Hermitian matrices. Assume that each random matrix satisfies (i) $\mathbf{P}(Z_t^2 \preceq A_t^2) = 1$ and (ii) $\mathbf{E}(Z_t) = 0$. Then, for all $\epsilon > 0$,

$$\mathbf{P}\left(\lambda_{\max}\left(\sum_t Z_t\right) \le \epsilon\right) \ge 1 - n \exp\left(\frac{-\epsilon^2}{8\sigma^2}\right),$$

where $\sigma^2 = \left\| \sum_t A_t^2 \right\|_2$.

The following two corollaries of Theorem 2.13 are easy to prove. The proofs are given in Appendix A.

Corollary 2.14 (Matrix Hoeffding Conditioned on Another Random Variable for a Nonzero Mean Hermitian Matrix): Given an α-length sequence of random matrices $\{Z_t\}$ of size $n \times n$, a r.v. X, and a set $\mathcal{C}$ of values that X can take. Assume that, for all $X \in \mathcal{C}$, (i) Z_t's are conditionally independent given X; (ii) $\mathbf{P}(b_1 I \preceq Z_t \preceq b_2 I | X) = 1$ and (iii) $b_3 I \preceq \frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X) \preceq b_4 I$. Then for all $\epsilon > 0$,

$$\mathbf{P}\left(\lambda_{\max}\left(\frac{1}{\alpha}\sum_t Z_t\right) \leq b_4 + \epsilon \middle| X\right) \geq 1 - n\exp\left(\frac{-\alpha\epsilon^2}{8(b_2-b_1)^2}\right) \text{ for all } X \in \mathcal{C}$$

$$\mathbf{P}\left(\lambda_{\min}\left(\frac{1}{\alpha}\sum_t Z_t\right) \geq b_3 - \epsilon \middle| X\right) \geq 1 - n\exp\left(\frac{-\alpha\epsilon^2}{8(b_2-b_1)^2}\right) \text{ for all } X \in \mathcal{C}$$

The proof is in Appendix A.

Corollary 2.15 (Matrix Hoeffding Conditioned on Another Random Variable for an Arbitrary Nonzero Mean Matrix): Given an α-length sequence $\{Z_t\}$ of random Hermitian matrices of size $n \times n$, a r.v. X, and a set $\mathcal{C}$ of values that X can take. Assume that, for all $X \in \mathcal{C}$, (i) Z_t's are conditionally independent given X; (ii) $\mathbf{P}(\|Z_t\|_2 \leq b_1|X) = 1$ and (iii) $\|\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X)\|_2 \leq b_2$. Then, for all $\epsilon > 0$,

$$\mathbf{P}\left(\left\|\frac{1}{\alpha}\sum_t Z_t\right\|_2 \leq b_2 + \epsilon \middle| X\right) \geq 1 - (n_1+n_2)\exp\left(\frac{-\alpha\epsilon^2}{32 b_1^2}\right) \text{ for all } X \in \mathcal{C}$$

The proof is in Appendix A.

III. Problem Definition and Model Assumptions

We give the problem definition below followed by the model and then describe the two key assumptions.

A. Problem Definition

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

$$M_t = L_t + S_t \tag{4}$$

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, i.e. $L_t = P_{(t)} a_t$ where $P_{(t)}$ is an $n \times r_{(t)}$ basis matrix with $r_{(t)} < n$, that changes every so often according to 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 $\text{span}(\mathcal{L}_{t_{\text{train}}})$, i.e. $\text{span}(P_0) = \text{span}(\mathcal{L}_{t_{\text{train}}})$. Also, for the first t_{train} time instants, S_t is zero. The goal is

1) to estimate both S_t and L_t at each time $t > t_{\text{train}}$, and
2) to estimate $\text{span}(\mathcal{L}_t)$ every so often, i.e. compute $\hat{P}_{(t)}$ so that the subspace estimation error, $\text{SE}_{(t)} := \|(I - \hat{P}_{(t)}\hat{P}_{(t)}')P_{(t)}\|_2$ is small.

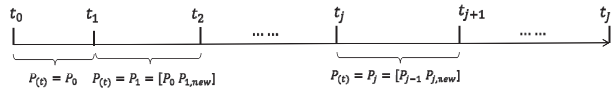

Fig. 1. The subspace change model explained in Sec III-A. Here $t_0 = 0$ and $0 < t_{\text{train}} < t_1$.

We assume a subspace change model that allows the subspace to change at certain change times t_j rather than continuously at each time. It should be noted that this is only a model for reality. In practice there will typically be some changes at every time t; however this is difficult to model in a simple fashion. Moreover the analysis for such a model will be a lot more complicated. However, we do allow the variance of the projection of L_t along the subspace directions to change continuously. The projection along the new directions is assumed to be small initially and allowed to gradually increase to a large value (see Sec III-B).

Signal Model 3.1 (Model on L_t):

1) *We assume that $L_t = P_{(t)} a_t$ with $P_{(t)} = P_j$ for all $t_j \leq t < t_{j+1}$, $j = 0, 1, 2 \cdots J$. Here P_j is an $n \times r_j$ basis matrix with $r_j < \min(n, (t_{j+1} - t_j))$ that 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

$$r_j = \text{rank}(P_j) = r_{j-1} + c_{j,\text{new}}.$$

We let $t_0 = 0$. Also t_{J+1} can be the length of the sequence or $t_{J+1} = \infty$. This model is illustrated in Figure 1.

2) *The vector of coefficients, $a_t := P_{(t)}{}' L_t$, is a zero mean random variable (r.v.) with mutually uncorrelated entries, i.e. $\mathbf{E}[a_t] = 0$ and $\mathbf{E}[(a_t)_i (a_t)_j] = 0$ for $i \neq j$.*

Definition 3.2: Define the covariance matrix of a_t to be the diagonal matrix

$$\Lambda_t := \text{Cov}[a_t] = \mathbf{E}(a_t a_t').$$

Define For $t_j \leq t < t_{j+1}$, a_t is an r_j length vector which can be split as

$$a_t = P_j{}' L_t = \begin{bmatrix} a_{t,*} \\ a_{t,\text{new}} \end{bmatrix}$$

where $a_{t,*} := P_{j-1}{}' L_t$ and $a_{t,\text{new}} := P_{j,\text{new}}{}' L_t$. Thus, for this interval, L_t can be rewritten as

$$L_t = \begin{bmatrix} P_{j-1} & P_{j,\text{new}} \end{bmatrix} \begin{bmatrix} a_{t,*} \\ a_{t,\text{new}} \end{bmatrix} = P_{j-1} a_{t,*} + P_{j,\text{new}} a_{t,\text{new}}$$

Also, Λ_t can be split as

$$\Lambda_t = \begin{bmatrix} (\Lambda_t)_* & 0 \\ 0 & (\Lambda_t)_{\text{new}} \end{bmatrix}$$

where $(\Lambda_t)_* = \mathrm{Cov}[a_{t,*}]$ *and* $(\Lambda_t)_{\mathrm{new}} = \mathrm{Cov}[a_{t,\mathrm{new}}]$ *are diagonal matrices. Define*

$$\lambda^- := \inf_t \lambda_{\min}(\Lambda_t), \quad \lambda^+ := \sup_t \lambda_{\max}(\Lambda_t),$$

and

$$\lambda^-_{\mathrm{new}} := \inf_t \lambda_{\min}((\Lambda_t)_{\mathrm{new}}), \quad \lambda^+_{\mathrm{new}} := \sup_t \lambda_{\max}((\Lambda_t)_{\mathrm{new}}).$$

Also let,

$$f := \frac{\lambda^+}{\lambda^-}$$

and

$$g := \frac{\lambda^+_{\mathrm{new}}}{\lambda^-_{\mathrm{new}}}.$$

The above simple model only allows new additions to the subspace and hence the rank of P_j can only grow over time. The ReProCS algorithm designed for this model can be interpreted as a recursive algorithm for solving the robust PCA problem studied in [6] and other batch robust PCA works. At time t we estimate the subspace spanned by $L_1, L_2, \ldots L_t$. For the above model, the subspace dimension is bounded by $r_0 + Jc_{\max}$. Thus a bound on J is needed to keep the subspace dimension small at all times. We remove this limitation in Sec VII where we also allow for subspace deletions and correspondingly design a ReProCS algorithm that does the same thing. For that algorithm, as we will see, we will not need a bound on the number of changes, J, as long as the separation between the subspace change times is allowed to grow logarithmically with J and a clustering assumption holds.

Define the following quantities for the sparse part.

Definition 3.3: Let $T_t := \{i : (S_t)_i \neq 0\}$ *denote the support of* S_t*. Define*

$$S_{\min} := \min_{t > t_{\mathrm{train}}} \min_{i \in T_t} |(S_t)_i|, \text{ and } s := \max_t |T_t|.$$

B. Slow Subspace Change

By slow subspace change we mean all of the following.

First, the delay between consecutive subspace change times is large enough, i.e., for a d large enough,

$$t_{j+1} - t_j \geq d \tag{5}$$

Second, the magnitude of the projection of L_t along the newly added directions, $a_{t,\mathrm{new}}$, is initially small but can increase gradually. We model this as follows. Assume that for an $\alpha > 0$[1] the following holds

$$\|a_{t,\mathrm{new}}\|_\infty \leq \min\left(v^{\frac{t-t_j}{\alpha}-1}\gamma_{\mathrm{new}}, \gamma_*\right) \tag{6}$$

when $t \in [t_j, t_{j+1}-1]$ for a $v > 1$ but not too large and with $\gamma_{\mathrm{new}} < \gamma_*$ and $\gamma_{\mathrm{new}} < S_{\min}$. Clearly, the above assumption implies that

$$\|a_{t,\mathrm{new}}\|_\infty \leq \gamma_{\mathrm{new},k} := \min(v^{k-1}\gamma_{\mathrm{new}}, \gamma_*)$$

for all $t \in [t_j + (k-1)\alpha, t_j + k\alpha - 1]$. This assumption is verified for real video data in Sec. IX-A.

[1]As we will see in the algorithm α is the number of previous frames used to get a new estimate of $P_{j,\mathrm{new}}$.

Third, the number of newly added directions is small, i.e. $c_{j,\mathrm{new}} \leq c_{\max} \ll r_0$. This is also verified in Sec. IX-A.

Remark 3.4 (Large f): Since our problem definition allows large noise, L_t, but assumes slow subspace change, thus the maximum condition number of $\mathrm{Cov}[L_t]$*, which is bounded by f, cannot be bounded by a small value. The reason is as follows. Slow subspace change implies that the projection of L_t along the new directions is initially small, i.e. γ_{new} is small. Since $\lambda^- \leq \gamma_{\mathrm{new}}$, this means that λ^- is small. Since* $\mathbf{E}[\|L_t\|^2] \leq r_{\max}\lambda^+$ *and $r_{\max}$ is small (low-dimensional), thus, large L_t means that λ^+ needs to be large. As a result $f = \lambda^+/\lambda^-$ cannot be upper bounded by a small value.*

C. Measuring Denseness of a Matrix and Its Relation With RIC

Before we can state the denseness assumption, we need to define the denseness coefficient.

Definition 3.5 (Denseness Coefficient): For a matrix or a vector B, define

$$\kappa_s(B) = \kappa_s(\mathrm{span}(B)) := \max_{|T| \leq s} \|I_T{}'\mathrm{basis}(B)\|_2 \tag{7}$$

where $\|.\|_2$ is the vector or matrix ℓ_2-norm.

Clearly, $\kappa_s(B) \leq 1$. First consider an n-length vector B. Then κ_s measures the denseness (non-compressibility) of B. A small value indicates that the entries in B are spread out, i.e. it is a dense vector. A large value indicates that it is compressible (approximately or exactly sparse). The worst case (largest possible value) is $\kappa_s(B) = 1$ which indicates that B is an s-sparse vector. The best case is $\kappa_s(B) = \sqrt{s/n}$ and this will occur if each entry of B has the same magnitude. Similarly, for an $n \times r$ matrix B, a small κ_s means that most (or all) of its columns are dense vectors.

Remark 3.6: The following facts should be noted about $\kappa_s(.)$:

1) *For a given matrix B, $\kappa_s(B)$ is an non-decreasing function of s.*
2) *$\kappa_s([B_1]) \leq \kappa_s([B_1 \ B_2])$ i.e. adding columns cannot decrease κ_s.*
3) *A bound on $\kappa_s(B)$ is $\kappa_s(B) \leq \sqrt{s}\kappa_1(B)$. This follows because $\|B\|_2 \leq \left\|\left[\|b_1\|_2 \ldots \|b_r\|_2\right]\right\|_2$ where b_i is the i^{th} column of B.*

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

Lemma 3.7: For an $n \times r$ basis matrix P (i.e P satisfying $P'P = I$),

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

In other words, if P is dense enough (small κ_s), then the RIC of $I - PP'$ is small.

In this work, we assume an upper bound on $\kappa_{2s}(P_j)$ for all j, and a tighter upper bound on $\kappa_{2s}(P_{j,\mathrm{new}})$, i.e., there exist $\kappa^+_{2s,*} < 1$ and a $\kappa^+_{2s,\mathrm{new}} < \kappa^+_{2s,*}$ such that

$$\max_j \kappa_{2s}(P_{j-1}) \leq \kappa^+_{2s,*} \tag{8}$$

$$\max_j \kappa_{2s}(P_{j,\mathrm{new}}) \leq \kappa^+_{2s,\mathrm{new}} \tag{9}$$

Additionally, we also assume denseness of another matrix, $D_{j,\text{new},k}$, whose columns span the currently unestimated part of $\text{span}(P_{j,\text{new}})$ (see Theorem 4.2).

The denseness coefficient $\kappa_s(B)$ is related to the denseness assumption required by PCP [6]. That work uses $\kappa_1(B)$ to quantify denseness.

IV. Recursive Projected CS (ReProCS) and Its Performance Guarantees

In this section we introduce the ReProCS algorithm and state the performance guarantee for it. We begin by first stating the result in IV-A, and then describe and explain the algorithm in Section IV-C. In Section IV-B we describe the projection-PCA algorithm that is used in the ReProCS algorithm. The assumptions used by the result are discussed in Section IV-D.

A. Performance Guarantees

We state the main result here and then discuss it in Section IV-D. Definitions needed for the proof are given in Section V and the actual proof is given in Section VI.

Definition 4.1: We define here the parameters that will be used in Theorem 4.2.

1) *Let* $c := c_{\max}$ *and* $r := r_0 + (J-1)c$.
2) *Define* $K = K(\zeta) := \left\lceil \frac{\log(0.6c\zeta)}{\log 0.6} \right\rceil$
3) *Define* $\xi_0(\zeta) := \sqrt{c}\gamma_{\text{new}} + \sqrt{\zeta}(\sqrt{r} + \sqrt{c})$
4) *Define*

$$\alpha_{\text{add}}(\zeta) := \Bigg\lceil (\log 6KJ + 11\log n)\frac{8\cdot 24^2}{\zeta^2(\lambda^-)^2}\cdot \max\Bigg(\min(1.2^{4K}\gamma_{\text{new}}^4, \gamma_*^4), \frac{16}{c^2}, 4(0.186\gamma_{\text{new}}^2 + 0.0034\gamma_{\text{new}} + 2.3)^2\Bigg)\Bigg\rceil$$

We note that α_{add} is the number of data points, α, used for one projection PCA step and is chosen to ensure that the conclusions of Theorem 4.2 hold with probability at least $(1-n^{-10})$. If γ_* is large enough ($\gamma_*{}^4 > 16$), a simpler but larger value for $\alpha_{add}(\zeta)$ is

$$\alpha_{\text{add}}(\zeta) = \left\lceil (\log 6KJ + 11\log n)\frac{8\cdot 24^2\gamma_*^4}{\zeta^2(\lambda^-)^2}\right\rceil$$

Theorem 4.2: Consider Algorithm 2. Pick a ζ that satisfies

$$\zeta \le \min\left(\frac{10^{-4}}{r^2}, \frac{1.5\times 10^{-4}}{r^2 f}, \frac{1}{r^3\gamma_*^2}\right)$$

Assume that the initial subspace estimate is accurate enough, i.e. $\|(I-\hat{P}_0\hat{P}_0')P_0\| \le r_0\zeta$. If the following conditions hold:

1) *The algorithm parameters are set as $\xi = \xi_0(\zeta)$, $7\xi \le \omega \le S_{\min} - 7\xi$, $K = K(\zeta)$, $\alpha \ge \alpha_{add}(\zeta)$*
2) *L_t satisfies Signal Model 3.1 with*
   a) $0 \le c_{j,\text{new}} \le c_{\max}$ *for all j (thus $r_j \le r_{\max} := r_0 + Jc_{\max}$),*
   b) *the a_t's mutually independent over t,*
   c) $\|a_t\|_\infty \le \gamma_*$ *for all t (a_t's bounded);,*
   d) $0 < \lambda^- \le \lambda^+ < \infty$, *and*
   e) $g \le g^+ = \sqrt{2}$;

Algorithm 1 Projection-PCA: $Q \leftarrow \text{proj-PCA}(\mathcal{D}, P, r)$

1) Projection: compute $\mathcal{D}_{\text{proj}} \leftarrow (I - PP')\mathcal{D}$
2) PCA: compute $\frac{1}{\alpha}\mathcal{D}_{\text{proj}}\mathcal{D}_{\text{proj}}{}' \overset{EVD}{=} \begin{bmatrix} Q & Q_\perp \end{bmatrix}\begin{bmatrix}\Lambda & 0\\ 0 & \Lambda_\perp\end{bmatrix}\begin{bmatrix} Q' \\ Q_\perp{}'\end{bmatrix}$ where Q is an $n \times r$ basis matrix and α is the number of columns in $\mathcal{D}$.

---

3) *slow subspace change holds: (5) holds with $d = K\alpha$; (6) holds with $v = 1.2$; and c and γ_{new} are small enough so that $14\xi_0(\zeta) \le S_{\min}$.*
4) *denseness holds: equation* (8) *holds with $\kappa_{2s,*}^+ = 0.3$ and equation* (9) *holds with $\kappa_{2s,\text{new}}^+ = 0.15$*
5) *the matrices*

$$D_{j,\text{new},k} := (I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,\text{new},k}\hat{P}_{j,\text{new},k}')P_{j,\text{new}}$$

and

$$Q_{j,\text{new},k} := (I - P_{j,\text{new}}P_{j,\text{new}}{}')\hat{P}_{j,\text{new},k}$$

satisfy

$$\max_j \max_{1\le k\le K} \kappa_s(D_{j,\text{new},k}) \le \kappa_s^+ := 0.152$$

$$\max_j \max_{1\le k\le K} \kappa_{2s}(Q_{j,\text{new},k}) \le \tilde{\kappa}_{2s}^+ := 0.15$$

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

1) *at all times, t,*

$$\hat{T}_t = T_t \quad \text{and}$$
$$\|e_t\|_2 = \|L_t - \hat{L}_t\|_2 = \|\hat{S}_t - S_t\|_2 \le 0.18\sqrt{c}\gamma_{\text{new}} + 1.2\sqrt{\zeta}(\sqrt{r} + 0.06\sqrt{c}).$$

2) *the subspace error $SE_{(t)} := \|(I - \hat{P}_{(t)}\hat{P}_{(t)}')P_{(t)}\|_2$ satisfies*

$$\text{SE}_{(t)} \le \begin{cases} (r_0 + (j-1)c)\zeta + 0.4c\zeta + 0.6^{k-1} \\ \qquad \text{if } t \in \mathcal{I}_{j,k},\ k = 1,2\ldots K \\ (r_0 + jc)\zeta \quad \text{if } t \in \mathcal{I}_{j,K+1} \end{cases}$$
$$\le \begin{cases} 10^{-2}\sqrt{\zeta} + 0.6^{k-1} \\ \qquad \text{if } t \in \mathcal{I}_{j,k},\ k = 1,2\ldots K \\ 10^{-2}\sqrt{\zeta} \quad \text{if } t \in \mathcal{I}_{j,K+1} \end{cases}$$

3) *the error $e_t = \hat{S}_t - S_t = L_t - \hat{L}_t$ satisfies the following at various times*

$$\|e_t\|_2 \le \begin{cases} 0.18\sqrt{c}0.72^{k-1}\gamma_{\text{new}} + \\ \quad 1.2(\sqrt{r} + 0.06\sqrt{c})(r_0 + (j-1)c)\zeta\gamma_* \\ \qquad \text{if } t \in \mathcal{I}_{j,k},\ k = 1,2\ldots K \\ 1.2(r_0 + jc)\zeta\sqrt{r}\gamma_* \quad \text{if } t \in \mathcal{I}_{j,K+1} \end{cases}$$
$$\le \begin{cases} 0.18\sqrt{c}0.72^{k-1}\gamma_{\text{new}} + 1.2(\sqrt{r} + 0.06\sqrt{c})\sqrt{\zeta} \\ \qquad \text{if } t \in \mathcal{I}_{j,k},\ k = 1,2\ldots K \\ 1.2\sqrt{r}\sqrt{\zeta} \quad \text{if } t \in \mathcal{I}_{j,K+1} \end{cases}$$

Remark 4.3: Consider the last assumption. We actually also need a similar denseness of $\kappa_s(D_{j,\text{new}})$ where $D_{j,\text{new}} = D_{j,\text{new},0} = (I - \hat{P}_{j-1}\hat{P}_{j-1}')P_{j,\text{new}}$. Conditioned on the fact that $\text{span}(P_{j-1})$ has been accurately estimated, this follows easily from the denseness of $P_{j,\text{new}}$ (see Lemma 6.10).

The above result says the following. Consider Algorithm 2. Assume that the initial subspace error is small enough. If the algorithm parameters are appropriately set, if slow subspace

Algorithm 2 Recursive Projected CS (ReProCS)

Parameters: algorithm parameters: ξ, ω, α, K, model parameters: t_j, $c_{j,\text{new}}$
(set as in Theorem 4.2)
Input: M_t, *Output:* $\hat{S}_t$, $\hat{L}_t$, $\hat{P}_{(t)}$
Initialization: Compute $\hat{P}_0 \leftarrow$ proj-PCA$\big([L_1, L_2, \cdots, L_{t_{\text{train}}}], [.], r_0\big)$ where $r_0 = \text{rank}([L_1, L_2, \cdots, L_{t_{\text{train}}}])$.
Set $\hat{P}_{(t)} \leftarrow \hat{P}_0$, $j \leftarrow 1$, $k \leftarrow 1$.
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 $\min_x \|x\|_1 \ s.t. \ \|y_t - \Phi_{(t)} x\|_2 \le \xi$
   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 $\hat{P}_{(t)}$: K Projection PCA steps.
   a) If $t = t_j + k\alpha - 1$,
      i) $\hat{P}_{j,\text{new},k} \leftarrow$ proj-PCA$\left(\left[\hat{L}_{t_j+(k-1)\alpha}, \ldots, \hat{L}_{t_j+k\alpha-1}\right], \hat{P}_{j-1}, c_{j,\text{new}}\right)$.
      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.

change holds, if the subspaces are dense, if the condition number of Cov$[a_{t,\text{new}}]$ is small enough, and if the currently unestimated part of the newly added subspace is dense enough (this is an assumption on the algorithm estimates), 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, $\gamma_{\text{new}} < S_{\min}$, 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.

As we explain in Section IV-D, the most important limitation of our result is that it requires an assumption on $D_{\text{new},k}$ and $Q_{\text{new},k}$ which depend on algorithm estimates. Moreover, it studies an algorithm that requires knowledge of model parameters.

B. Projection-PCA Algorithm for ReProCS

Given a data matrix $\mathcal{D}$, a basis matrix P and an integer r, projection-PCA (proj-PCA) applies PCA on $\mathcal{D}_{\text{proj}} := (I - PP')\mathcal{D}$, i.e., it computes the top r eigenvectors (the eigenvectors with the largest r eigenvalues) of $\frac{1}{\alpha}\mathcal{D}_{\text{proj}}{\mathcal{D}_{\text{proj}}}'$. Here α is the number of column vectors in $\mathcal{D}$. This is summarized in Algorithm 1.

If $P = [.]$, then projection-PCA reduces to standard PCA, i.e. it computes the top r eigenvectors of $\frac{1}{\alpha}\mathcal{D}\mathcal{D}'$.

The reason we need projection PCA algorithm in step 3 of Algorithm 2 is because the error $e_t = \hat{L}_t - L_t = S_t - \hat{S}_t$ is correlated with L_t; and the maximum condition number of Cov(L_t), which is bounded by f, cannot be bounded by a small value (see Remark 3.4). This issue is explained in detail in Appendix X. Most other works that analyze standard PCA, see [33] and references therein, do not face this issue because they assume uncorrelated-ness of the noise/error and the true data vector. With this assumption, one only needs to increase the PCA data length α to deal with the larger condition number.

We should mention that the idea of projecting perpendicular to a partly estimated subspace has been used in other different contexts in past work [14] and [41].

C. Recursive Projected CS (ReProCS)

We summarize the Recursive Projected CS (ReProCS) algorithm in Algorithm 2. It uses the following definition.

Definition 4.4: Define the time interval $\mathcal{I}_{j,k} := [t_j + (k-1)\alpha, t_j + k\alpha - 1]$ for $k = 1, \ldots K$ and $\mathcal{I}_{j,K+1} := [t_j + K\alpha, t_{j+1} - 1]$.

The key idea of ReProCS is as follows. First, consider a time t when the current basis matrix $P_{(t)} = P_{(t-1)}$ and this 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 the measurement vector, 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 $r_* = \text{rank}(\hat{P}_{(t-1)})$, therefore y_t has only $n - r_*$ "effective" measurements[2], even though its

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

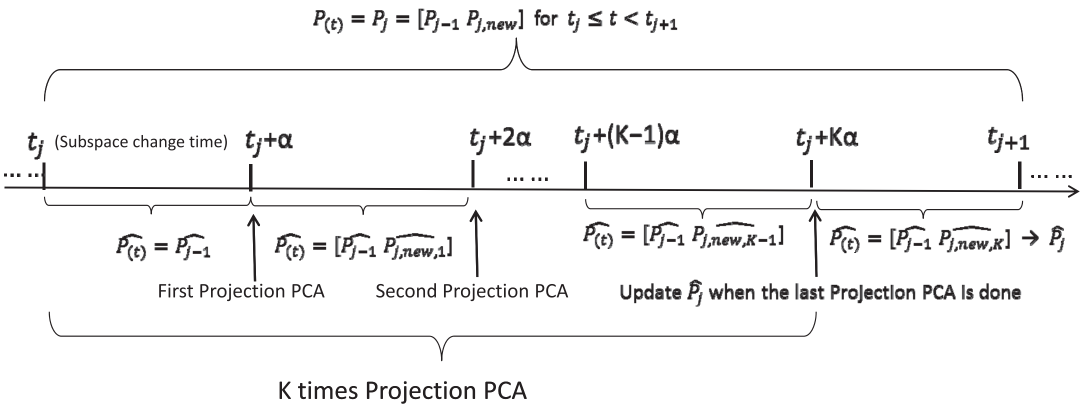

Fig. 2. The K projection PCA steps.

length is 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 $\|(I - \hat{P}_{(t-1)} \hat{P}'_{(t-1)}) P_{(t-1)}\|_2$ 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 [34], [37], [42]–[45]. We use ℓ_1 minimization to recover it (step 1b). If the current basis matrix $P_{(t)}$, and hence its estimate, $\hat{P}_{(t-1)}$, is dense enough, then, by Lemma 3.7, the RIC of $\Phi_{(t)}$ is small enough. Using Theorem 2.5, 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 [46]. This $\hat{S}_t$ is used to estimate L_t as $\hat{L}_t = M_t - \hat{S}_t$. As we explain in the proof of Lemma 6.4, if $S_{\min}$ is large enough and the support estimation threshold, ω, is chosen appropriately, we can get exact support recovery, i.e. $\hat{T}_t = T_t$. In this case, the error $e_t := \hat{S}_t - S_t = L_t - \hat{L}_t$ has the following simple expression:

$$e_t = I_{T_t} (\Phi_{(t)})_{T_t}{}^{\dagger} \beta_t = I_{T_t} [(\Phi_{(t)})'_{T_t} (\Phi_{(t)})_{T_t}]^{-1} I_{T_t}{}' \Phi_{(t)} L_t \quad (10)$$

The second equality follows because $(\Phi_{(t)})_T{}' \Phi_{(t)} = (\Phi_{(t)} I_T)' \Phi_{(t)} = I_T{}' \Phi_{(t)}$ for any set T.

Now consider a time t when $P_{(t)} = P_j = [P_{j-1}, P_{j,\text{new}}]$ and P_{j-1} has been accurately estimated but $P_{j,\text{new}}$ has not been estimated, i.e. consider a $t \in \mathcal{I}_{j,1}$. At this time, $\hat{P}_{(t-1)} = \hat{P}_{j-1}$ and so $\Phi_{(t)} = \Phi_{j,0} := I - \hat{P}_{j-1} \hat{P}'_{j-1}$. Let $r_* := r_0 + (j-1) c_{\max}$ (We remove subscript j for ease of notation.) , and $c := c_{\max}$. Assume that the delay between change times is large enough so that by $t = t_j$, $\hat{P}_{j-1}$ is an accurate enough estimate of P_{j-1}, i.e. $\|\Phi_{j,0} P_{j-1}\|_2 \le r_* \zeta \ll 1$. It is easy to see using Lemma 2.10 that $\kappa_s(\Phi_0 P_{\text{new}}) \le \kappa_s(P_{\text{new}}) + r_* \zeta$, i.e. $\Phi_0 P_{\text{new}}$ is dense because P_{new} is dense and because $\hat{P}_{j-1}$ is an accurate estimate of P_{j-1} (which is perpendicular to P_{new}). Moreover, using Lemma 3.7, it can be shown that $\phi_0 := \max_{|T| \le s} \|[(\Phi_0)'_T (\Phi_0)_T]^{-1}\|_2 \le \frac{1}{1-\delta_s(\Phi_0)} \le \frac{1}{1-(\kappa_s(P_{j-1})+r_*\zeta)^2}$. The error e_t still satisfies (10) although its magnitude is not as small. Using the above facts in (10), we get that

$$\|e_t\|_2 \le \frac{\kappa_s(P_{\text{new}}) \sqrt{c} \gamma_{\text{new}} + r_* \zeta (\sqrt{r_*} \gamma_* + \sqrt{c} \gamma_{\text{new}})}{1 - (\kappa_s(P_{j-1}) + r\zeta)^2}$$

If $\sqrt{\zeta} < 1/\gamma_*$, all terms containing ζ can be ignored and we get that the above is approximately upper bounded by $\frac{\kappa_s(P_{\text{new}})}{1-\kappa_s^2(P_{j-1})} \sqrt{c} \gamma_{\text{new}}$. Using the denseness assumption, this quantity is a small constant times $\sqrt{c} \gamma_{\text{new}}$, e.g. with the numbers assumed in Theorem 4.2 we get a bound of $0.18 \sqrt{c} \gamma_{\text{new}}$. Since $\gamma_{\text{new}} \ll S_{\min}$ and c is assumed to be small, thus, $\|e_t\|_2 = \|S_t - \hat{S}_t\|_2$ is small compared with $\|S_t\|_2$, i.e. S_t is recovered accurately. With each projection PCA step, as we explain below, the error e_t becomes even smaller.

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). We illustrate the projection PCA algorithm in Figure 2. In the first projection PCA step, we get the first estimate of $P_{j,\text{new}}$, $\hat{P}_{j,\text{new},1}$. For the next α frame interval, $\hat{P}_{(t-1)} = [\hat{P}_{j-1}, \hat{P}_{j,\text{new},1}]$ and so $\Phi_{(t)} = \Phi_{j,1} = I - \hat{P}_{j-1} \hat{P}'_{j-1} - \hat{P}_{\text{new},1} \hat{P}'_{\text{new},1}$. Using this in the projected CS step reduces the projection noise, β_t, and hence the reconstruction error, e_t, for this interval, as long as $\gamma_{\text{new},k}$ increases slowly enough. Smaller e_t makes the perturbation seen by the second projection PCA step even smaller, thus resulting in an improved second estimate $\hat{P}_{j,\text{new},2}$. Within K updates (K chosen as given in Theorem 4.2), it can be shown that both $\|e_t\|_2$ and the subspace error drop down to a constant times $\sqrt{\zeta}$. At this time, we update $\hat{P}_j$ as $\hat{P}_j = [\hat{P}_{j-1}, \hat{P}_{j,\text{new},K}]$.

D. Discussion

First consider the choices of α and of K. 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. This means that we also require a larger delay between subspace change times, i.e. larger $t_{j+1} - t_j$.

Now consider the assumptions used in the result. We assume slow subspace change, i.e. the delay between change times is large enough, $\|a_{t,\text{new}}\|_\infty$ is initially below γ_{new} and increases gradually, and $14\xi_0 \le S_{\min}$ which holds if $c_{\max}$ and γ_{new} are small enough. Small $c_{\max}$, small initial $a_{t,\text{new}}$ (i.e. small γ_{new}) and its gradual increase are verified for real video data in Section IX-A. As explained there, one cannot estimate the delay between change times unless one has access to an ensemble of videos of a given type and hence the first assumption cannot be verified.

We also assume denseness of P_{j-1} and $P_{j,\text{new}}$. This is a subset of the denseness assumptions used in earlier work [6]. As explained there, this is valid for the video application because typically the changes of the background sequence are global, e.g. due to illumination variation affecting the entire image or due to textural changes such as water motion or tree leaves' motion etc. We quantify this denseness using the parameter κ_s. The way it is defined, bounds on κ_s simultaneously place restrictions on denseness of L_t, $r = \text{rank}(P_J)$, and s (the maximum sparsity of any S_t). To compare our assumptions with those of Candès et. al. in [6], we could assume $\kappa_1(P_J) \leq \sqrt{\frac{\mu r}{n}}$, where μ is any value between 1 and $\frac{n}{r}$. Using the bound $\kappa_s(P) \leq \sqrt{s}\kappa_1(P)$, we see that if $\frac{2sr}{n} \leq \mu^{-1}(0.3)^2$, then our assumption of $\kappa_{2s}(P_J) \leq 0.3$ will be satisfied. Up to differences in the constants, this is the same requirement found in [47], even though [47] studies a batch approach (PCP) while we study an online algorithm. From this we can see that if s grows linearly with n, then r must be constant. Similarly, if r grows linearly with n, then s must be constant. This is a stronger assumption than required by [6] where s is allowed to grow linearly with n, and r is simultaneously allowed to grow as $\frac{n}{\log(n)^2}$. However, the comparison with [6] is not direct because we do not need denseness of the right singular vectors or a bound on the vector infinity norm of UV'. The reason for the stronger requirement on the product sr is because we study an online algorithm that recovers the sparse vector S_t at each time t rather than in a batch or a piecewise batch fashion. Because of this the sparse recovery step does not use the low dimensional structure of the new (and still unestimated) subspace.

We assume the independence of a_t's, and hence of L_t's, over time. This is typically not valid in practice; however, it allows us to simplify the problem and hence the derivation of the performance guarantees. In particular it allows us to use the matrix Hoeffding inequality to bound the terms in the subspace error bound. In ongoing work by Zhan and Vaswani [48], we are seeing that, with some more work, this can be replaced by a more realistic assumption: an autoregressive model on the a_t's, i.e. assume $a_t = ba_{t-1} + \nu_t$ where ν_t's are independent over time and $b < 1$. We can work with this model in two ways. If we assume b is known, then a simple change to the algorithm (in the subspace update step, replace $\hat{L}_t$ by $\hat{L}_t - b\hat{L}_{t-1}$ everywhere) allows us to get a result that is almost the same as the current one using exactly the same approach. Alternatively if b is unknown, as long as b is bounded by a $b_* < 1$, we can use the matrix Azuma inequality to still get a result similar to the current one. It will require a larger α though and some other changes.

The most limiting assumption is the assumption on $D_{j,\text{new},k}$ and $Q_{j,\text{new},k}$ because these are functions of algorithm estimates. The denseness assumption on $Q_{j,\text{new},k}$ is actually not essential, it is possible to prove a slightly more complicated version of Theorem 4.2 without it. We use this assumption only in Lemma 6.6. However, if we use tighter bounds on other quantities such as g and $\kappa_s(P_{j,\text{new}})$, and if we analyze the first projection-PCA step differently from the others, we can get a tighter bound on $\zeta_{j,1}$ (and hence $\zeta_{j,k}$ for $k \geq 1$) and then we will not need this assumption.

Consider denseness of $D_{j,\text{new},k}$. Our proof actually only needs smallness of $\max_{t \in \mathcal{I}_{j,k+1}} d_t$ where $d_t = \|I_{T_t}{}' D_{j,\text{new},k}\|_2 / \|D_{j,\text{new},k}\|_2$ for $t \in \mathcal{I}_{j,k+1}$ for $k = 1, 2 \dots K$. Since this quantity is upper bounded by $\kappa_s(D_{j,\text{new},k})$, we have just assumed a bound on this for simplicity. Note also that densenss of $D_{j,\text{new},0}$ does not need to be assumed, this follows from denseness of $P_{j,\text{new}}$ conditioned on the fact that P_{j-1} has been accurately estimated. We attempted to verify the smallness of d_t in simulations done with a dense P_j and $P_{j,\text{new}}$ and involving correlated support change of S_t's. We observed that, as long as there was a support change every few frames, this quantity was small. For example, with $n = 2048$, $s = 20$, $r_0 = 36$, $c_{\text{new}} = 1$, support change by one index every 2 frames was sufficient to ensure a small d_t at all times (see Sec IX-B). Even one index change every 50 frames was enough to ensure that the errors decayed down to small enough values, although in this case d_t was large at certain times and the decay of the subspace error was not exponential. It should be possible to use a similar idea to modify our result as well. The first thing to point out is that the max of d_t can be replaced by its average over $t \in \mathcal{I}_{j,k}$ with a minor change to the proof of Lemma 6.11. Moreover, if we try to show linear decay of the subspace error (instead of exponential decay), and if we analyze the first projection-PCA interval differently from the others, we will need a looser bound on the d_t's, which will be easier to obtain under a certain support change assumption. In the first interval, the subspace error is large since P_{new} has not been estimated but $D_{\text{new},0}$ is dense (see Remark 4.3). In the later intervals, the subspace error is lower but $D_{\text{new},k}$ may not be as dense.

Finally, Algorithm 2 assumes knowledge of certain model parameters and these may not always be available. It needs to know $c_{j,\text{new}}$, which is the number of new directions added at subspace change time j, and it needs knowledge of γ_{new} (in order to set ξ and ω), which is the bound on the infinity norm of the projection of a_t along the new directions for the first α frames. It also needs to know the subspace change times t_j, and this is the most restrictive.

A practical version of Algorithm 2 (that provides reasonable heuristics for setting its parameters without model knowledge) is given in [26]. As explained there, $\hat{t}_j + \alpha - 1$ can be estimated by taking the last set of α estimates $\hat{L}_t$, projecting them perpendicular to $\hat{P}_{j-1}$ and checking if any of the singular values of the resulting matrix is above $\sqrt{\hat{\lambda}^-}$. It should be possible to prove in future work that this happens only after an actual change and within a short delay of it.

Lastly, note that, because the subspace change model only allows new additions to the subspace, the rank of the subspace basis matrix P_j can only grow over time. The same is true for its ReProCS estimate. Thus, $\max_j \kappa_{2s}(P_j) = \kappa_{2s}(P_J)$ and a bound on this imposes a bound on the number of allowed subspace change times, J, or equivalently on the maximum rank of $\mathcal{L}_t$ for any t. A similar bound is also needed by PCP [6] and all batch approaches. In Sec VII, we explain how we can remove the bound on J and hence on the rank of $\mathcal{L}_t$ if an extra clustering assumption holds.

V. Definitions Needed for Proving Theorem 4.2

A few quantities are already defined in the model (Section III-A), Definition 4.4, Algorithm 2, and Theorem 4.2. Here we define more quantities needed for the proofs.

Definition 5.1: In the sequel, we let

1) $r := r_{\max} = r_0 + Jc_{\max}$ *and* $c := c_{\max} = \max_j c_{j,\text{new}}$,
2) $\kappa_{s,*} := \max_j \kappa_s(P_{j-1})$, $\kappa_{s,\text{new}} := \max_j \kappa_s(P_{j,\text{new}})$, $\kappa_{s,k} := \max_j \kappa_s(D_{j,\text{new},k})$, $\tilde{\kappa}_{s,k} := \max_j \kappa_s((I - P_{j,\text{new}}P_{j,\text{new}}')\hat{P}_{j,\text{new},k})$,
3) $\kappa_{2s,*}^+ := 0.3$, $\kappa_{2s,\text{new}}^+ := 0.15$, $\kappa_s^+ := 0.152$, $\tilde{\kappa}_{2s}^+ := 0.15$ *and* $g^+ := \sqrt{2}$ *are the upper bounds assumed in Theorem 4.2 on* $\max_j \kappa_{2s}(P_j)$, $\max_j \kappa_{2s}(P_{j,\text{new}})$, $\max_j \max_k \kappa_s(D_{j,\text{new},k})$, $\max_j \kappa_{2s}(Q_{j,\text{new},k})$ *and* g *respectively.*
4) $\phi^+ := 1.1735$
5) $\gamma_{\text{new},k} := \min(1.2^{k-1}\gamma_{\text{new}}, \gamma_*)$ *(recall that this is defined in Sec III-B).*

Definition 5.2: Define the following:

1) $\zeta_{j,*}^+ := (r_0 + (j-1)c)\zeta$
2) *Define the sequence* $\{\zeta_{j,k}{}^+\}_{k=0,1,2,\dots,K}$ *recursively as follows:*

$$\zeta_{j,0}^+ := 1$$
$$\zeta_{j,k}^+ := \frac{b + 0.125c\zeta}{1 - (\zeta_{j,*}^+)^2 - (\zeta_{j,*}^+)^2 f - 0.125c\zeta - b} \quad \text{for } k \geq 1, \tag{11}$$

where

$$b := C\kappa_s^+ g^+ \zeta_{j,k-1}^+ + \tilde{C}(\kappa_s^+)^2 g^+ (\zeta_{k-1}^+)^2 + C' f (\zeta_{j,*}^+)^2$$
$$C := \frac{2\kappa_s^+\phi^+}{\sqrt{1-(\zeta_{j,*}^+)^2}} + \phi^+,$$
$$C' := (\phi^+)^2 + \frac{2\phi^+}{\sqrt{1-(\zeta_{j,*}^+)^2}} + 1 + \phi^+ + \frac{\kappa_s^+\phi^+}{\sqrt{1-(\zeta_{j,*}^+)^2}} + \frac{\kappa_s^+(\phi^+)^2}{\sqrt{1-(\zeta_{j,*}^+)^2}},$$
$$\tilde{C} := (\phi^+)^2 + \frac{\kappa_s^+(\phi^+)^2}{\sqrt{1-(\zeta_{j,*}^+)^2}}.$$

As we will see, $\zeta_{j,*}^+$ *and* $\zeta_{j,k}^+$ *are the high probability upper bounds on* $\zeta_{j,*}$ *and* $\zeta_{j,k}$ *(defined in Definition 5.4) under the assumptions of Theorem 4.2.*

Definition 5.3: We define the noise seen by the sparse recovery step at time t *as*

$$\beta_t := (I - \hat{P}_{(t-1)}\hat{P}_{(t-1)}')L_t.$$

Also define the reconstruction error of S_t *as*

$$e_t := \hat{S}_t - S_t.$$

Here $\hat{S}_t$ *is the final estimate of* S_t *after the LS step in Algorithm 2. Notice that* e_t *also satisfies* $e_t = L_t - \hat{L}_t$.

Definition 5.4: We define the subspace estimation errors as follows. Recall that $\hat{P}_{j,\text{new},0} = [.]$ *(empty matrix).*

$$\text{SE}_{(t)} := \|(I - \hat{P}_{(t)}\hat{P}_{(t)}')P_{(t)}\|_2,$$
$$\zeta_{j,*} := \|(I - \hat{P}_{j-1}\hat{P}_{j-1}')P_{j-1}\|_2$$
$$\zeta_{j,k} := \|(I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,\text{new},k}\hat{P}_{j,\text{new},k}')P_{j,\text{new}}\|_2$$

Remark 5.5: Recall from the model given in Sec III-A and from Algorithm 2 that

1) $\hat{P}_{j,\text{new},k}$ *is orthogonal to* $\hat{P}_{j-1}$, *i.e.* $\hat{P}_{j,\text{new},k}'\hat{P}_{j-1} = 0$
2) $\hat{P}_{j-1} := [\hat{P}_0, \hat{P}_{1,\text{new},K}, \dots \hat{P}_{j-1,\text{new},K}]$ *and* $P_{j-1} := [P_0, P_{1,\text{new}}, \dots P_{j-1,\text{new}}]$
3) *for* $t \in \mathcal{I}_{j,k+1}$, $\hat{P}_{(t)} = [\hat{P}_{j-1}, \hat{P}_{j,\text{new},k}]$ *and* $P_{(t)} = P_j = [P_{j-1}, P_{j,\text{new}}]$.
4) $\Phi_{(t)} := I - \hat{P}_{(t-1)}\hat{P}_{(t-1)}'$

Then it is easy to see that

1) $\zeta_{j,*} \leq \zeta_{j-1,*} + \zeta_{j,K} = \zeta_{1,*} + \sum_{j'=1}^{j-1}\zeta_{j',K}$
2) $\text{SE}_{(t)} \leq \zeta_{j,*} + \zeta_{j,k} \leq \zeta_{1,*} + \sum_{j'=1}^{j-1}\zeta_{j',K} + \zeta_{j,k}$ for $t \in \mathcal{I}_{j,k+1}$.

Definition 5.6: Define the following

1) $\Phi_{j,k}$, $\Phi_{j,0}$ *and* ϕ_k
   a) $\Phi_{j,k} := I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,\text{new},k}\hat{P}_{j,\text{new},k}'$ *is the CS matrix for* $t \in \mathcal{I}_{j,k+1}$, *i.e.* $\Phi_{(t)} = \Phi_{j,k}$ *for this duration.*
   b) $\Phi_{j,0} := I - \hat{P}_{j-1}\hat{P}_{j-1}'$ *is the CS matrix for* $t \in \mathcal{I}_{j,1}$, *i.e.* $\Phi_{(t)} = \Phi_{j,0}$ *for this duration.* $\Phi_{j,0}$ *is also the projection matrix used in all of the projection PCA steps for* $t \in [t_j, t_{j+1} - 1]$.
   c) $\phi_k := \max_j \max_{T:|T|\leq s} \|((\Phi_{j,k})_T'(\Phi_{j,k})_T)^{-1}\|_2$. *It is easy to see that* $\phi_k \leq \frac{1}{1-\max_j \delta_s(\Phi_{j,k})}$ *[34].*
2) $D_{j,\text{new},k}$, $D_{j,\text{new}}$, $D_{j,*,k}$ *and* $D_{j,*}$
   a) $D_{j,\text{new},k} := \Phi_{j,k}P_{j,\text{new}}$. $\text{span}(D_{j,\text{new},k})$ *is the unestimated part of the newly added subspace for any* $t \in \mathcal{I}_{j,k+1}$.
   b) $D_{j,\text{new}} := D_{j,\text{new},0} = \Phi_{j,0}P_{j,\text{new}}$. $\text{span}(D_{j,\text{new}})$ *is interpreted similarly for any* $t \in \mathcal{I}_{j,1}$.
   c) $D_{j,*,k} := \Phi_{j,k}P_{j-1}$. $\text{span}(D_{j,*,k})$ *is the unestimated part of the existing subspace for any* $t \in \mathcal{I}_{j,k}$
   d) $D_{j,*} := D_{j,*,0} = \Phi_{j,0}P_{j-1}$. $\text{span}(D_{j,*,k})$ *is interpreted similarly for any* $t \in \mathcal{I}_{j,1}$
   e) *Notice that* $\zeta_{j,0} = \|D_{j,\text{new}}\|_2$, $\zeta_{j,k} = \|D_{j,\text{new},k}\|_2$, $\zeta_{j,*} = \|D_{j,*}\|_2$. *Also, clearly,* $\|D_{j,*,k}\|_2 \leq \zeta_{j,*}$.

Definition 5.7:

1) *Let* $D_{j,\text{new}} \overset{QR}{=} E_{j,\text{new}}R_{j,\text{new}}$ *denote its reduced QR decomposition, i.e. let* $E_{j,\text{new}}$ *be a basis matrix for* $\text{span}(D_{j,\text{new}})$ *and let* $R_{j,\text{new}} = E_{j,\text{new}}'D_{j,\text{new}}$.
2) *Let* $E_{j,\text{new},\perp}$ *be a basis matrix for the orthogonal complement of* $\text{span}(E_{j,\text{new}}) = \text{span}(D_{j,\text{new}})$. *To be precise,* $E_{j,\text{new},\perp}$ *is a* $n \times (n - c_{j,\text{new}})$ *basis matrix that satisfies* $E_{j,\text{new},\perp}'E_{j,\text{new}} = 0$.

3) *Using* $E_{j,\text{new}}$ *and* $E_{j,\text{new},\perp}$, *define* $A_{j,k}$, $A_{j,k,\perp}$, $H_{j,k}$, $H_{j,k,\perp}$ *and* $B_{j,k}$ *as*

$$A_{j,k} := \frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}} E_{j,\text{new}}{}'\Phi_{j,0}L_t L_t{}'\Phi_{j,0}E_{j,\text{new}}$$

$$A_{j,k,\perp} := \frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}} E_{j,\text{new},\perp}{}'\Phi_{j,0}L_t L_t{}'\Phi_{j,0}E_{j,\text{new},\perp}$$

$$H_{j,k} := \frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}} E_{j,\text{new}}{}'\Phi_{j,0}(e_t e_t{}' - L_t e_t{}' - e_t L_t{}')\Phi_{j,0}E_{j,\text{new}}$$

$$H_{j,k,\perp} := \frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}} E_{j,\text{new},\perp}{}'\Phi_{j,0}(e_t e_t{}' - L_t e_t{}' - e_t L_t{}')\Phi_{j,0}E_{j,\text{new},\perp}$$

$$B_{j,k} := \frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}} E_{j,\text{new},\perp}{}'\Phi_{j,0}\hat{L}_t\hat{L}_t'\Phi_{j,0}E_{j,\text{new}} = \frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}} E_{j,\text{new},\perp}{}'\Phi_{j,0}(L_t - e_t)(L_t{}' - e_t{}')\Phi_{j,0}E_{j,\text{new}}$$

4) *Define*

$$\mathcal{A}_{j,k} := \begin{bmatrix} E_{j,\text{new}} & E_{j,\text{new},\perp}\end{bmatrix}\begin{bmatrix} A_{j,k} & 0 \\ 0 & A_{j,k,\perp}\end{bmatrix}\begin{bmatrix} E_{j,\text{new}}{}' \\ E_{j,\text{new},\perp}{}'\end{bmatrix}$$

$$\mathcal{H}_{j,k} := \begin{bmatrix} E_{j,\text{new}} & E_{j,\text{new},\perp}\end{bmatrix}\begin{bmatrix} H_{j,k} & B_{j,k}{}' \\ B_{j,k} & H_{j,k,\perp}\end{bmatrix}\begin{bmatrix} E_{j,\text{new}}{}' \\ E_{j,\text{new},\perp}{}'\end{bmatrix}$$

Remark 5.8: 1) *From the above, it is easy to see that*

$$\mathcal{A}_{j,k} + \mathcal{H}_{j,k} = \frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}} \Phi_{j,0}\hat{L}_t\hat{L}_t'\Phi_{j,0}.$$

2) *Recall from Algorithm 2 that*

$$\mathcal{A}_{j,k} + \mathcal{H}_{j,k} \overset{EVD}{=} \begin{bmatrix} \hat{P}_{j,\text{new},k} & \hat{P}_{j,\text{new},k,\perp}\end{bmatrix}\begin{bmatrix} \Lambda_k & 0 \\ 0 & \Lambda_{k,\perp}\end{bmatrix}\begin{bmatrix} \hat{P}_{j,\text{new},k}' \\ \hat{P}_{j,\text{new},k,\perp}'\end{bmatrix}$$

is the EVD of $\mathcal{A}_{j,k} + \mathcal{H}_{j,k}$.

3) *Using the above,* $\mathcal{A}_{j,k} + \mathcal{H}_{j,k}$ *can be decomposed in two ways as follows:*

$$\begin{aligned}\mathcal{A}_{j,k} + \mathcal{H}_{j,k} &= \begin{bmatrix} \hat{P}_{j,\text{new},k} & \hat{P}_{j,\text{new},k,\perp}\end{bmatrix}\begin{bmatrix} \Lambda_k & 0 \\ 0 & \Lambda_{k,\perp}\end{bmatrix}\begin{bmatrix} \hat{P}_{j,\text{new},k}' \\ \hat{P}_{j,\text{new},k,\perp}'\end{bmatrix} \\ &= \begin{bmatrix} E_{j,\text{new}} & E_{j,\text{new},\perp}\end{bmatrix}\begin{bmatrix} A_{j,k}+H_{j,k} & B_{j,k}' \\ B_{j,k} & A_{j,k,\perp}+H_{j,k,\perp}\end{bmatrix}\begin{bmatrix} E_{j,\text{new}}{}' \\ E_{j,\text{new},\perp}{}'\end{bmatrix}\end{aligned}$$

Definition 5.9: Define the random variable $X_{j,k} := \{a_1, a_2, \cdots, a_{t_j+k\alpha-1}\}$.

Recall that the a_t's are mutually independent over t, hence $X_{j,k}$ and $\{a_{t_j+k\alpha}, \ldots, a_{t_j+(k+1)\alpha-1}\}$ are mutually independent.

Definition 5.10: Define the set $\check{\Gamma}_{j,k}$ *as follows:*

$$\check{\Gamma}_{j,k} := \{X_{j,k} : \zeta_{j,k} \le \zeta_k^+ \text{ and } \hat{T}_t = T_t \text{ for all } t\in\mathcal{I}_{j,k}\}$$
$$\check{\Gamma}_{j,K+1} := \{X_{j+1,0} : \hat{T}_t = T_t \text{ for all } t\in\mathcal{I}_{j,K+1}\}$$

Definition 5.11: Recursively define the sets $\Gamma_{j,k}$ *as follows:*

$$\begin{aligned}\Gamma_{1,0} &:= \{X_{1,0} : \zeta_{1,*} \le r\zeta \text{ and } \hat{T}_t = T_t \text{ for all } t\in[t_{\text{train}}+1 : t_1 - 1]\} \\ \Gamma_{j,0} &:= \{X_{j,0} : \zeta_{j',*} \le \zeta_{j',*}^+ \text{ for all } j' = 1,2,\ldots,j \text{ and } \hat{T}_t = T_t \text{ for all } t \le t_{j-1}\} \\ \Gamma_{j,k} &:= \Gamma_{j,k-1}\cap\check{\Gamma}_{j,k}\ \ k = 1,2,\ldots K+1\end{aligned}$$

Remark 5.12: Whenever $\hat{T}_t = T_t$ *we have an exact expression for* e_t:

$$e_t = I_{T_t}[(\Phi_{(t)})_{T_t}'(\Phi_{(t)})_{T_t}]^{-1}I_{T_t}{}'\Phi_{(t)}L_t \tag{12}$$

Recall that $L_t = P_j a_t = P_{j-1}a_{t,*} + P_{j,\text{new}}a_{t,\text{new}}$.

Definition 5.13: Define $P_{j,*} := P_{j-1}$ *and* $\hat{P}_{j,*} := \hat{P}_{j-1}$.

Remark 5.14: Notice that the subscript j always appears as the first subscript, while k is the last one. At many places in the rest of the paper, we remove the subscript j for simplicity, e.g., Φ_0 *refers to* $\Phi_{j,0}$, $\hat{P}_{\text{new},k}$ *refers to* $\hat{P}_{j,\text{new},k}$, P_* *refers to* $P_{j,*} := P_{j-1}$ *and so on.*

VI. Proof of Theorem 4.2

A. Two Main Lemmas and Proof of Theorem 4.2

The proof of Theorem 4.2 essentially follows from two main lemmas that we state below. Lemma 6.1 gives an exponentially decaying upper bound on ζ_k^+ defined in Definition 5.2. ζ_k^+ will be shown to be a high probability upper bound for ζ_k under the assumptions of the Theorem. Lemma 6.2 says that conditioned on $X_{j,k-1}\in\Gamma_{j,k-1}$, $X_{j,k}$ will be in $\Gamma_{j,k}$ w.h.p.. In words this says that if, during the time interval $\mathcal{I}_{j,k-1}$, the algorithm has worked well (recovered the support of S_t exactly and recovered the background subspace with subspace recovery error below $\zeta_{k-1}^+ + \zeta_*^+$), then it will also work well in $\mathcal{I}_{j,k}$ w.h.p..

Lemma 6.1 (Exponential Decay of ζ_k^+*): Assume that the bounds on* ζ *from Theorem 4.2 hold. Define the sequence* ζ_k^+ *as in Definition 5.2. Then*

1) $\zeta_0^+ = 1$ *and* $\zeta_k^+ \le 0.6^k + 0.4c\zeta$ *for all* $k = 1,2,\ldots,K$,
2) *the denominator of* ζ_k^+ *is positive for all* $k = 1,2,\ldots,K$.

We prove this lemma in Section VI-B.

Lemma 6.2: Assume that all the conditions of Theorem 4.2 hold. Also assume that $\mathbf{P}(\Gamma_{j,k-1}^e) > 0$. *Then*

$$\mathbf{P}(\Gamma_{j,k}^e|\Gamma_{j,k-1}^e) \ge p_k(\alpha,\zeta) \ge p_K(\alpha,\zeta)$$

for all $k = 1,2,\ldots,K$, *and*

$$\mathbf{P}(\Gamma_{j,K+1}^e|\Gamma_{j,K}^e) = 1$$

where $p_k(\alpha,\zeta)$ *is defined in equation* (13).

We prove this lemma in Section VI-C.

Remark 6.3: Using Lemma 6.1 and Remark 5.5 and the value of K given in the theorem, it is easy to see that, under the assumptions of Theorem 4.2,

$$\Gamma_{j,0}\cap(\cap_{k=1}^{K+1}\check{\Gamma}_{j,k}) \subseteq \Gamma_{j+1,0}.$$

Thus $\mathbf{P}(\Gamma_{j+1,0}^e|\Gamma_{j,0}^e) \ge \mathbf{P}(\check{\Gamma}_{j,1}^e, \ldots \check{\Gamma}_{j,K+1}^e|\Gamma_{j,0}^e)$.

Proof of Theorem 4.2:

The theorem is a direct consequence of Lemmas 6.1, 6.2, and Lemma 2.12. From Remark 6.3, $\mathbf{P}(\Gamma_{j+1,0}^e|\Gamma_{j,0}^e) \ge$

$\mathbf{P}(\check{\Gamma}^e_{j,1},\ldots\check{\Gamma}^e_{j,K+1}|\Gamma^e_{j,0}) = \prod_{k=1}^{K+1} P(\check{\Gamma}^e_{j,k}|\Gamma^e_{j,k-1})$. Also, since $\Gamma_{j+1,0} \subseteq \Gamma_{j,0}$, using Lemma 2.12, $\mathbf{P}(\Gamma^e_{J+1,0}|\Gamma^e_{1,0}) = \prod_{j=1}^{J}\mathbf{P}(\Gamma^e_{j+1,0}|\Gamma^e_{j,0})$. Thus,

$$\mathbf{P}(\Gamma^e_{J+1,0}|\Gamma^e_{1,0}) \geq \prod_{j=1}^{J}\prod_{k=1}^{K+1}\mathbf{P}(\check{\Gamma}^e_{j,k}|\Gamma^e_{j,k-1})$$

Using Lemma 6.2, and the fact that $p_k(\alpha,\zeta) \geq p_K(\alpha,\zeta)$ (see their respective definitions in Lemma 6.11 and equation (13) and observe that $p_k(\alpha,\zeta)$ is decreasing in k), we get

$$\mathbf{P}(\Gamma^e_{J+1,0}|\Gamma_{1,0}) \geq p_K(\alpha,\zeta)^{KJ}.$$

Also, $\mathbf{P}(\Gamma^e_{1,0}) = 1$. This follows by the assumption on $\hat{P}_0$ and Lemma 6.4. Thus, $\mathbf{P}(\Gamma^e_{J+1,0}) \geq p_K(\alpha,\zeta)^{KJ}$.

Using the definition of α_{add}, and $\alpha \geq \alpha_{\text{add}}$, we get that

$$\mathbf{P}(\Gamma^e_{J+1,0}) \geq p_K(\alpha,\zeta)^{KJ} \geq 1 - n^{-10}$$

The event $\Gamma^e_{J+1,0}$ implies that $\hat{T}_t = T_t$ and e_t satisfies (10) for all $t < t_{J+1}$. Using Remarks 5.5 and 6.3, $\Gamma^e_{J+1,0}$ implies that all the bounds on the subspace error hold. Using these, $\|a_{t,\text{new}}\|_2 \leq \sqrt{c}\gamma_{\text{new},k}$, and $\|a_t\|_2 \leq \sqrt{r}\gamma_*$, $\Gamma^e_{J+1,0}$ implies that all the bounds on $\|e_t\|_2$ hold (the bounds are obtained in Lemma 6.4).

Thus, all conclusions of the the result hold w.p. at least $1 - n^{-10}$. ■

B. Proof of Lemma 6.1

Proof: First recall the definition of ζ_k^+ (Definition 5.2). Recall from Definition 5.1 that $\kappa_s^+ := 0.15$, $\phi^+ := 1.1735$, and $g^+ := \sqrt{2}$. So we can make these substitutions directly. Notice that ζ_k^+ is an increasing function of $\zeta_*^+, \zeta, c,$ and f. Therefore we can use upper bounds on each of these quantities to get an upper bound on ζ_k^+. From the definition of ζ in Theorem 4.2 and $\zeta_{j,*}^+ := (r_0 + (j-1)c)\zeta$ we get

- $\zeta_{j,*}^+ \leq 10^{-4}$
- $\zeta_{j,*}^+ f \leq 1.5 \times 10^{-4}$
- $c\zeta \leq 10^{-4}$
- $\dfrac{\zeta_{j,*}^+}{c\zeta} = \dfrac{(r_0 + (j-1)c)\zeta}{c\zeta} \leq \dfrac{r_0 + (J-1)c}{c} = \dfrac{r}{c} \leq r$
  (Without loss of generality we can assume that $c = c_{\max} \geq 1$ because if $c = 0$ then there is no subspace estimation problem to be solved. $c = 0$ is the trivial case where all conclusions of Theorem 4.2 will hold just using Lemma 6.4.)
- $\zeta_{j,*}^+ f r \leq r^2 f\zeta \leq 1.5 \times 10^{-4}$

First we prove by induction that $\zeta_k^+ \leq \zeta_{k-1}^+ \leq 0.6$ for all $k \geq 1$. Notice that $\zeta_0^+ = 1$ by definition.

- Base case ($k = 1$): Using the above bounds we get that $\zeta_1^+ < 0.5985 < 1 = \zeta_0^+$.
- For the induction step, assume that $\zeta_{k-1}^+ \leq \zeta_{k-2}^+$. Then because ζ_k^+ is increasing in ζ_{k-1}^+ (denote the increasing function by f_{inc}) we get that $\zeta_k^+ = f_{inc}(\zeta_{k-1}^+) \leq f_{inc}(\zeta_{k-2}^+) = \zeta_{k-1}^+$.

1) To prove the first claim, first rewrite ζ_k^+ as

$$\zeta_k^+ = \zeta_{k-1}^+ \frac{C\kappa_s^+ g^+ + \tilde{C}(\kappa_s^+)^2 g^+(\zeta_{k-1}^+)}{1 - (\zeta_*^+)^2 - (\zeta_*^+)^2 f - 0.125c\zeta - b} + c\zeta \frac{C(\zeta_*^+ f)\frac{(\zeta_*^+)}{c\zeta} + .125}{1 - (\zeta_*^+)^2 - (\zeta_*^+)^2 f - 0.125c\zeta - b}$$

where $C, \tilde{C}$, and b are as in Definition 5.2. Using the above bounds including $\zeta_{k-1}^+ \leq .6$ we get that

$$\begin{aligned}\zeta_k^+ &\leq \zeta_{k-1}^+(0.6) + c\zeta(0.16)\\ &= \zeta_0^+(0.6)^k + \sum_{i=0}^{k-1}(0.6)^k(0.16)c\zeta\\ &\leq \zeta_0^+(0.6)^k + \sum_{i=0}^{\infty}(0.6)^k(0.16)c\zeta\\ &\leq 0.6^k + 0.4c\zeta\end{aligned}$$

2) To see that the denominator is positive, observe that the denominator is decreasing in all of its arguments: $\zeta_{j,*}^+, \zeta_{j,*}^+ f, c\zeta$, and b. Using the same upper bounds as before, we get that the denominator is greater than or equal to $0.78 > 0$. ■

C. Proof of Lemma 6.2

The proof of Lemma 6.2 follows from two lemmas. The first, Lemma 6.4, is the final conclusion for the projected CS step for $t \in \mathcal{I}_{j,k}$. Its proof follows using Lemmas 6.1, 3.7, 2.10, the CS error bound (Theorem 2.5) and some straightforward steps. The second, Lemma 6.5, is the final conclusion for one projection PCA step, i.e. for $t \in \mathcal{I}_{j,k}$. Its proof is much longer. It first uses a lemma based on the $\sin\theta$ and Weyl theorems (Theorems 2.7 and 2.8) to get a bound on ζ_k. This is Lemma 6.9. Next we bound $\kappa_s(D_{\text{new}})$ in Lemma 6.10. Finally in Lemma 6.11, we use the expression for e_t from Lemma 6.4, the matrix Hoeffding inequalities (Corollaries 2.14 and 2.15) and the bound from Lemma 6.10 to bound each of the terms in the bound on ζ_k to finally show that, conditioned on $\Gamma^e_{j,k-1}$, $\zeta_k \leq \zeta_k^+$ w.h.p.. We state the two lemmas first and then proceed to prove them in order.

Lemma 6.4 (Projected CS Lemma): Assume that all conditions of Theorem 4.2 hold.

1) *For all $t \in \mathcal{I}_{j,k}$, for any $k = 1, 2, \ldots K$, if $X_{j,k-1} \in \Gamma_{j,k-1}$,*
   a) *the projection noise β_t satisfies $\|\beta_t\|_2 \leq \zeta_{k-1}^+\sqrt{c}\gamma_{\text{new},k} + \zeta_*^+\sqrt{r}\gamma_* \leq \sqrt{c}0.72^{k-1}\gamma_{\text{new}} + 1.06\sqrt{\zeta} \leq \xi_0$.*
   b) *the CS error satisfies $\|\hat{S}_{t,cs} - S_t\|_2 \leq 7\xi_0$.*
   c) *$\hat{T}_t = T_t$*
   d) *e_t satisfies (10) and $\|e_t\|_2 \leq \phi^+[\kappa_s^+\zeta_{k-1}^+\sqrt{c}\gamma_{\text{new},k} + \zeta_*^+\sqrt{r}\gamma_*] \leq 0.18 \cdot 0.72^{k-1}\sqrt{c}\gamma_{\text{new}} + 1.17 \cdot 1.06\sqrt{\zeta}$. Recall that (10) is*

$$I_{T_t}(\Phi_{(t)})_{T_t}{}^{\dagger}\beta_t = I_{T_t}[(\Phi_{(t)})'_{T_t}(\Phi_{(t)})_{T_t}]^{-1}I_{T_t}{}'\Phi_{(t)}L_t$$

2) *For all* $k = 1, 2, \dots K$, $\mathbf{P}(\hat{T}_t = T_t$ *and* e_t *satisfies (10) for all* $t \in \mathcal{I}_{j,k} | X_{j,k-1}) = 1$ *for all* $X_{j,k-1} \in \Gamma_{j,k-1}$.

Lemma 6.5 (Projection PCA Lemma): Assume that all the conditions of Theorem 4.2 hold. Then, for all $k = 1, 2, \dots K$,

$$\mathbf{P}(\zeta_k \leq \zeta_k^+ | \Gamma_{j,k-1}^e) \geq p_k(\alpha, \zeta)$$

where ζ_k^+ *is defined in Definition 5.2 and* $p_k(\alpha,\zeta)$ *is defined in* (13).

Proof of Lemma 6.2: Observe that $\mathbf{P}(\Gamma_{j,k}|\Gamma_{j,k-1}) = \mathbf{P}(\check{\Gamma}_{j,k}|\Gamma_{j,k-1})$. The lemma then follows by combining Lemma 6.5 and item 2 of Lemma 6.4 and Lemma 2.11. ∎

D. Proof of Lemma 6.4

We begin by first bounding the RIC of the CS matrix Φ_k.

Lemma 6.6 (Bounding the RIC of Φ_k*): Recall that* $\zeta_* := \|(I - \hat{P}_* \hat{P}_*')P_*\|_2$. *The following hold.*

1) *Suppose that a basis matrix* P *can be split as* $P = [P_1, P_2]$ *where* P_1 *and* P_2 *are also basis matrices. Then* $\kappa_s^2(P) = \max_{T:|T|\leq s} \|I_T' P\|_2^2 \leq \kappa_s^2(P_1) + \kappa_s^2(P_2)$.
2) $\kappa_s^2(\hat{P}_*) \leq \kappa_{s,*}^2 + 2\zeta_*$
3) $\kappa_s(\hat{P}_{\text{new},k}) \leq \kappa_{s,\text{new}} + \tilde{\kappa}_{s,k}\zeta_k + \zeta_*$
4) $\delta_s(\Phi_0) = \kappa_s^2(\hat{P}_*) \leq \kappa_{s,*}^2 + 2\zeta_*$
5) $\delta_s(\Phi_k) = \kappa_s^2([\hat{P}_* \ \hat{P}_{\text{new},k}]) \leq \kappa_s^2(\hat{P}_*) + \kappa_s^2(\hat{P}_{\text{new},k}) \leq \kappa_{s,*}^2 + 2\zeta_* + (\kappa_{s,\text{new}} + \tilde{\kappa}_{s,k}\zeta_k + \zeta_*)^2$ for $k \geq 1$

Proof:

1) Since P is a basis matrix, $\kappa_s^2(P) = \max_{|T|\leq s} \|I_T{}'P\|_2^2$. Also, $\|I_T{}'P\|_2^2 = \|I_T{}'[P_1,P_2][P_1,P_2]'I_T\|_2 = \|I_T{}'(P_1P_1' + P_2P_2')I_T\|_2 \leq \|I_T{}'P_1P_1'I_T\|_2 + \|I_T{}'P_2P_2'I_T\|_2$. Thus, the inequality follows.
2) For any set T with $|T| \leq s$, $\|I_T{}'\hat{P}_*\|_2^2 = \|I_T{}'\hat{P}_*\hat{P}_*'I_T\|_2 = \|I_T{}'(\hat{P}_*\hat{P}_*' - P_*P_*{}' + P_*P_*{}')I_T\|_2 \leq \|I_T{}'(\hat{P}_*\hat{P}_*' - P_*P_*{}')I_T\|_2 + \|I_T{}'P_*P_*{}'I_T\|_2 \leq 2\zeta_* + \kappa_{s,*}^2$. The last inequality follows using Lemma 2.10 with $P = P_*$ and $\hat{P} = \hat{P}_*$.
3) By Lemma 2.10 with $P = P_*$, $\hat{P} = \hat{P}_*$ and $Q = P_{\text{new}}$, $\|P_{\text{new}}{}'\hat{P}_*\|_2 \leq \zeta_*$. By Lemma 2.10 with $P = P_{\text{new}}$ and $\hat{P} = \hat{P}_{\text{new},k}$, $\|(I - P_{\text{new}}P_{\text{new}}')\hat{P}_{\text{new},k}\|_2 = \|(I - \hat{P}_{\text{new},k}\hat{P}_{\text{new},k}')P_{\text{new}}\|_2$. For any set T with $|T| \leq s$, $\|I_T{}'\hat{P}_{\text{new},k}\|_2 \leq \|I_T{}'(I - P_{\text{new}}P_{\text{new}}')\hat{P}_{\text{new},k}\|_2 + \|I_T{}'P_{\text{new}}P_{\text{new}}'\hat{P}_{\text{new},k}\|_2 \leq \tilde{\kappa}_{s,k}\|(I - P_{\text{new}}P_{\text{new}}{}')\hat{P}_{\text{new},k}\|_2 + \|I_T{}'P_{\text{new}}\|_2 = \tilde{\kappa}_{s,k}\|(I - \hat{P}_{\text{new},k}\hat{P}_{\text{new},k}')P_{\text{new}}\|_2 + \|I_T{}'P_{\text{new}}\|_2 \leq \tilde{\kappa}_{s,k}\|D_{\text{new},k}\|_2 + \tilde{\kappa}_{s,k}\|\hat{P}_*\hat{P}_*'P_{\text{new}}\|_2 + \|I_T{}'P_{\text{new}}\|_2 \leq \tilde{\kappa}_{s,k}\zeta_k + \tilde{\kappa}_{s,k}\zeta_* + \kappa_{s,\text{new}} \leq \tilde{\kappa}_{s,k}\zeta_k + \zeta_* + \kappa_{s,\text{new}}$. Taking max over $|T| \leq s$ the claim follows.
4) This follows using Lemma 3.7 and the second claim of this lemma.
5) This follows using Lemma 3.7 and the first three claims of this lemma. ∎

Corollary 6.7: If the conditions of Theorem 4.2 are satisfied, and $X_{j,k-1} \in \Gamma_{j,k-1}$, *then*

1) $\delta_s(\Phi_0) \leq \delta_{2s}(\Phi_0) \leq {\kappa_{2s,*}^+}^2 + 2\zeta_*^+ < 0.1 < 0.1479$
2) $\delta_s(\Phi_{k-1}) \leq \delta_{2s}(\Phi_{k-1}) \leq {\kappa_{2s,*}^+}^2 + 2\zeta_*^+ + (\kappa_{2s,\text{new}}^+ + \tilde{\kappa}_{2s,k-1}^+\zeta_{k-1}^+ + \zeta_*^+)^2 < 0.1479$
3) $\phi_{k-1} \leq \frac{1}{1-\delta_s(\Phi_{k-1})} < \phi^+$

Proof: This follows using Lemma 6.6, the definition of $\Gamma_{j,k-1}$, and the bound on ζ_{k-1}^+ from Lemma 6.1. ∎

The following are straightforward bounds that will be useful for the proof of Lemma 6.4 and later.

Fact 6.8: Under the assumptions of Theorem 4.2:

1) $\zeta\gamma_* \leq \frac{\sqrt{\zeta}}{(r_0+(J-1)c)^{3/2}} \leq \sqrt{\zeta}$
2) $\zeta_{j,*}^+ \leq \frac{10^{-4}}{(r_0+(J-1)c)} \leq 10^{-4}$
3) $\zeta_{j,*}^+\gamma_*^2 \leq \frac{1}{(r_0+(J-1)c)^2} \leq 1$
4) $\zeta_{j,*}^+\gamma_* \leq \frac{\sqrt{\zeta}}{\sqrt{r_0+(J-1)c}} \leq \sqrt{\zeta}$
5) $\zeta_{j,*}^+ f \leq \frac{1.5\times10^{-4}}{r_0+(J-1)c} \leq 1.5\times10^{-4}$
6) $\zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$ (from Lemma 6.1)
7) $\zeta_{k-1}^+\gamma_{\text{new},k} \leq (0.6\cdot1.2)^{k-1}\gamma_{\text{new}} + 0.4c\zeta\gamma_* \leq 0.72^{k-1}\gamma_{\text{new}} + \frac{0.4\sqrt{\zeta}}{\sqrt{r_0+(J-1)c}} \leq 0.72^{k-1}\gamma_{\text{new}} + 0.4\sqrt{\zeta}$
8) $\zeta_{k-1}^+\gamma_{\text{new},k}^2 \leq (0.6\cdot1.2^2)^{k-1}\gamma_{\text{new}}^2 + 0.4c\zeta\gamma_*^2 \leq 0.864^{k-1}\gamma_{\text{new}}^2 + \frac{0.4}{(r_0+(J-1)c)^2} \leq 0.864^{k-1}\gamma_{\text{new}}^2 + 0.4$

Proof of Lemma 6.4: Recall that $X_{j,k-1} \in \Gamma_{j,k-1}$ *implies that* $\zeta_{j,*} \leq \zeta_{j,*}^+$ *and* $\zeta_{k-1} \leq \zeta_{k-1}^+$.

1)
   a) For $t \in \mathcal{I}_{j,k}$, $\beta_t := (I - \hat{P}_{(t-1)}\hat{P}_{(t-1)}')L_t = D_{*,k-1}a_{t,*} + D_{\text{new},k-1}a_{t,\text{new}}$. Thus, using Fact 6.8
   $$\begin{aligned}\|\beta_t\|_2 &\leq \zeta_{j,*}\sqrt{r}\gamma_* + \zeta_{k-1}\sqrt{c}\gamma_{\text{new},k}\\ &\leq \sqrt{\zeta}\sqrt{r} + (0.72^{k-1}\gamma_{\text{new}} + .4\sqrt{\zeta})\sqrt{c}\\ &= \sqrt{c}0.72^{k-1}\gamma_{\text{new}} + \sqrt{\zeta}(\sqrt{r} + 0.4\sqrt{c}) \leq \xi_0.\end{aligned}$$
   b) By Corollary 6.7, $\delta_{2s}(\Phi_{k-1}) < 0.15 < \sqrt{2}-1$. Given $|T_t| \leq s$, $\|\beta_t\|_2 \leq \xi_0 = \xi$, by Theorem 2.5, the CS error satisfies
   $$\|\hat{S}_{t,\text{cs}} - S_t\|_2 \leq \frac{4\sqrt{1+\delta_{2s}(\Phi_{k-1})}}{1-(\sqrt{2}+1)\delta_{2s}(\Phi_{k-1})}\xi_0 < 7\xi_0.$$
   c) Using the above, $\|\hat{S}_{t,\text{cs}} - S_t\|_\infty \leq 7\xi_0$. Since $\min_{i\in T_t}|(S_t)_i| \geq S_{\min}$ and $(S_t)_{T_t^c} = 0$, $\min_{i\in T_t}|(\hat{S}_{t,cs})_i| \geq S_{\min} - 7\xi_0$ and $\min_{i\in T_t^c}|(\hat{S}_{t,\text{cs}})_i| \leq 7\xi_0$. If $\omega < S_{\min} - 7\xi_0$, then $\hat{T}_t \supseteq T_t$. On the other hand, if $\omega > 7\xi_0$, then $\hat{T}_t \subseteq T_t$. Since $S_{\min} > 14\xi_0$ (condition 3 of the theorem) and ω satisfies $7\xi_0 \leq \omega \leq S_{\min} - 7\xi_0$ (condition 1 of the theorem), then the support of S_t is exactly recovered, i.e. $\hat{T}_t = T_t$.
   d) Given $\hat{T}_t = T_t$, the LS estimate of S_t satisfies $(\hat{S}_t)_{T_t} = [(\Phi_{k-1})_{T_t}]^\dagger y_t = [(\Phi_{k-1})_{T_t}]^\dagger(\Phi_{k-1}S_t + \Phi_{k-1}L_t)$ and $(\hat{S}_t)_{T_t^c} = 0$ for $t \in \mathcal{I}_{j,k}$. Also, $(\Phi_{k-1})_{T_t}{}'\Phi_{k-1} = I_{T_t}{}'\Phi_{k-1}$ (this follows since $(\Phi_{k-1})_{T_t} = \Phi_{k-1}I_{T_t}$ and $\Phi_{k-1}'\Phi_{k-1} = \Phi_{k-1}$). Using this, the LS error $e_t := \hat{S}_t - S_t$ satisfies (10). Thus, using Fact 6.8 and condition 2 of the theorem,
   $$\begin{aligned}\|e_t\|_2 &\leq \phi^+(\zeta_{j,*}^+\sqrt{r}\gamma_* + \kappa_{s,k-1}\zeta_{k-1}^+\sqrt{c}\gamma_{\text{new},k})\\ &\leq 1.2\Big(\sqrt{r}\sqrt{\zeta} + \sqrt{c}0.15(0.72)^{k-1}\\ &\qquad + \sqrt{c}0.06\sqrt{\zeta}\Big)\\ &= 0.18\sqrt{c}0.72^{k-1}\gamma_{\text{new}} +\\ &\qquad 1.2\sqrt{\zeta}(\sqrt{r} + 0.06\sqrt{c}).\end{aligned}$$

2) The second claim is just a restatement of the first. ∎

E. Proof of Lemma 6.5

The proof of Lemma 6.5 will use the next three lemmas

Lemma 6.9: If $\lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2 > 0$, then

$$\zeta_k \le \frac{\|\mathcal{R}_k\|_2}{\lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2} \le \frac{\|\mathcal{H}_k\|_2}{\lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2}$$

where $\mathcal{R}_k := \mathcal{H}_k E_{\text{new}}$ and A_k, $A_{k,\perp}$, $\mathcal{H}_k$ are defined in Definition 5.7.

Proof: Since $\lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2 > 0$, so $\lambda_{\min}(A_k) > \|A_{k,\perp}\|_2$. Since A_k is of size $c_{\text{new}} \times c_{\text{new}}$ and $\lambda_{\min}(A_k) > \|A_{k,\perp}\|_2$, $\lambda_{c_{\text{new}}+1}(\mathcal{A}_k) = \|A_{k,\perp}\|_2$. By definition of EVD, and since Λ_k is a $c_{\text{new}} \times c_{\text{new}}$ matrix, $\lambda_{\max}(\Lambda_{k,\perp}) = \lambda_{c_{\text{new}}+1}(\mathcal{A}_k + \mathcal{H}_k)$. By Weyl's theorem (Theorem 2.8), $\lambda_{c_{\text{new}}+1}(\mathcal{A}_k + \mathcal{H}_k) \le \lambda_{c_{\text{new}}+1}(\mathcal{A}_k) + \|\mathcal{H}_k\|_2 = \|A_{k,\perp}\|_2 + \|\mathcal{H}_k\|_2$. Therefore, $\lambda_{\max}(\Lambda_{k,\perp}) \le \|A_{k,\perp}\|_2 + \|\mathcal{H}_k\|_2$ and hence $\lambda_{\min}(A_k) - \lambda_{\max}(\Lambda_{k,\perp}) \ge \lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2 > 0$. Apply the $\sin\theta$ theorem (Theorem 2.7) with $\lambda_{\min}(A_k) - \lambda_{\max}(\Lambda_{k,\perp}) > 0$, we get

$$\|(I - \hat{P}_{\text{new},k}\hat{P}'_{\text{new},k})E_{\text{new}}\|_2 \le \frac{\|\mathcal{R}_k\|_2}{\lambda_{\min}(A_k) - \lambda_{\max}(\Lambda_{k,\perp})} \le \frac{\|\mathcal{H}_k\|_2}{\lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2}$$

Since $\zeta_k = \|(I - \hat{P}_{\text{new},k}\hat{P}'_{\text{new},k})D_{\text{new}}\|_2 = \|(I - \hat{P}_{\text{new},k}\hat{P}'_{\text{new},k})E_{\text{new}}R_{\text{new}}\|_2 \le \|(I - \hat{P}_{\text{new},k}\hat{P}'_{\text{new},k})E_{\text{new}}\|_2$, the result follows. The last inequality follows because $\|R_{\text{new}}\|_2 = \|E'_{\text{new}}D_{\text{new}}\|_2 \le 1$. ∎

Lemma 6.10: Assume that the assumptions of Theorem 4.2 hold. Conditioned on $\Gamma^e_{j,k-1}$,

$$\kappa_s(D_{\text{new}}) \le \frac{\kappa_s(P_{\text{new}}) + \zeta_*^+}{\sqrt{1-\zeta_*^+}} \le \frac{\kappa^+_{2s,\text{new}} + 0.0015}{\sqrt{1-0.0015}} \approx 0.1516 \le \kappa_s^+.$$

Proof: Recall that $D_{\text{new}} = D_{\text{new},0} = (I - \hat{P}_{j-1}\hat{P}'_{j-1})P_{\text{new}}$. Also $D_{\text{new}} \overset{\text{QR}}{=} E_{\text{new}}R_{\text{new}}$. By lemma 2.10 $\|R_{\text{new}}{}^{-1}\|_2 \le \frac{1}{\sqrt{1-\zeta_*^+}}$. $\kappa_s(D_{\text{new}}) = \kappa_s(E_{\text{new}}) = \max_{|T|\le s}\|I_T' D_{\text{new}} R_{\text{new}}{}^{-1}\|_2 \le \max_{|T|\le s}\|I_T' D_{\text{new}}\|_2 \|R_{\text{new}}{}^{-1}\|_2 \le \frac{\kappa_s(P_{\text{new}})+\zeta_*}{\sqrt{1-\zeta_*^+}}$. The event $\Gamma^e_{j,k-1}$ implies that $\zeta_* \le \zeta_*^+ \le 0.0015$. Thus, the lemma follows. ∎

Lemma 6.11 (High Probability Bounds for Each of the Terms in the ζ_k Bound (6.9)): Assume the conditions of Theorem 4.2 hold. Also assume that $\mathbf{P}(\Gamma^e_{j,k-1}) > 0$ for all $1 \le k \le K+1$. Then, for all $1 \le k \le K$

1) $\mathbf{P}\left(\lambda_{\min}(A_k) \ge \lambda^-_{\text{new},k}\left(1 - (\zeta^+_{j,*})^2 - \frac{c\zeta}{12}\right) \middle| \Gamma^e_{j,k-1}\right) > 1 - p_{a,k}(\alpha,\zeta)$ where

$$p_{a,k}(\alpha,\zeta) := c\exp\left(\frac{-\alpha\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot\min(1.2^{4k}\gamma_{\text{new}}^4, \gamma_*^4)}\right) + c\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot 4^2}\right)$$

2) $\mathbf{P}\left(\lambda_{\max}(A_{k,\perp}) \le \lambda^-_{\text{new},k}\left((\zeta^+_{j,*})^2 f + \frac{c\zeta}{24}\right) \middle| \Gamma^e_{j,k-1}\right) > 1 - p_b(\alpha,\zeta)$ where

$$p_b(\alpha,\zeta) := (n-c)\exp\left(\frac{-\alpha c^2\zeta(\lambda^-)^2}{8\cdot 24^2}\right)$$

3) $\mathbf{P}\left(\|\mathcal{H}_k\|_2 \le \lambda^-_{\text{new},k}(b + 0.125c\zeta) \middle| \Gamma^e_{j,k-1}\right) \ge 1 - p_c(\alpha,\zeta)$ where b is as defined in Definition 5.2 and

$$\begin{aligned} p_c(\alpha,\zeta) &:= n\exp\left(\frac{-\alpha\zeta^2(\lambda^-)^2}{8\cdot 24^2(.0324\gamma_{\text{new}}^2 + .0072\gamma_{\text{new}} + .0004)^2}\right) \\ &+ n\exp\left(\frac{-\alpha\zeta^2(\lambda^-)^2}{32\cdot 24^2(.06\gamma_{\text{new}}^2 + .0006\gamma_{\text{new}} + .4)^2}\right) \\ &+ n\exp\left(\frac{-\alpha\zeta^2(\lambda^-)^2\epsilon^2}{32\cdot 24^2(.186\gamma_{\text{new}}^2 + .00034\gamma_{\text{new}} + 2.3)^2}\right). \end{aligned}$$

Proof: The proof is quite long and hence is given in Appendix C. The first two claims are obtained by simplifying the terms and then appropriately applying the Hoeffding corollaries. The third claim first uses Lemma 6.4 to argue that conditioned on $X_{j,k-1} \in \Gamma_{j,k-1}$, e_t satisfies (10). It then simplifies the resulting expressions and eventually uses the Hoeffding corollaries. The simplification also uses the bound on $\kappa_s(D_{\text{new}})$ from Lemma 6.10. ∎

Proof of Lemma 6.5: Lemma 6.5 now follows by combining Lemmas 6.9 and 6.11 and defining

$$p_k(\alpha,\zeta) := 1 - p_{a,k}(\alpha,\zeta) - p_b(\alpha,\zeta) - p_c(\alpha,\zeta). \quad (13)$$

∎

VII. ReProCS With Cluster PCA

The ReProCS approach studied so far is designed under the assumption that the subspace in which L_t lies can only grow over time. In practice, usually, the dimension of this subspace typically remains roughly constant. A simple way to model this is to assume that at every change time, t_j, some new directions can get added and some directions from the existing subspace can get deleted and to assume an upper bound on the difference between the total number of added and deleted directions. We specify this model next.

Signal Model 7.1: Assume that $L_t = P_{(t)}a_t$ where $P_{(t)} = P_j$ for all $t_j \le t < t_{j+1}$, $j = 0, 1, 2\cdots J$, P_j is an $n \times r_j$ basis matrix with $r_j \ll \min(n, (t_{j+1} - t_j))$. We let $t_0 = 0$ and t_{J+1} equal the sequence length. This can be infinity also.

1) *At the change times, t_j, P_j changes as*

$$P_j = [(P_{j-1}R_j \setminus P_{j,\text{old}}) \quad P_{j,\text{new}}]$$

Here, R_j is a rotation matrix, $P_{j,\text{new}}$ is an $n \times c_{j,\text{new}}$ basis matrix with $P'_{j,\text{new}}P_{j-1} = 0$ and $P_{j,\text{old}}$ contains $c_{j,\text{old}}$ columns of $P_{j-1}R_j$. Thus $r_j = r_{j-1} + c_{j,\text{new}} - c_{j,\text{old}}$. Also, $0 < t_{train} \le t_1$. This model is illustrated in Figure 3.

2) *There exist constants $c_{\max}$ and c_{dif} such that $0 \le c_{j,\text{new}} \le c_{\max}$ and $\sum_{i=1}^{j}(c_{i,\text{new}} - c_{i,\text{old}}) \le c_{dif}$ for all j. Thus, $r_j = r_0 + \sum_{i=1}^{j}(c_{i,\text{new}} - c_{i,\text{old}}) \le r_{\max} := r_0 + c_{dif}$, i.e., the rank of P_j is upper bounded by $r_{\max}$.*

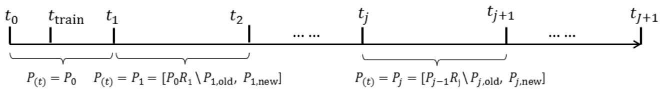

Fig. 3. The subspace change model given in Signal Model 7.1. Here $t_0 = 0$.

The ReProCS algorithm (Algorithm 2) still applies for the above more general model. We can conclude the following for it.

Corollary 7.2: Consider Algorithm 2 for the model given above. The result of Theorem 4.2 applies with the following change: we also need $\kappa_{2s}([P_0, P_{1,\text{new}}, \ldots, P_{J-1,\text{new}}]) \le 0.3$.

Because Algorithm 2 never deletes directions, the rank of $\hat{P}_{(t)}$ keeps increasing with every subspace change time (even though the rank of $P_{(t)}$ is now bounded by $r_0 + c_{\text{dif}}$). As a result, the performance guarantee above still requires a bound on J that is imposed by the denseness assumption. In this section, we address this limitation by re-estimating the current subspace after the newly added directions have been accurately estimated. This helps to "delete" span(P_{old}) from the subspaces estimate. For the resulting algorithm, as we will see, we do not need a bound on the number of changes, J, as long as the separation between the subspace change times is allowed to grow logarithmically with J.

One simple way to re-estimate the current subspace would be by standard PCA: at $t = \tilde{t}_j + \tilde{\alpha} - 1$, compute $\hat{P}_j \leftarrow$ proj-PCA$([\hat{L}_t; \tilde{\mathcal{I}}_{j,1}], [.], r_j)$ and let $\hat{P}_{(t)} \leftarrow \hat{P}_j$. Using the $\sin\theta$ theorem [31] and the matrix Hoeffding inequality [32], and using the procedure used earlier to analyze projection PCA, it can be shown that, as long as f, a bound on the maximum condition number of Cov$[L_t]$, is small enough, doing this is guaranteed to give an accurate estimate of span(P_j). However as explained in Remark 3.4, f cannot be small because our problem definition allows large noise, L_t, but assumes slow subspace change. In other works that analyze standard PCA, see [33] and references therein, the large condition number does not cause a problem because they assume that the error (e_t in our case) in the observed data vector ($\hat{L}_t$) is uncorrelated with the true data vector (L_t). Under this assumption, one only needs to increase the PCA data length α to deal with larger condition numbers. However, in our case, because e_t is correlated with L_t, this strategy does not work. This issue is explained in detail in Appendix B.

In this section, we introduce a generalization of the above strategy called cluster-PCA that removes the requirement that f be small, but instead only requires that the eigenvalues of Cov(L_t) be clustered for the times when the changed subspace has stabilized. Under this assumption, cluster-PCA recovers one cluster of entries of P_j at a time by using an approach that generalizes the projection PCA step developed earlier. We first explain the clustering assumption in Sec VII-A below and then give the cluster-PCA algorithm.

A. Clustering Assumption

For positive integers K and α, let $\tilde{t}_j := t_j + K\alpha$. We set their values in, Theorem 7.7. Recall from the model on L_t and the slow subspace change assumption that new directions, $P_{j,\text{new}}$, get added at $t = t_j$ and initially, for the first α frames, the projection of L_t along these directions is small (and thus their variances are small), but can increase gradually. It is fair to assume that within $K\alpha$ frames, i.e. by $t = \tilde{t}_j$, the variances along these new directions have stabilized and do not change much for $t \in [\tilde{t}_j, t_{j+1} - 1]$. It is also fair to assume that the same is true for the variances along the existing directions, P_{j-1}. In other words, we assume that the matrix Λ_t is either constant or does not change much during this period. Under this assumption, we assume that we can cluster its eigenvalues (diagonal entries) into a few clusters such that the distance between consecutive clusters is large and the distance between the smallest and largest element of each cluster is small. We make this precise below.

Assumption 7.3: Assume the following.

1) *Either* $\Lambda_t = \Lambda_{\tilde{t}_j}$ *for all* $t \in [\tilde{t}_j, t_{j+1} - 1]$ *or* Λ_t *changes very little during this period so that for each* $i = 1, 2, \cdots, r_j$, $\min_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_i(\Lambda_t) \ge \max_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_{i+1}(\Lambda_t)$.
2) *Let* $\mathcal{G}_{j,(1)}, \mathcal{G}_{j,(2)}, \cdots, \mathcal{G}_{j,(\vartheta_j)}$ *be a partition of the index set* $\{1, 2, \ldots r_j\}$ *so that* $\min_{i \in \mathcal{G}_{j,(k)}} \min_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_i(\Lambda_t) > \max_{i \in \mathcal{G}_{j,(k+1)}} \max_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_i(\Lambda_t)$, *i.e. the first group/cluster contains the largest set of eigenvalues, the second one the next smallest set and so on (see Figure 4). Let*
   a) $G_{j,k} := (P_j)_{\mathcal{G}_{j,(k)}}$ *be the corresponding cluster of eigenvectors, then* span$(P_j) =$ span$([G_{j,1}, G_{j,2}, \cdots, G_{j,\vartheta_j}])$;
   b) $\tilde{c}_{j,k} := |\mathcal{G}_{j,(k)}|$ *be the number of elements in* $\mathcal{G}_{j,(k)}$, *then* $\sum_{k=1}^{\vartheta_j} \tilde{c}_{j,k} = r_j$;
   $\tilde{c}_{\min} := \min_j \min_{k=1,2,\cdots,\vartheta_j} \tilde{c}_{j,k}$
   c) $\lambda_{j,k}{}^- := \min_{i \in \mathcal{G}_{j,(k)}} \min_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_i(\Lambda_t)$, $\lambda_{j,k}{}^+ := \max_{i \in \mathcal{G}_{j,(k)}} \max_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_i(\Lambda_t)$ *and* $\lambda_{j,\vartheta_j+1}{}^+ := 0$;
   d) $\tilde{g}_{j,k} := \lambda_{j,k}{}^+ / \lambda_{j,k}{}^-$ *(notice that* $\tilde{g}_{j,k} \ge 1$*)*;
   e) $\tilde{h}_{j,k} := \lambda_{j,k+1}{}^+ / \lambda_{j,k}{}^-$ *(notice that* $\tilde{h}_{j,k} < 1$*)*;
   f) $\tilde{g}_{\max} := \max_j \max_{k=1,2,\cdots,\vartheta_j} \tilde{g}_{j,k}$,
   $\tilde{h}_{\max} := \max_j \max_{k=1,2,\cdots,\vartheta_j} \tilde{h}_{j,k}$,
   g) $\vartheta_{\max} := \max_j \vartheta_j$

   We assume that $\tilde{g}_{\max}$ *is small enough (the distance between the smallest and largest eigenvalues of a cluster is small) and* $\tilde{h}_{\max}$ *is small enough (distance between consecutive clusters is large). We quantify this in Theorem 7.7.*

Remark 7.4: In order to address a reviewer's concern, we should clarify the following point. The above assumption still allows the newly added eigenvalues to become large and hence still allows the subspace of L_t *to change significantly over time. The above requires the covariance matrix of* L_t *to be constant or nearly constant only for the time between* $\tilde{t}_j := t_j + K\alpha$ *and the next change time,* t_{j+1} *and not for the first* $K\alpha$ *frames. Slow subspace change assumes that the projection of* L_t *along the new directions is initially small for the first* α *frames but then can increase gradually over the next* $K-1$ *intervals of duration* α. *The variance along the new directions can increase by as much as* 1.2^{2K} *times the initial variance. Thus by* $t = \tilde{t}_j = t_j + K\alpha$, *the variances along the new directions can have already increased to large enough values.*

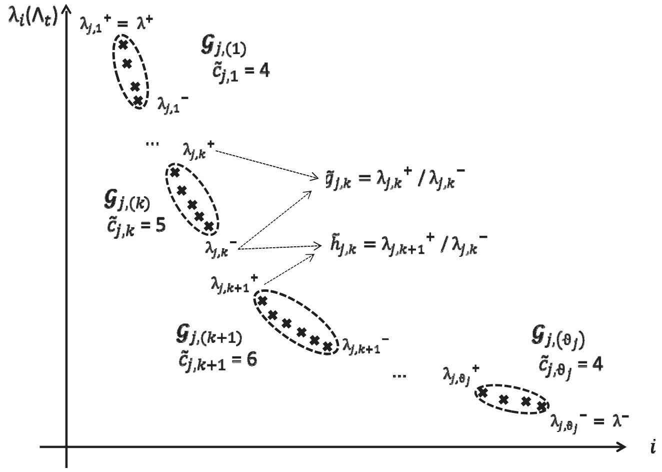

Fig. 4. We illustrate the clustering assumption. Assume $\Lambda_t = \Lambda_{\tilde{t}_j}$.

We can allow the variances to increase for even longer with the following simple change: re-define $\tilde{t}_j$ as $\tilde{t}_j := t_{j+1} - \vartheta_j \tilde{\alpha}$ in both the clustering assumption and the algorithm. With this redefinition, we will be doing cluster-PCA at the very end of the current subspace interval.

Lastly, note that the projection along the new directions can further increase in the later subspace change periods also.

B. The ReProCS With Cluster PCA Algorithm

ReProCS-cPCA is summarized in Algorithm 3. It uses the following definition.

Definition 7.5: Let $\tilde{t}_j := t_j + K\alpha$. Define the following time intervals

1) $\mathcal{I}_{j,k} := [t_j + (k-1)\alpha, t_j + k\alpha - 1]$ for $k = 1, 2, \cdots, K$.
2) $\tilde{\mathcal{I}}_{j,k} := [\tilde{t}_j + (k-1)\tilde{\alpha}, \tilde{t}_j + k\tilde{\alpha} - 1]$ for $k = 1, 2, \cdots, \vartheta_j$.
3) $\tilde{\mathcal{I}}_{j,\vartheta_j+1} := [\tilde{t}_j + \vartheta_j \tilde{\alpha}, t_{j+1} - 1]$.

Steps 1, 2, 3a and 3b of ReProCS-cPCA are the same as Algorithm 2. As shown earlier, within K proj-PCA updates (K chosen as given in Theorem 7.7) $\|e_t\|_2$ and the subspace error, $\text{SE}_{(t)}$, drop down to a constant times ζ. In particular, if at $t = t_j - 1$, $\text{SE}_{(t)} \leq r\zeta$, then at $t = \tilde{t}_j := t_j + K\alpha$, we can show that $\text{SE}_{(t)} \leq (r + c_{\max})\zeta$. Here $r := r_{\max} = r_0 + c_{\text{dif}}$. To bring $\text{SE}_{(t)}$ down to $r\zeta$ before t_{j+1}, we proceed as follows. The main idea is to recover one cluster of entries of P_j at a time. For each batch we use a new set of $\tilde{\alpha}$ frames. The entire procedure is done at $t = \tilde{t}_j + \vartheta_j \tilde{\alpha} - 1$ (since we cannot update $\hat{P}_{(t)}$ until all clusters are recovered). We proceed as follows. In the first iteration, we use standard PCA to estimate the first cluster, $\text{span}(G_{j,1})$. In the k^{th} iteration, we apply proj-PCA on $[\hat{L}_{\tilde{t}_j+(k-1)\tilde{\alpha}}, \ldots, \hat{L}_{\tilde{t}_j+k\tilde{\alpha}-1}]$ with $P \leftarrow [\hat{G}_{j,1}, \hat{G}_{j,2}, \ldots \hat{G}_{j,k-1}]$ to estimate $\text{span}(G_{j,k})$. By modifying the approach used to prove Theorem 4.2, we can show that since $\tilde{g}_{j,k}$ and $\tilde{h}_{j,k}$ are small enough, $\text{span}(G_{j,k})$ will be accurately recovered, i.e. $\|(I - \sum_{i=1}^{k} \hat{G}_{j,i}\hat{G}_{j,i}')G_{j,k}\|_2 \leq \tilde{c}_{j,k}\zeta$. We do this ϑ_j times and finally we set $\hat{P}_j \leftarrow [\hat{G}_{j,1}, \hat{G}_{j,2} \ldots \hat{G}_{j,\vartheta_j}]$ and $\hat{P}_{(t)} \leftarrow \hat{P}_j$. Thus, at $t = \tilde{t}_j + \vartheta_j\tilde{\alpha} - 1$, $\text{SE}_{(t)} \leq \sum_{k=1}^{\vartheta_j} \|(I - \sum_{i=1}^{k} \hat{G}_{j,i}\hat{G}_{j,i}')G_{j,k}\|_2 \leq \sum_{k=1}^{\vartheta_j} \tilde{c}_{j,k}\zeta = r_j\zeta \leq r\zeta$. Under the assumption that $t_{j+1} - t_j \geq K\alpha + \vartheta_{\max}\tilde{\alpha}$, this means that before the next subspace change time, t_{j+1}, $\text{SE}_{(t)}$ is below $r\zeta$.

We illustrate the ideas of subspace estimation by addition proj-PCA and cluster-PCA in Fig. 5. The connection between proj-PCA done in the addition step and for the cluster-PCA (in deletion) step is given in Table I.

C. Performance Guarantees

Definition 7.6: We need the following definitions for stating the main result.

1) *We define $\alpha_{del}(\zeta)$ as*

$$\alpha_{\text{del}}(\zeta) := \left\lceil (\log 6\vartheta_{\max} J + 11 \log n) \cdot \frac{8 \cdot 10^2}{(\zeta\lambda^-)^2} \max(4.2^2, 4b_7^2) \right\rceil$$

where $b_7 := (\sqrt{r}\gamma_* + \phi^+\sqrt{\zeta})^2$ and $\phi^+ = 1.1732$. We choose α_{del} so that if , $\tilde{\alpha} \geq \alpha_{del}$, then the conclusions of the theorem will hold wth probability at least $(1 - 2n^{-10})$.

2) *Define*

$$\begin{aligned} f_{inc}(\tilde{g}, \tilde{h}, \kappa_{s,e}^+, \kappa_{s,D}^+) &:= (r + c) \\ &\times \zeta \Bigg[\max(3\kappa_{s,e}^+\kappa_{s,D}^+\phi^+\tilde{g}, \kappa_{s,e}^+\phi^+\tilde{h}) \\ &+ \Big[\kappa_{s,e}^+\phi^+ + \kappa_{s,e}^+(1 + 2\phi^+)\frac{r^2\zeta^2}{\sqrt{1 - r^2\zeta^2}}\Big]\tilde{h} \\ &+ \Big[\frac{r^2}{r + c}\zeta + 4r\zeta\kappa_{s,e}^+\phi^+ + 2(r + c)\zeta(1 + {\kappa_{s,e}^+}^2){\phi^+}^2\Big] f \\ &+ 0.2\frac{1}{r + c}\Bigg], \\ f_{dec}(\tilde{g}, \tilde{h}, \kappa_{s,e}^+, \kappa_{s,D}^+) &:= 1 - \tilde{h} - 0.2\zeta - r^2\zeta^2 f - r^2\zeta^2 \\ &- f_{inc}(\tilde{g}, \tilde{h}, \kappa_{s,e}^+, \kappa_{s,D}^+) \end{aligned}$$

Notice that $f_{inc}(.)$ is an increasing function of $\tilde{g}, \tilde{h}$ and $f_{dec}(.)$ is a decreasing function of $\tilde{g}, \tilde{h}$.

Theorem 7.7: Consider Algorithm 3. Let $c := c_{\max}$ and $r := r_{\max} = r_0 + c_{dif}$. Pick a ζ that satisfies

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

Assume that the initial subspace estimate is accurate enough, i.e. $\|(I - \hat{P}_0\hat{P}_0')P_0\| \leq r_0\zeta$. If the following conditions hold:

1) *All of the conditions of Theorem 4.2 hold with L_t satisfying Signal model 7.1,*
2) $\tilde{\alpha} \geq \alpha_{\text{del}}(\zeta)$,
3) $\min_j(t_{j+1} - t_j) > K\alpha + \vartheta_{\max}\tilde{\alpha}$
4) *algorithm estimates $\hat{P}_{j-1}$ and $\hat{P}_{j,\text{new},K}$ satisfy*

$$\max_j \kappa_s((I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,\text{new},K}\hat{P}_{j,\text{new},K}')P_j) \leq \kappa_{s,e}^+$$

5) *(clustered eigenvalues) Assumption 7.3 holds with $\tilde{g}_{\max}, \tilde{h}_{\max}, \tilde{c}_{\min}$ satisfying $f_{dec}(\tilde{g}_{\max}, \tilde{h}_{\max}, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta) - \frac{f_{inc}(\tilde{g}_{\max}, \tilde{h}_{\max}, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta)}{\tilde{c}_{\min}\zeta} > 0.$*

then, with probability at least $1 - 2n^{-10}$, at all times, t,

Algorithm 3 Recursive Projected CS With Cluster-PCA (ReProCS-cPCA)

Parameters: algorithm parameters: ξ, ω, α, $\tilde{\alpha}$, K, model parameters: t_j, $c_{j,\text{new}}$, ϑ_j and $\tilde{c}_{j,i}$

Input: $n \times 1$ vector, M_t, and $n \times r_0$ basis matrix $\hat{P}_0$. **Output:** $n \times 1$ vectors $\hat{S}_t$ and $\hat{L}_t$, and $n \times r_{(t)}$ basis matrix $\hat{P}_{(t)}$.

Initialization: Let $\hat{P}_{(t_{\text{train}})} \leftarrow \hat{P}_0$. 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 $\Phi_{(t)} \leftarrow I - \hat{P}_{(t-1)}\hat{P}_{(t-1)}'$, $y_t \leftarrow \Phi_{(t)} M_t$
   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** $\hat{P}_{(t)}$:
   a) If $t \neq t_j + q\alpha - 1$ for any $q = 1, 2, \ldots K$ and $t \neq t_j + K\alpha + \vartheta_j \tilde{\alpha} - 1$,
      i) set $\hat{P}_{(t)} \leftarrow \hat{P}_{(t-1)}$
   b) **Addition: Estimate** $\text{span}(P_{j,\text{new}})$ **iteratively using proj-PCA:** If $t = t_j + k\alpha - 1$
      i) $\hat{P}_{j,\text{new},k} \leftarrow$ proj-PCA$([\hat{L}_{t_j+(k-1)\alpha}, \ldots, \hat{L}_{t_j+k\alpha-1}], \hat{P}_{j-1}, c_{j,\text{new}})$
      ii) set $\hat{P}_{(t)} \leftarrow [\hat{P}_{j-1} \ \hat{P}_{j,\text{new},k}]$.
      iii) If $k = K$, reset $k \leftarrow 1$; else increment $k \leftarrow k + 1$.
   c) **Deletion: Estimate** $\text{span}(P_j)$ **by cluster-PCA:** If $t = t_j + K\alpha + \vartheta_j \tilde{\alpha} - 1$,
      i) set $\hat{G}_{j,0} \leftarrow [.]$
      ii) For $i = 1, 2, \cdots, \vartheta_j$,
         - $\hat{G}_{j,i} \leftarrow$ proj-PCA$([\hat{L}_{\tilde{t}_j+(i-1)\tilde{\alpha}}, \ldots, \hat{L}_{\tilde{t}_j+i\tilde{\alpha}-1}], [\hat{G}_{j,1}, \hat{G}_{j,2}, \ldots \hat{G}_{j,i-1}], \tilde{c}_{j,i})$

         End for
      iii) set $\hat{P}_j \leftarrow [\hat{G}_{j,1}, \cdots, \hat{G}_{j,\vartheta_j}]$ and set $\hat{P}_{(t)} \leftarrow \hat{P}_j$.
      iv) increment $j \leftarrow j + 1$.

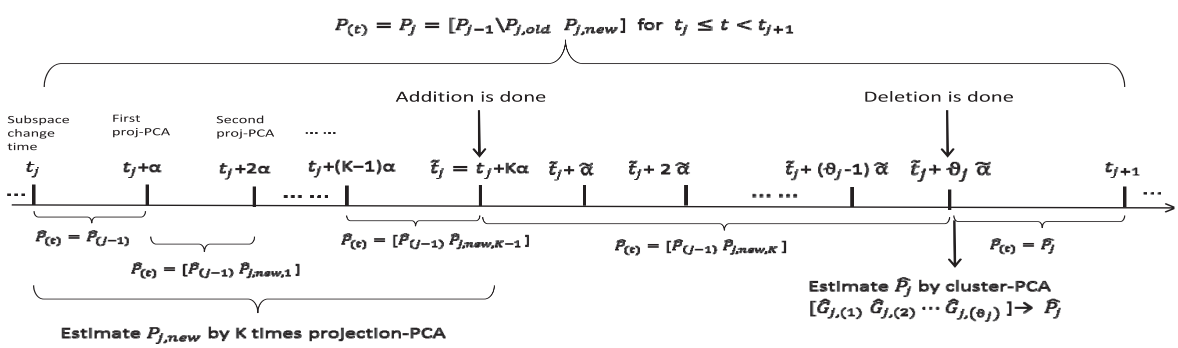

Fig. 5. A diagram illustrating subspace estimation by ReProCS-cPCA.

1) $\hat{T}_t = T_t$ and $\|e_t\|_2 = \|L_t - \hat{L}_t\|_2 = \|\hat{S}_t - S_t\|_2 \leq 0.18\sqrt{c}\gamma_{\text{new}} + 1.24\sqrt{\zeta}$.
2) *the subspace error, $\text{SE}_{(t)}$ satisfies*

$$\text{SE}_{(t)} \leq \begin{cases} 0.6^{k-1} + r\zeta + 0.4c\zeta & \text{if } t \in \mathcal{I}_{j,k}, \ k = 1, 2, \cdots, K \\ (r+c)\zeta & \text{if } t \in \tilde{\mathcal{I}}_{j,k}, \ k = 1, 2, \cdots, \vartheta_j \\ r\zeta & \text{if } t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases}$$

3) *the error $e_t = \hat{S}_t - S_t = L_t - \hat{L}_t$ satisfies the following at various times*

$$\|e_t\|_2 \leq \begin{cases} 1.17[0.15 \cdot 0.72^{k-1}\sqrt{c}\gamma_{\text{new}} + \\ \qquad 0.15 \cdot 0.4c\zeta\sqrt{c}\gamma_* + r\zeta\sqrt{r}\gamma_*] \\ \qquad \text{if } t \in \mathcal{I}_{j,k}, \ k = 1, 2, \cdots, K \\ 1.17(r+c)\zeta\sqrt{r}\gamma_* \\ \qquad \text{if } \ t \in \tilde{\mathcal{I}}_{j,k}, \ k = 1, 2, \cdots, \vartheta_j \\ 1.17r\zeta\sqrt{r}\gamma_* \ \text{ if } \ t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases}$$

D. Special Case When f is Small

If in a problem, L_t has small magnitude for all times t or if its subspace does not change, then f can be small. In this case, the clustering assumption is not needed, or in fact it trivially holds with $\vartheta_j = 1$, $\tilde{c}_{j,1} = r_j$, $\tilde{g}_{\max} = \tilde{g}_{j,1} = f$ and $\tilde{h}_{\max} = h_{j,1} = 0$. Thus, $\vartheta_{\max} = 1$. With this, the following corollary holds.

Corollary 7.8: Assume that all conditions of Theorem 7.7 hold except the last one (clustering assumption). If f is small enough so that $f_{inc}(f, 0, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta) \leq f_{dec}(f, 0, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta) r_j \zeta$, then, all conclusions of Theorem 7.7 hold.

E. Discussion

Notice from Definition 4.1 that $K = K(\zeta)$ is larger if ζ is smaller. Also, both $\alpha_{\text{add}}(\zeta)$ and $\alpha_{\text{del}}(\zeta)$ are inversely

TABLE I

COMPARING AND CONTRASTING THE ADDITION PROJ-PCA STEP AND PROJ-PCA USED IN THE DELETION STEP (CLUSTER-PCA)

k^{th} **iteration of addition proj-PCA**	k^{th} **iteration of cluster-PCA in the deletion step**
done at $t = t_j + k\alpha - 1$	done at $t = t_j + K\alpha + \vartheta_j\tilde{\alpha} - 1$
goal: keep improving estimates of $\text{span}(P_{j,\text{new}})$	goal: re-estimate $\text{span}(P_j)$ and thus "delete" $\text{span}(P_{j,\text{old}})$
compute $\hat{P}_{j,\text{new},k}$ by proj-PCA on $[\hat{L}_t : t \in \mathcal{I}_{j,k}]$ with $P = \hat{P}_{j-1}$	compute $\hat{G}_{j,k}$ by proj-PCA on $[\hat{L}_t : t \in \tilde{\mathcal{I}}_{j,k}]$ with $P = \hat{G}_{j,\text{det},k} = [\hat{G}_{j,1}, \cdots, \hat{G}_{j,k-1}]$
start with $\|(I - \hat{P}_{j-1}\hat{P}_{j-1}')P_{j-1}\|_2 \leq r\zeta$ and $\zeta_{j,k-1} \leq \zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$	start with $\|(I - \hat{G}_{j,\text{det},k}\hat{G}_{j,\text{det},k}')G_{j,\text{det},k}\|_2 \leq r\zeta$ and $\zeta_{j,K} \leq c\zeta$
need small g which is the maximum condition number of $\text{Cov}(P_{j,\text{new}}'L_t)$	need small $\tilde{g}_{\max}$ which is the maximum of the maximum condition number of $\text{Cov}(G_{j,k}'L_t)$
no undetected subspace	extra issue: ensure perturbation due to $\text{span}(G_{j,\text{undet},k})$ is small; need small $\tilde{h}_{j,k}$ to ensure the above
$\zeta_{j,k}$ is the subspace error in estimating $\text{span}(P_{j,\text{new}})$ after the k^{th} step	$\tilde{\zeta}_{j,k}$ is the subspace error in estimating $\text{span}(G_{j,k})$ after the k^{th} step
end with $\zeta_{j,k} \leq \zeta_k^+ \leq 0.6^k + 0.4c\zeta$ w.h.p.	end with $\tilde{\zeta}_{j,k} \leq \tilde{c}_{j,k}\zeta$ w.h.p.
stop when $k = K$ with K chosen so that $\zeta_{j,K} \leq c\zeta$	stop when $k = \vartheta_j$ and $\tilde{\zeta}_{j,k} \leq \tilde{c}_{j,k}\zeta$ for all $k = 1, 2, \cdots, \vartheta_j$
after K^{th} iteration: $\hat{P}_{(t)} \leftarrow [\hat{P}_{j-1}\ \hat{P}_{j,\text{new},K}]$ and $SE_{(t)} \leq (r+c)\zeta$	after ϑ_j^{th} iteration: $\hat{P}_{(t)} \leftarrow [\hat{G}_{j,1}, \cdots, \hat{G}_{j,\vartheta_j}]$ and $SE_{(t)} \leq r\zeta$

proportional to ζ. Thus, if we want to achieve a smaller lowest error level, ζ, we need to compute both addition proj-PCA and cluster-PCA's over larger durations, α and $\tilde{\alpha}$ respectively, and we will need more number of addition proj-PCA steps K. This means that we also require a larger delay between subspace change times, i.e. larger $t_{j+1} - t_j$.

Let us first compare the above result with that for ReProCS for the same subspace change model, i.e. the result from Corollary 7.2. The most important difference is that ReProCS requires $\kappa_{2s}([P_0, P_{1,\text{new}}, \ldots P_{J,\text{new}}]) \leq 0.3$ whereas ReProCS-cPCA only requires $\max_j \kappa_{2s}(P_j) \leq 0.3$. Moreover in case of ReProCS, the denominator in the bound on ζ also depends on J whereas in case of ReProCS-cPCA, it only depends on $r_{max} + c_{\max}$. Because of this, in Theorem 7.7 for ReProCS-cPCA, the only place where J appears is in the definitions of α_{add} and α_{del}. These govern the delay between subspace change times, $t_{j+1} - t_j$. Thus, with ReProCS-cPCA, J can keep increasing, as long as $\min_j(t_{j+1} - t_j)$ also increases accordingly. Moreover, notice that the dependence of α_{add} and α_{del} on J is only logarithmic and thus $\min_j(t_{j+1} - t_j)$ needs to only increase in proportion to $\log J$. The main extra assumptions that ReProCS-cPCA needs are the clustering assumption; a longer delay between subspace change times; and a denseness assumption similar to that on $D_{j,\text{new},k}$. We verify the clustering assumption in Sec IX-A. The ReProCS-cPCA algorithm also needs to know the cluster sizes of the eigenvalues. These can, however, be estimated by computing the eigenvalues of the estimated covariance matrix at $t = \tilde{t}_j + \tilde{\alpha}$ and clustering them.

Comparison With the PCP Result From [6]: Our results need many more assumptions compared with the PCP result [6] which only assumes independent support change of the sparse part and a denseness assumption on the low-rank part. The most important limitation of our work is that both our results need an assumption on the algorithm estimates, thus neither can be called a correctness result. Moreover, both the results assume that the algorithms know the model parameters while the result for PCP does not. The key limiting aspect here is the knowledge of the subspace change times. The advantages of our results w.r.t. that for PCP are as follows. (a) Both results are for online algorithms; and (b) both need weaker denseness assumptions on the singular vectors of $\mathcal{L}_t$ as compared to PCP. PCP [6] requires denseness of both the left and right singular vectors of $\mathcal{L}_t$ and it requires a bound on $\|UV'\|_\infty$ where U and V denote the left and right singular vectors. Denseness of only the left singular vectors is needed in our case (notice that $U = [P_{j-1}, P_{j,\text{new}}]$). (c) Finally, the most important advantage of the ReProCS-cPCA result is that it does not need a bound on J (number of subspace change times) as long as $\min_j(t_{j+1} - t_j)$ increases in proportion to $\log J$, and equivalently, does not need a bound on the rank of $\mathcal{L}_t$. However PCP needs a tight bound on the rank of $\mathcal{L}_t$.

VIII. PROOF OF THEOREM 7.7

We first give some new definitions next. We then give the key lemmas leading to the proof of the theorem and the proof itself. Finally we prove these lemmas.

A. Some New Definitions

Unless redefined here, all previous definitions still apply.

Definition 8.1: Define the following:

1) $r = r_{\max} = r_0 + c_{\text{dif}}$ (*Note that this is a redefinition from Definition 5.1*)
2) $\zeta_{j,*}^+ := r\zeta$ (*Note that this is a redefinition from Definition 5.2*)
3) *define the sequence* $\{\tilde{\zeta}_k^+\}_{k=1,2,\cdots,\vartheta_j}$ *as follows*

$$\tilde{\zeta}_k^+ := \frac{f_{inc}(\tilde{g}_k, \tilde{h}_k, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta)}{f_{dec}(\tilde{g}_k, \tilde{h}_k, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta)}$$

where $f_{inc}(.)$ *and* $f_{dec}(.)$ *are defined in Definition 7.6.*

Definition 8.2: Define

1) $\Psi_{j,k} := I - \sum_{i=0}^{k} \hat{G}_{j,i}\hat{G}_{j,i}'$.
2) $G_{j,det,k} := [G_{j,1} \cdots, G_{j,k-1}]$ *and* $\hat{G}_{j,det,k} := [\hat{G}_{j,1} \cdots, \hat{G}_{j,k-1}]$. *Notice that* $\Psi_{j,k} = I - \hat{G}_{j,det,k+1}\hat{G}_{j,det,k+1}'$.
3) $G_{j,undet,k} := [G_{j,k+1} \cdots, G_{j,\vartheta_j}]$.
4) $D_{j,k} := \Psi_{j,k-1}G_{j,k}$, $D_{j,det,k} := \Psi_{j,k-1}G_{j,det,k}$ *and* $D_{j,undet,k} := \Psi_{j,k-1}G_{j,undet,k}$.

Definition 8.3:

1) *Let* $D_{j,k} \overset{QR}{=} E_{j,k}R_{j,k}$ *denote its reduced QR decomposition, i.e. let* $E_{j,k}$ *be a basis matrix for* $\text{span}(D_{j,k})$ *and let* $R_{j,k} := E_{j,k}'D_{j,k}$.

2) *Let $E_{j,k,\perp}$ be a basis matrix for the orthogonal complement of* $\operatorname{span}(E_{j,k}) = \operatorname{span}(D_{j,k})$. *To be precise, $E_{j,k,\perp}$ is a $n \times (n - \tilde{c}_{j,k})$ basis matrix that satisfies $E_{j,k,\perp}{}' E_{j,k} = 0$.*

3) *Using $E_{j,k}$ and $E_{j,k,\perp}$, define $\tilde{A}_{j,k}$, $\tilde{A}_{j,k,\perp}$, $\tilde{H}_{j,k}$, $\tilde{H}_{j,k,\perp}$ and $\tilde{B}_{j,k}$ as*

$$\tilde{A}_{j,k} := \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k}{}' \Psi_{j,k-1} L_t L_t{}' \Psi_{j,k-1} E_{j,k}$$

$$\tilde{A}_{j,k,\perp} := \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k,\perp}{}' \Psi_{j,k-1} L_t L_t{}' \Psi_{j,k-1} E_{j,k,\perp}$$

$$\tilde{H}_{j,k} := \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k}{}' \Psi_{j,k-1} (e_t e_t{}' - L_t e_t{}' - e_t L_t{}') \Psi_{j,k-1} E_{j,k}$$

$$\tilde{H}_{j,k,\perp} := \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k,\perp}{}' \Psi_{j,k-1} (e_t e_t{}' - L_t e_t{}' - e_t L_t{}') \Psi_{j,k-1} E_{j,k,\perp}$$

$$\tilde{B}_{j,k} := \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k,\perp}{}' \Psi_{j,k-1} \hat{L}_t \hat{L}_t' \Psi_{j,k-1} E_{j,k} = \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,k}} E_{j,k,\perp}{}' \Psi_{j,k-1} (L_t - e_t) (L_t{}' - e_t{}') \Psi_{j,k-1} E_{j,k}$$

4) *Define*

$$\tilde{\mathcal{A}}_{j,k} := \begin{bmatrix} E_{j,k} & E_{j,k,\perp} \end{bmatrix} \begin{bmatrix} \tilde{A}_{j,k} & 0 \\ 0 & \tilde{A}_{j,k,\perp} \end{bmatrix} \begin{bmatrix} E_{j,k}{}' \\ E_{j,k,\perp}{}' \end{bmatrix}$$

$$\tilde{\mathcal{H}}_{j,k} := \begin{bmatrix} E_{j,k} & E_{j,k,\perp} \end{bmatrix} \begin{bmatrix} \tilde{H}_{j,k} & \tilde{B}'_{j,k} \\ \tilde{B}_{j,k} & \tilde{H}_{j,k,\perp} \end{bmatrix} \begin{bmatrix} E_{j,k}{}' \\ E_{j,k,\perp}{}' \end{bmatrix}$$

5) *From the above, it is easy to see that*

$$\tilde{\mathcal{A}}_{j,k} + \tilde{\mathcal{H}}_{j,k} = \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{\mathcal{I}}_{j,k}} \Psi_{j,k-1} \hat{L}_t \hat{L}_t' \Psi_{j,k-1}.$$

6) *Recall from Algorithm 3 that*

$$\tilde{\mathcal{A}}_{j,k} + \tilde{\mathcal{H}}_{j,k} = \frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{\mathcal{I}}_{j,k}} \Psi_{j,k-1} \hat{L}_t \hat{L}_t' \Psi_{j,k-1} \overset{EVD}{=} \begin{bmatrix} \hat{G}_{j,k} & \hat{G}_{j,k,\perp} \end{bmatrix} \begin{bmatrix} \Lambda_{j,k} & 0 \\ 0 & \Lambda_{j,k,\perp} \end{bmatrix} \begin{bmatrix} \hat{G}'_{j,k} \\ \hat{G}'_{j,k,\perp} \end{bmatrix}$$

is the EVD of $\tilde{\mathcal{A}}_{j,k} + \tilde{\mathcal{H}}_{j,k}$. Here Λ_k is a $\tilde{c}_{j,k} \times \tilde{c}_{j,k}$ diagonal matrix.

Definition 8.4: For $k = 1, 2, \cdots, \vartheta_j$, define

$$\tilde{\zeta}_{j,k} := \left\| \left(I - \sum_{i=1}^{k} \hat{G}_{j,i} \hat{G}'_{j,i} \right) G_{j,k} \right\|_2$$

This is the error in estimating $\operatorname{span}(G_{j,k})$ *after the k^{th} iteration of the cluster-PCA step.*

Remark 8.5:

1) *Notice that $\zeta_{j,0} = \|D_{j,\text{new}}\|_2$, $\zeta_{j,k} = \|D_{j,\text{new},k}\|_2$ and $\tilde{\zeta}_{j,k} = \|(I - \hat{G}_k \hat{G}'_k) D_{j,k}\|_2 = \|\Psi_{j,k} G_{j,k}\|_2$.*

2) *Notice from the algorithm that (i) $\hat{P}_{j,\text{new},k}$ is perpendicular to $\hat{P}_{j,*} = \hat{P}_{j-1}$; and (ii) $\hat{G}_{j,k}$ is perpendicular to $[\hat{G}_{j,1}, \hat{G}_{j,2}, \ldots \hat{G}_{j,k-1}]$.*

3) *For $t \in \mathcal{I}_{j,k}$, $P_{(t)} = P_j = [(P_{j-1} R_j \setminus P_{j,\text{old}}),\ P_{j,\text{new}}]$, $\hat{P}_{(t)} = [\hat{P}_{j-1}\ \hat{P}_{j,\text{new},k}]$ and*

$$\begin{aligned} SE_{(t)} &= \|(I - \hat{P}_{j-1} \hat{P}'_{j-1} - \hat{P}_{j,\text{new},k} \hat{P}'_{j,\text{new},k}) P_j\|_2 \\ &\leq \|(I - \hat{P}_{j-1} \hat{P}'_{j-1} - \hat{P}_{j,\text{new},k} \hat{P}'_{j,\text{new},k}) [P_{j-1}\ P_{j,\text{new}}]\|_2 \\ &\leq \zeta_{j,*} + \zeta_{j,k} \end{aligned}$$

for $k = 1, 2 \ldots K$. The last inequality uses the first item of this remark.

4) *For $t \in \tilde{\mathcal{I}}_{j,k}$, $P_{(t)} = P_j$, $\hat{P}_{(t)} = [\hat{P}_{j-1}\ \hat{P}_{j,\text{new},K}]$ and*

$$SE_{(t)} = SE_{(t_j + K\alpha - 1)} \leq \zeta_{j,*} + \zeta_{j,K}$$

5) *For $t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1}$, $P_{(t)} = P_j$,* $\operatorname{span}(P_j) = \operatorname{span}([G_{j,1}, \cdots, G_{j,\vartheta_j}])$, *$\hat{P}_{(t)} = \hat{P}_j = [\hat{G}_{j,1}, \cdots, \hat{G}_{j,\vartheta_j}]$, and*

$$SE_{(t)} = \zeta_{j+1,*} \leq \sum_{k=1}^{\vartheta_j} \tilde{\zeta}_{j,k}$$

The last inequality uses the first item of this remark.

Definition 8.6: Recall the definition of $\Phi_{j,k}$ from Definition 5.6. Define $\Phi_{(t)}$ as

$$\Phi_{(t)} := \begin{cases} \Phi_{j,k-1} & t \in \mathcal{I}_{j,k},\ k = 1, 2 \ldots K \\ \Phi_{j,K} & t \in \tilde{\mathcal{I}}_{j,k},\ k = 1, 2 \ldots \vartheta_j \\ \Phi_{j+1,0} & t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases}$$

Definition 8.7: Define the random variable

$$\tilde{X}_{j,k} := \{a_1, a_2, \cdots, a_{t_j + K\alpha + k\tilde{\alpha} - 1}\}$$

Definition 8.8: Define the sets

$$\tilde{\check{\Gamma}}_{j,k} := \{\tilde{X}_{j,k} : \tilde{\zeta}_{j,k} \leq \tilde{c}_{j,k} \zeta, \text{ and } \hat{T}_t = T_t \text{ for all } t \in \tilde{\mathcal{I}}_{j,k}\},\quad k = 1, 2, \ldots \vartheta_j,\ \ j = 1, 2, 3, \ldots J$$

$$\tilde{\check{\Gamma}}_{j,\vartheta_j+1} := \{X_{j+1,0} : \hat{T}_t = T_t \text{ for all } t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1}\},\quad j = 1, 2, 3, \ldots J$$

Define the sets

$$\tilde{\Gamma}_{j,0} := \Gamma_{j,K}$$

$$\tilde{\Gamma}_{j,k} := \tilde{\Gamma}_{j,k-1} \cap \tilde{\check{\Gamma}}_{j,k},\ \ k = 1, 2, \ldots \vartheta_j,\ \ j = 1, 2, 3, \ldots J$$

Definition 8.9: Define $\kappa_{s,D} := \max_j \max_k \kappa_s(D_{j,k})$

Remark 8.10: Conditioned on $\tilde{\Gamma}^e_{j,k-1}$, it is easy to see that

$$\begin{aligned} \kappa_{s,D} &:= \max_j \max_k \kappa_s(D_{j,k}) \\ &\leq \max_j \max_k (\kappa_s(G_{j,k}) + r\zeta) \\ &\leq \max_j \kappa_s(P_j) + r\zeta \leq \kappa^+_{s,D} := \kappa^+_{s,*} + r\zeta. \end{aligned}$$

In the above we have used $\kappa_s(G_{j,k}) \leq \kappa_s(P_j)$ and the same idea as in Lemma 6.10.

B. Two Main Lemmas

In this and the following subsections we remove the subscript j at most places. Also recall from earlier that $P_* = P_{j-1}$.

The theorem is a direct consequence of Lemmas 8.11 and 8.12 given below. Lemma 8.11 is a restatement of Lemmas 6.1 and 6.2 with using the new definition of ζ_*^+ and the new bound on ζ from Theorem 7.7. It summarizes the final conclusions of the addition step for ReProCS-cPCA.

Lemma 8.11 (Final Lemma for Addition Step): Assume that all the conditions in Theorem 7.7 holds. Also assume that $\mathbf{P}(\Gamma^e_{j,k-1}) > 0$. *Then*

1) $\zeta_0^+ = 1$, $\zeta_k^+ \le 0.6^k + 0.4c\zeta$ *for all* $k = 1, 2, \ldots K$;
2) $\mathbf{P}(\Gamma^e_{j,k} \mid \Gamma^e_{j,k-1}) \ge p_k(\alpha, \zeta) \ge p_K(\alpha, \zeta)$ *for all* $k = 1, 2, \ldots K$.

where ζ_k^+ *is defined in Definition 5.2 and* $p_k(\alpha, \zeta)$ *is defined in equation* (13).

The lemma below summarizes the final conclusions for the cluster-PCA step.

Lemma 8.12 (Final Lemma for Deletion (Cluster-PCA) Step): Assume that all the conditions in Theorem 7.7 hold. Also assume that $\mathbf{P}(\tilde{\Gamma}^e_{j,k-1}) > 0$. *Then,*

1) *for all* $k = 1, 2, \ldots \vartheta_j$, $\mathbf{P}(\tilde{\Gamma}^e_{j,k} \mid \tilde{\Gamma}^e_{j,k-1}) \ge \tilde{p}(\tilde{\alpha}, \zeta)$ *where* $\tilde{p}(\tilde{\alpha}, \zeta)$ *is defined in Lemma 8.19.*
2) $\mathbf{P}(\Gamma^e_{j+1,0} \mid \tilde{\Gamma}^e_{j,\vartheta_j}) = 1$.

Proof: Notice that $\mathbf{P}(\tilde{\Gamma}^e_{j,k} \mid \tilde{\Gamma}^e_{j,k-1}) = \mathbf{P}(\tilde{\zeta}_k \le \tilde{c}_k\zeta$ and $\hat{T}_t = T_t$ for all $t \in \tilde{\mathcal{I}}_{j,k} \mid \tilde{\Gamma}^e_{j,k-1})$ and $\mathbf{P}(\Gamma^e_{j+1,0} \mid \tilde{\Gamma}^e_{j,\vartheta_j}) = \mathbf{P}(\hat{T}_t = T_t$ for all $t \in \mathcal{I}_{j,\vartheta_j+1})$. The first claim of the lemma follows by combining Lemma 8.16 and the last claim of Lemma 6.4. The second claim follows using the last claim of Lemma 6.4. ■

Remark 8.13: Under the assumptions of Theorem 7.7,

$$\Gamma_{j,0} \cap (\cap_{k=1}^K \check{\Gamma}_{j,k}) \cap (\cap_{k=1}^{\vartheta_j} \tilde{\check{\Gamma}}_{j,k}) \subseteq \Gamma_{j+1,0}$$

This follows easily using Remark 8.5 and the fact that $\sum_k \tilde{c}_k = r_j \le r$.

Remark 8.14: Under the assumptions of Theorem 7.7, the following hold.

1) *For any* $k = 1, 2 \ldots \vartheta_j + 1$, $\tilde{\Gamma}^e_{j,k}$ *implies (i)* $\zeta_{j,K} \le c\zeta$, *(ii)* $\|\Phi_{j,K} P_j\|_2 \le (r + c)\zeta$.
   - (i) *follows from the first claim of Lemma 8.11 and the definition of* K, *(ii) follows using* $\|\Phi_{j,K} P_j\|_2 \le \|\Phi_{j,K}[P_*, P_{\text{new}}]\|_2 \le \zeta_* + \zeta_K \le \zeta_*^+ + \zeta_K^+ \le (r+c)\zeta$.
2) $\Gamma^e_{J+1,0}$ *implies (i)* $\zeta_{j,*} \le \zeta_*^+$ *for all* j, *(ii)* $\zeta_{j,k} \le 0.6^k + 0.4c\zeta$ *for all* $k = 1, \cdots, K$ *and all* j, *(iii)* $\zeta_{j,K} \le c\zeta$ *for all* j.

C. Proof of Theorem 7.7

Proof: From Remark 8.13,

$$\mathbf{P}(\Gamma^e_{j+1,0} | \Gamma^e_{j,0}) \ge \mathbf{P}(\check{\Gamma}^e_{j,1}, \ldots, \check{\Gamma}^e_{j,K}, \tilde{\check{\Gamma}}^e_{j,1}, \ldots, \tilde{\check{\Gamma}}^e_{j,\vartheta_j} | \Gamma_{j,0})$$
$$= \prod_{k=1}^{K} \mathbf{P}(\check{\Gamma}^e_{j,k} | \Gamma^e_{j,k-1}) \prod_{k=1}^{\vartheta_j} \mathbf{P}(\tilde{\check{\Gamma}}^e_{j,k} | \tilde{\Gamma}^e_{j,k-1})$$

Also, since $\Gamma_{j+1,0} \subseteq \Gamma_{j,0}$ using Lemma 2.12, $\mathbf{P}(\Gamma^e_{J+1,0} | \Gamma^e_{1,0}) = \prod_{j=1}^J \mathbf{P}(\Gamma^e_{j+1,0} | \Gamma^e_{j,0})$. Thus

$$\mathbf{P}(\Gamma^e_{J+1,0} | \Gamma^e_{1,0}) \ge \prod_{j=1}^{J} \times \left[\prod_{k=1}^{K} \mathbf{P}(\check{\Gamma}^e_{j,k} | \Gamma^e_{j,k-1}) \prod_{k=1}^{\vartheta_j} \mathbf{P}(\tilde{\check{\Gamma}}^e_{j,k} | \tilde{\Gamma}^e_{j,k-1}) \right]$$

Using Lemmas 8.11 and 8.12, and the fact that $p_k(\alpha, \zeta) \ge p_K(\alpha, \zeta)$, we get $\mathbf{P}(\Gamma^e_{J+1,0} | \Gamma_{1,0}) \ge p_K(\alpha, \zeta)^{KJ} \tilde{p}(\tilde{\alpha}, \zeta)^{\vartheta_{\max} J}$. Also, $\mathbf{P}(\Gamma^e_{1,0}) = 1$. This follows by the assumption on $\hat{P}_0$ and Lemma 6.4. Thus, $\mathbf{P}(\Gamma^e_{J+1,0}) \ge p_K(\alpha, \zeta)^{KJ} \tilde{p}(\tilde{\alpha}, \zeta)^{\vartheta_{\max} J}$.

Using the definitions of $\alpha_{\text{add}}(\zeta)$ and $\alpha_{\text{del}}(\zeta)$ and $\alpha \ge \alpha_{\text{add}}$ and $\tilde{\alpha} \ge \alpha_{\text{del}}$,

$$\mathbf{P}(\Gamma^e_{J+1,0}) \ge p_K(\alpha, \zeta)^{KJ} \tilde{p}(\tilde{\alpha}, \zeta)^{\vartheta_{\max} J}$$
$$\ge (1 - n^{-10})^2 \ge 1 - 2n^{-10}$$

The event $\Gamma^e_{J+1,0}$ implies that $\hat{T}_t = T_t$ for all $t < t_{J+1}$. Using Remark 8.5 and the last claim of Remark 8.14, $\Gamma^e_{J+1,0}$ implies that all the bounds on the subspace error hold. Using these, Remark 5.12, $\|a_{t,\text{new}}\|_2 \le \sqrt{c}\gamma_{\text{new},k}$ and $\|a_t\|_2 \le \sqrt{r}\gamma_*$, $\Gamma^e_{J+1,0}$ implies that all the bounds on $\|e_t\|_2$ hold (the bounds are obtained in Lemma 6.4).

Thus, all conclusions of the the result hold w.p. at least $1 - 2n^{-10}$. ■

D. A Lemma Needed for Getting High Probability Bounds on the Subspace Error

The following lemma is needed for bounding the subspace error, $\tilde{\zeta}_k$

Lemma 8.15: Assume that $\tilde{\zeta}_{k'} \le \tilde{c}_{k'}\zeta$ for $k' = 1, \cdots, k-1$. Then

1) $\|D_{\det,k}\|_2 = \|\Psi_{k-1} G_{\det,k}\|_2 \le r\zeta$.
2) $\|G_{\det,k} G_{\det,k}{}' - \hat{G}_{\det,k} \hat{G}'_{\det,k}\|_2 \le 2r\zeta$.
3) $0 < \sqrt{1 - r^2\zeta^2} \le \sigma_i(D_k) = \sigma_i(R_k) \le 1$. Thus, $\|D_k\|_2 = \|R_k\|_2 \le 1$ and $\|D_k^{-1}\|_2 = \|R_k^{-1}\|_2 \le 1/\sqrt{1 - r^2\zeta^2}$.
4) $\|D_{\text{undet},k}{}' E_k\|_2 = \|G_{\text{undet},k}{}' E_k\|_2 \le \frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}$.

Proof: The proof is given in Appendix D. ■

E. Bounding the Subspace Error, $\tilde{\zeta}_k$

Lemma 8.16 (High Probability Bound on $\tilde{\zeta}_k$*): Assume that the conditions of Theorem 7.7 hold. Then,*

$$\mathbf{P}(\tilde{\zeta}_k \le \tilde{c}_k\zeta \mid \tilde{\Gamma}^e_{j,k-1}) \ge \tilde{p}(\tilde{\alpha}, \zeta)$$

where $\tilde{p}(.)$ *is defined in Lemma 8.19.*

Proof: This follows by combining Lemma 8.17 and the last claim of Lemma 8.19, both of which are given below. ■

Lemma 8.17 (Bounding $\tilde{\zeta}_k{}^+$*):* If

$$f_{dec}(\tilde{g}_{\max}, \tilde{h}_{\max}, \kappa^+_{s,e}, \kappa^+_{s,*} + r\zeta) - \frac{f_{inc}(\tilde{g}_{\max}, \tilde{h}_{\max}, \kappa^+_{s,e}, \kappa^+_{s,*} + r\zeta)}{\tilde{c}_{\min}\zeta} > 0$$

then $f_{dec}(\tilde{g}_k, \tilde{h}_k, \kappa^+_{s,e}, \kappa^+_{s,*} + r\zeta) > 0$ and $\tilde{\zeta}_k^+ \le \tilde{c}_k\zeta$.

Proof: Recall from Definition 8.1 that $\tilde{\zeta}_k^+ := \frac{f_{inc}(\tilde{g}_k, \tilde{h}_k, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta)}{f_{dec}(\tilde{g}_k, \tilde{h}_k, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta)}$. Notice that $f_{inc}(.)$ is an increasing function of $\tilde{g}, \tilde{h}$, and $f_{dec}(.)$ is a decreasing function. Using the definition of $\tilde{g}_{\max}, \tilde{h}_{\max}, \tilde{c}_{\min}$ given in Assumption 7.3, the result follows. ∎

Lemma 8.18 (Bounding $\tilde{\zeta}_k$): If $\lambda_{\min}(\tilde{A}_k) - \lambda_{\max}(\tilde{A}_{k,\perp}) - \|\tilde{\mathcal{H}}_k\|_2 > 0$, then

$$\tilde{\zeta}_k \leq \frac{\|\tilde{\mathcal{H}}_k\|_2}{\lambda_{\min}(\tilde{A}_k) - \lambda_{\max}(\tilde{A}_{k,\perp}) - \|\tilde{\mathcal{H}}_k\|_2} \tag{14}$$

Proof: The proof is the same as that of Lemma 6.9. ∎

Lemma 8.19 (High Probability Bounds for Each of the Terms in the $\tilde{\zeta}_k$ Bound and for $\tilde{\zeta}_k$): Assume that the conditions of Theorem 7.7 hold. Also, assume that $\mathbf{P}(\tilde{\Gamma}_{j,k-1}^e) > 0$. Then, for all $1 \leq k \leq \vartheta_j$,

1) $\mathbf{P}(\lambda_{\min}(\tilde{A}_k) \geq \lambda_k^-(1 - r^2\zeta^2 - 0.1\zeta)|\tilde{\Gamma}_{j,k-1}^e) > 1 - \tilde{p}_1(\tilde{\alpha}, \zeta)$ with $\tilde{p}_1(\tilde{\alpha}, \zeta)$ given in (28).
2) $\mathbf{P}(\lambda_{\max}(\tilde{A}_{k,\perp}) \leq \lambda_k^-(\tilde{h}_k + r^2\zeta^2 f + 0.1\zeta)|\tilde{\Gamma}_{j,k-1}^e) > 1 - \tilde{p}_2(\tilde{\alpha}, \zeta)$ with $\tilde{p}_2(\tilde{\alpha}, \zeta)$ given in (30).
3) $\mathbf{P}(\|\tilde{\mathcal{H}}_k\|_2 \leq \lambda_k^- f_{inc}(\tilde{g}_k, \tilde{h}_k, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta)\ |\tilde{\Gamma}_{j,k-1}^e) \geq 1 - \tilde{p}_3(\tilde{\alpha}, \zeta)$ with $\tilde{p}_3(\tilde{\alpha}, \zeta)$ given in (35).
4) $\mathbf{P}(\lambda_{\min}(\tilde{A}_k) - \lambda_{\max}(\tilde{A}_{k,\perp}) - \|\tilde{\mathcal{H}}_k\|_2 \geq \lambda_k^- f_{dec}(\tilde{g}_k, \tilde{h}_k, \kappa_{s,e}^+, \kappa_{s,*}^+ + r\zeta)\ |\tilde{\Gamma}_{j,k-1}^e) \geq \tilde{p}(\tilde{\alpha}, \zeta) := 1 - \tilde{p}_1(\tilde{\alpha}, \zeta) - \tilde{p}_2(\tilde{\alpha}, \zeta) - \tilde{p}_3(\tilde{\alpha}, \zeta)$.
5) If $f_{dec}(\tilde{g}_k, \tilde{h}_k) > 0$, then $\mathbf{P}(\tilde{\zeta}_k \leq \tilde{\zeta}_k^+\ |\tilde{\Gamma}_{j,k-1}^e) \geq \tilde{p}(\tilde{\alpha}, \zeta)$

Proof: Recall that $f_{inc}(.)$, $f_{dec}(.)$ and $\tilde{\zeta}_k^+$ are defined in Definition 8.1. The proof of the first three claims is given in Appendix X-B. This proof uses Lemmas 8.15 and 6.4, Remark 8.10, and the Hoeffing corollaries. The fourth claim follows directly from the first three using the union bound on probabilities. The fifth claim follows from the fourth using Lemma 8.18. ∎

IX. Model Verification and Simulation Experiments

We first discuss model verification for real data in Sec IX-A. We then describe simulation experiments in Sec IX-B.

A. Model Verification for Real Data

We experimented with two background image sequence datasets. The first was a video of lake water motion. The second was a video of window curtains moving due to the wind. The curtain sequence is available at http://home.engineering.iastate.edu/chenlu/ReProCS/Fig2.mp4. For this sequence, the image size was $n = 5120$ and the number of images, $t_{\max} = 1755$. The lake sequence is available at http://home.engineering.iastate.edu/chenlu/ReProCS/ReProCS.htm (sequence 3). For this sequence, $n = 6480$ and the number of images, $t_{\max} = 1500$. Any given background image sequence will never be exactly low rank, but only approximately so. Let the data matrix with its empirical mean subtracted be $\mathcal{L}_{full}$. Thus $\mathcal{L}_{full}$ is a $n \times t_{\max}$ matrix. We first "low-rankified" this dataset by computing the EVD of $(1/t_{\max})\mathcal{L}_{full}\mathcal{L}_{full}'$; retaining the 90% eigenvectors' set (i.e. sorting eigenvalues in non-increasing order and retaining all eigenvectors until the sum of the corresponding eigenvalues exceeded 90% of the sum of all eigenvalues); and projecting the dataset into this subspace. To be precise, we computed P_{full} as the matrix containing these eigenvectors and we computed the low-rank matrix $\mathcal{L} = P_{full}P_{full}'\mathcal{L}_{full}$. Thus $\mathcal{L}$ is a $n \times t_{\max}$ matrix with $\text{rank}(\mathcal{L}) < \min(n, t_{\max})$. The curtains dataset is of size 5120×1755, but 90% of the energy is contained in only 34 directions, i.e. $\text{rank}(\mathcal{L}) = 34$. The lake dataset is of size 6480×1500 but 90% of the energy is contained in only 14 directions, i.e. $\text{rank}(\mathcal{L}) = 14$. This indicates that both datasets are indeed approximately low rank.

In practical data, the subspace does not just change as simply as in the model given in Sec. III-A. There are also rotations of the new and existing eigen-directions at each time which have not been modeled there. Moreover, with just one training sequence of a given type, it is not possible to compute $\text{Cov}(L_t)$ at each time t. Thus it is not possible to compute the delay between subspace change times. The only thing we can do is to assume that there may be a change every d frames, and that during these d frames the data is stationary and ergodic, and then estimate $\text{Cov}(L_t)$ for this period using a time average. We proceeded as follows. We took the first set of d frames, $\mathcal{L}_{1:d} := [L_1, L_2 \ldots L_d]$, estimated its covariance matrix as $(1/d)\mathcal{L}_{1:d}\mathcal{L}_{1:d}'$ and computed P_0 as the 99.99% eigenvectors' set. Also, we stored the lowest retained eigenvalue and called it λ^-. It is assumed that all directions with eigenvalues below λ^- are due to noise. Next, we picked the next set of d frames, $\mathcal{L}_{d+1:2d} := [L_{d+1}, L_{d+2}, \ldots L_{2d}]$; projected them perpendicular to P_0, i.e. computed $\mathcal{L}_{1,p} = (I - P_0P_0')\mathcal{L}_{d+1:2d}$; and computed $P_{1,\text{new}}$ as the eigenvectors of $(1/d)\mathcal{L}_{1,p}\mathcal{L}_{1,p}'$ with eigenvalues equal to or above λ^-. Then, $P_1 = [P_0, P_{1,\text{new}}]$. For the third set of d frames, we repeated the above procedure, but with P_0 replaced by P_1 and obtained P_2. A similar approach was repeated for each batch.

We used $d = 150$ for both the datasets. In each case, we computed $r_0 := \text{rank}(P_0)$, and $c_{\max} := \max_j \text{rank}(P_{j,\text{new}})$. For each batch of d frames, we also computed $a_{t,\text{new}} := P_{j,\text{new}}'L_t$, $a_{t,*} := P_{j-1}'L_t$ and $\gamma_* := \max_t \|a_t\|_\infty$. We got $c_{\max} = 3$ and $r_0 = 8$ for the lake sequence and $c_{\max} = 5$ and $r_0 = 29$ for the curtain sequence. Thus the ratio $c_{\max}/r_0$ is sufficiently small in both cases. In Fig 6, we plot $\|a_{t,\text{new}}\|_\infty/\gamma_*$ for one 150-frame period of the curtain sequence and for three 150-frame change periods of the lake sequence. If we take $\alpha = 40$, we observe that $\gamma_{\text{new}} := \max_j \max_{t_j \leq t < t_j+\alpha} ||a_{t,\text{new}}||_\infty = 0.125\gamma_*$ for the curtain sequence and $\gamma_{\text{new}} = 0.06\gamma_*$ for the lake sequence, i.e. the projection along the new directions is small for the initial α frames. Also, clearly, it increases slowly. In fact $\|a_{t,\text{new}}\|_\infty \leq \max(v^{k-1}\gamma_{\text{new}}, \gamma_*)$ for all $t \in \mathcal{I}_{j,k}$ also holds with $v = 1.5$ for the curtain sequence and $v = 1.8$ for the lake sequence.

Verifying the Clustering Assumption: We verified the clustering assumption for the lake video as follows. We first "low-rankified" it to 90% energy as explained above. Note that, with one sequence, it is not possible to estimate Λ_t (this would require an ensemble of sequences) and thus it is not possible to check if all Λ_t's in $[\tilde{t}_j, t_{j+1} - 1]$ are similar enough. However, by assuming that Λ_t is the same for a long enough sequence, one can estimate it using a time average and then verify if its

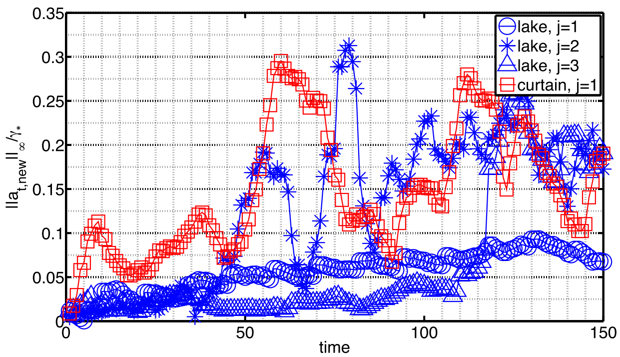

Fig. 6. Verification of slow subspace change. The figure is discussed in Sec IX-A.

eigenvalues are sufficiently clustered. When this was done, we observed that the clustering assumption holds with $\tilde{g}_{\max} = 7.2$, $\tilde{h}_{\max} = 0.34$ and $\vartheta_{\max} = 7$.

B. Simulation Experiments

The simulated data is generated as follows. The measurement matrix $\mathcal{M}_t := [M_1, M_2, \cdots, M_t]$ is of size 2048×4200. It can be decomposed as a sparse matrix $\mathcal{S}_t := [S_1, S_2, \cdots, S_t]$ plus a low rank matrix $\mathcal{L}_t := [L_1, L_2, \cdots, L_t]$.

The sparse matrix $\mathcal{S}_t := [S_1, S_2, \ldots, S_t]$ is generated as follows.

1) For $1 \le t \le t_{\text{train}} = 200$, $S_t = 0$.
2) For $t_{\text{train}} < t \le 5200$, S_t has s nonzero elements. The initial support $T_0 = \{1, 2, \ldots s\}$. Every Δ time instants we increment the support indices by 1. For example, for $t \in [t_{\text{train}} + 1, t_{\text{train}} + \Delta - 1]$, $T_t = T_0$, for $t \in [t_{\text{train}} + \Delta, t_{\text{train}} + 2\Delta - 1]$. $T_t = \{2, 3, \ldots s + 1\}$ and so on. Thus, the support set changes in a highly correlated fashion over time and this results in the matrix $\mathcal{S}_t$ being low rank. The larger the value of Δ, the smaller will be the rank of $\mathcal{S}_t$ (for $t > t_{\text{train}} + \Delta$).
3) The signs of the nonzero elements of S_t are ± 1 with equal probability and the magnitudes are uniformly distributed between 2 and 3. Thus, $S_{\min} = 2$.

The low rank matrix $\mathcal{L}_t := [L_1, L_2, \ldots, L_t]$ where $L_t := P_{(t)} a_t$ is generated as follows:

1) There are a total of $J = 2$ subspace change times, $t_1 = 301$ and $t_2 = 2701$. Let U be an $2048 \times (r_0 + c_{1,\text{new}} + c_{2,\text{new}})$ orthonormalized random Gaussian matrix.
   a) For $1 \le t \le t_1 - 1$, $P_{(t)} = P_0$ has rank r_0 with $P_0 = U_{[1,2,\ldots,r_0]}$.
   b) For $t_1 \le t \le t_2 - 1$, $P_{(t)} = P_1 = [P_0\ P_{1,\text{new}}]$ has rank $r_1 = r_0 + c_{1,\text{new}}$ with $P_{1,\text{new}} = U_{[r_0+1,\ldots,r_0+c_{1,\text{new}}]}$.
   c) For $t \ge t_2$, $P_{(t)} = P_2 = [P_1\ P_{2,\text{new}}]$ has rank $r_2 = r_1 + c_{2,\text{new}}$ with $P_{2,\text{new}} = U_{[r_0+c_{1,\text{new}}+1,\ldots,r_0+c_{1,\text{new}}+c_{2,\text{new}}]}$.
2) a_t is independent over t. The various $(a_t)_i$'s are also mutually independent for different i.
   a) For $1 \le t < t_1$, we let $(a_t)_i$ be uniformly distributed between $-\gamma_{i,t}$ and $\gamma_{i,t}$, where

$$\gamma_{i,t} = \begin{cases} 400 & \text{if } i = 1, 2, \ldots, r_0/4, \forall t, \\ 30 & \text{if } i = r_0/4+1, r_0/4+2, \ldots, r_0/2, \forall t. \\ 2 & \text{if } i = r_0/2+1, r_0/2+2, \ldots, 3r_0/4, \forall t. \\ 1 & \text{if } i = 3r_0/4+1, 3r_0/4+2, \ldots, r_0, \forall t. \end{cases}$$

   b) For $t_1 \le t < t_2$, $a_{t,*}$ is an r_0 length vector, $a_{t,\text{new}}$ is a $c_{1,\text{new}}$ length vector and $L_t := P_{(t)} a_t = P_1 a_t = P_0 a_{t,*} + P_{1,\text{new}} a_{t,\text{new}}$. $(a_{t,*})_i$ is uniformly distributed between $-\gamma_{i,t}$ and $\gamma_{i,t}$ and $a_{t,\text{new}}$ is uniformly distributed between $-\gamma_{r_1,t}$ and $\gamma_{r_1,t}$, where

$$\gamma_{r_1,t} = \begin{cases} 1.1^{k-1} & \text{if } t_1 + (k-1)\alpha \le t \\ & \le t_1 + k\alpha - 1 \\ & k = 1, 2, 3, 4 \\ 1.1^{4-1} = 1.331 & \text{if } t \ge t_1 + 4\alpha. \end{cases}$$

   c) For $t \ge t_2$, $a_{t,*}$ is an $r_1 = r_0 + c_{1,\text{new}}$ length vector, $a_{t,\text{new}}$ is a $c_{2,\text{new}}$ length vector and $L_t := P_{(t)} a_t = P_2 a_t = [P_0\ P_{1,\text{new}}] a_{t,*} + P_{2,\text{new}} a_{t,\text{new}}$. Also, $(a_{t,*})_i$ is uniformly distributed between $-\gamma_{i,t}$ and $\gamma_{i,t}$ for $i = 1, 2, \cdots, r_0$ and is uniformly distributed between $-\gamma_{r_1,t}$ and $\gamma_{r_1,t}$ for $i = r_0 + 1, \ldots r_1$. $a_{t,\text{new}}$ is uniformly distributed between $-\gamma_{r_2,t}$ and $\gamma_{r_2,t}$, where

$$\gamma_{r_2,t} = \begin{cases} 1.1^{k-1} & \text{if } t_2 + (k-1)\alpha \le t \\ & \le t_2 + k\alpha - 1, \\ & k = 1, 2, \cdots, 7 \\ 1.1^{7-1} = 1.7716 & \text{if } t \ge t_2 + 7\alpha. \end{cases}$$

Thus for the above model, $\gamma_* = 400$, $\gamma_{\text{new}} = 1$, $\lambda^+ = 53333$, $\lambda^- = 0.3333$ and $f := \frac{\lambda^+}{\lambda^-} = 1.6 \times 10^5$. Also, $S_{\min} = 2$.

We used $\mathcal{L}_{t_{\text{train}}} + \mathcal{N}_{t_{\text{train}}}$ as the training sequence to estimate $\hat{P}_0$. Here $\mathcal{N}_{t_{\text{train}}} = [N_1, N_2, \cdots, N_{t_{\text{train}}}]$ is i.i.d. random noise with each $(N_t)_i$ uniformly distributed between -10^{-3} and 10^{-3}. This is done to ensure that $\text{span}(\hat{P}_0) \neq \text{span}(P_0)$ but only approximates it.

Figure 7 shows the results of applying Algorithm 2 (ReProCS) to data generated according to the above model. The model parameters used were $s = 20$, $r_0 = 36$ and $c_{1,\text{new}} = c_{2,\text{new}} = 1$, and each subfigure corresponds to a different value of Δ. Because of the correlated support change, the $2048 \times t$ sparse matrix $\mathcal{S}_t = [S_1, S_2, \cdots, S_t]$ is rank deficient in either case, e.g. for Fig. 7(a), $\mathcal{S}_t$ has rank $69, 119, 169, 1219$ at $t = 300, 400, 500, 2600$; for Fig. 7(b), $\mathcal{S}_t$ has rank $29, 39, 49, 259$ at $t = 300, 400, 500, 2600$. We plot the subspace error $\text{SE}_{(t)}$ and the normalized error for S_t, $\frac{\|\hat{S}_t - S_t\|_2}{\|S_t\|_2}$ averaged over 100 Monte Carlo simulations. We also plot the ratio $d_t := \frac{\|I_{T_t}{}' D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$. This serves as a proxy for $\kappa_s(D_{j,\text{new},k})$ (which has exponential computational complexity). In fact, in our proofs, we only need this ratio to be small.

As can be seen from Figs. 7(a) and 7(b), the subspace error $\text{SE}_{(t)}$ of ReProCS decreased exponentially and stabilized

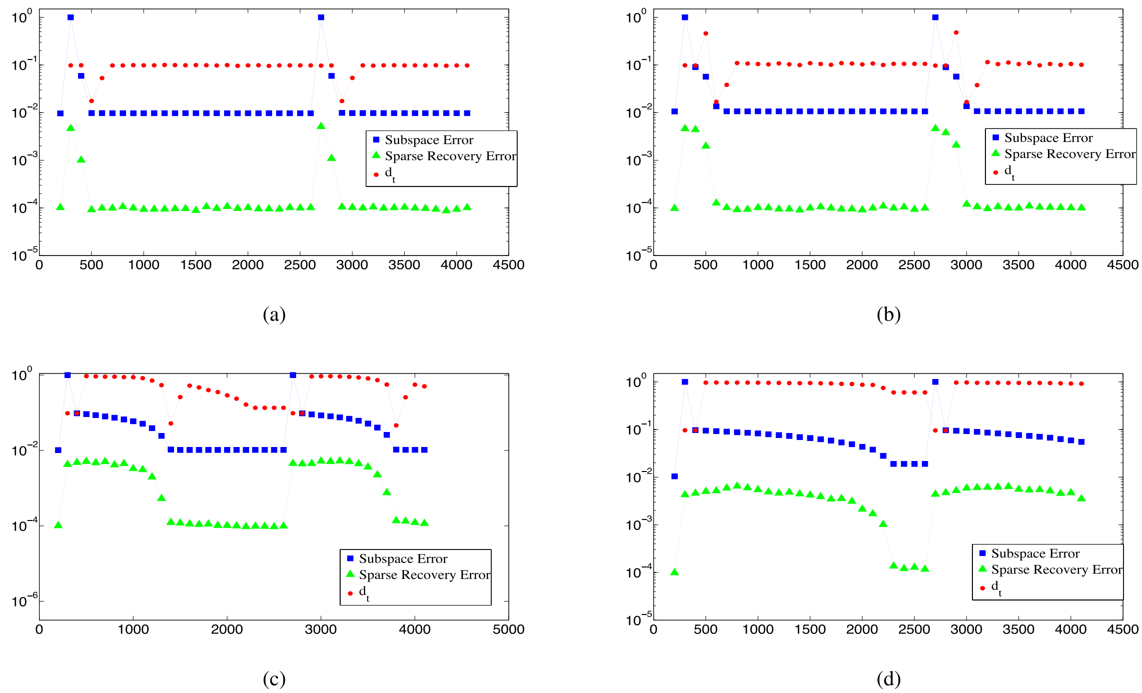

Fig. 7. Plots of d_t, SE and e_t for simulated data with $r_0 = 36$, $s = \max_t |T_t| = 20$. (a) $\Delta = 2$. (b) $\Delta = 10$. (c) $\Delta = 50$. (d) $\Delta = 100$.

after about 4 projection PCA update steps. The averaged normalized error for S_t followed a similar trend. In Fig. 7(b) where $\Delta = 10$, the subspace error $\mathrm{SE}_{(t)}$ also decreased but the decrease was a bit slower as compared to Fig. 7(a) where $\Delta = 2$.

In Fig. 7(d) we set $\Delta = 100$. In this case $\mathcal{S}_t$ is very low rank. The rank of $\mathcal{S}_t$ at $t = 300, 1000, 2600$ is $20, 27, 43$. We can see here that the subspace error decays rather slowly and does not return all the way to .01 within the $K\alpha$ frames.

Finally, if we set $\Delta = \infty$, the ratio $\frac{\|I_{T_t}{}' D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$ was 1 always. As a result, the subspace error and hence the reconstruction error of ReProCS did not decrease from its initial value at the subspace change time.

We also did one experiment in which we generated T_t of size $s = 100$ uniformly at random from all possible s-size subsets of $\{1, 2, \dots n\}$. T_t at different times t was also generated independently. In this case, the reconstruction error of ReProCS is $\frac{1}{5000}\sum_{t=201}^{5200} \frac{\|\hat{S}_t - S_t\|_2}{\|S_t\|_2} = 2.8472 \times 10^{-4}$. The error for PCP was 3.5×10^{-3} which is also quite small.

The data for figure 8 was generated the same as above except that we use the more general subspace model that allows for deletion of directions. Here, for $1 \le t \le t_1 - 1$, $P_{(t)} = P_0$ has rank r_0 with $P_0 = U_{[1,2,\cdots,36]}$. For $t_1 \le t \le t_2 - 1$, $P_{(t)} = P_1 = [P_0 \setminus P_{1,\text{old}}\ P_{1,\text{new}}]$ has rank $r_1 = r_0 + c_{1,\text{new}} - c_{1,\text{old}} = 34$ with $P_{1,\text{new}} = U_{[37]}$ and $P_{1,\text{old}} = U_{[9,18,36]}$. For $t \ge t_2$, $P_{(t)} = P_2 = [P_1 \setminus P_{2,\text{old}}\ P_{2,\text{new}}]$ has rank $r_2 = r_1 + c_{2,\text{new}} - c_{2,\text{old}} = 32$ with $P_{2,\text{new}} = U_{[38]}$ and $P_{1\text{old}} = U_{[8,17,35]}$. Again, we average over 100 Monte Carlo simulations.

As can be seen from Figure 8, the normalized sparse recovery error of ReProCS and ReProCS-cPCA decreased exponentially and stabilized. Furthermore, ReProCS-cPCA outperforms over ReProCS greatly when deletion steps are done.

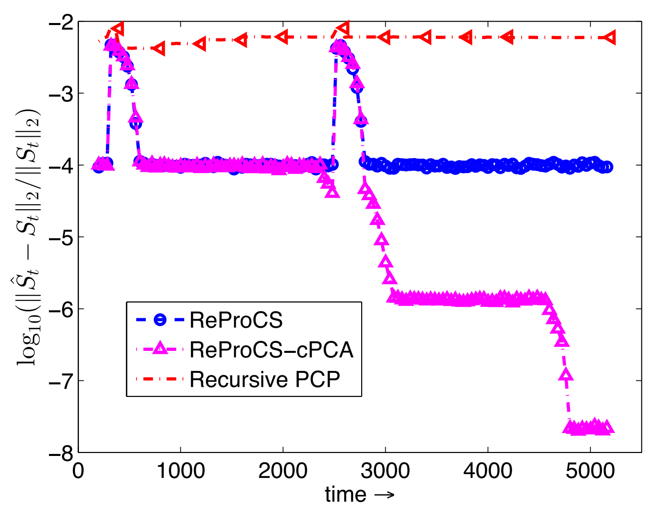

Fig. 8. Reconstruction errors of S_t with $r_0 = 36$, $s = \max_t |T_t| = 20$. The times at which PCP is done are marked by red triangles. Δ : 10, comparing PCP with ReProCS and ReProCS-cPCA.

We also compared against PCP [6]. At every $t = t_j + 4k\alpha$, we solved (1) with $\lambda = 1/\sqrt{\max(n, t)}$ as suggested in [6] to recover $\mathcal{S}_t$ and $\mathcal{L}_t$. We used the estimates of S_t for the last 4α frames as the final estimates of $\hat{S}_t$. So, the $\hat{S}_t$ for $t = t_j + 1, \dots t_j + 4\alpha$ is obtained from PCP done at $t = t_j + 4\alpha$, the $\hat{S}_t$ for $t = t_j + 4\alpha + 1, \dots t_j + 8\alpha$ is obtained from PCP done at $t = t_j + 8\alpha$ and so on. Because of the correlated support change, the error of PCP was larger in both cases.

X. Conclusions and Future Work

In this 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 (noise that is dense and lies in a "slowly changing" low dimensional subspace). We analyzed a novel solution approach called Recursive Projected CS or ReProCS that was introduced in our earlier work [1], [25], and [26]. The ReProCS algorithm that we analyze assumes knowledge of the subspace change model on the L_t's. We showed that, under mild assumptions and a denseness assumption on the currently unestimated subspace, span$(D_{j,\text{new},k})$ (this assumption depends on algorithm estimates), w.h.p., ReProCS can exactly recover the support set of S_t at all times; the reconstruction errors of both S_t and L_t are upper bounded by a time-invariant and small value; and after every subspace change time, w.h.p., the subspace recovery error decays to a small enough value within a finite delay. The most important open question that is being addressed in ongoing work is how to make our result a correctness result, i.e. how to remove the denseness assumption on $D_{j,\text{new},k}$ (see a forthcoming paper). Two other issues being studied are (i) how to get a result for the correlated L_t's case [48], and (ii) how to analyze the ReProCS algorithm when subspace change times are not known. Finally, an open question is how to to bound the sparse recovery error even when the support set is not exactly recovered. The undersampled measurements' case is also being studied [49].

Appendix A
Proofs of Preliminary Lemmas

Proof of Lemma 2.10: Because P, Q and $\hat{P}$ are basis matrix, $P'P = I$, $Q'Q = I$ and $\hat{P}'\hat{P} = I$.

1) Using $P'P = I$ and $\|M\|_2^2 = \|MM'\|_2$, $\|(I - \hat{P}\hat{P}')PP'\|_2 = \|(I - \hat{P}\hat{P}')P\|_2$. Similarly, $\|(I - PP')\hat{P}\hat{P}'\|_2 = \|(I - PP')\hat{P}\|_2$. Let $D_1 = (I - \hat{P}\hat{P}')PP'$ and let $D_2 = (I - PP')\hat{P}\hat{P}'$. Notice that $\|D_1\|_2 = \sqrt{\lambda_{\max}(D_1'D_1)} = \sqrt{\|D_1'D_1\|_2}$ and $\|D_2\|_2 = \sqrt{\lambda_{\max}(D_2'D_2)} = \sqrt{\|D_2'D_2\|_2}$. So, in order to show $\|D_1\|_2 = \|D_2\|_2$, it suffices to show that $\|D_1'D_1\|_2 = \|D_2'D_2\|_2$. Let $P'\hat{P} \overset{SVD}{=} U\Sigma V'$. Then, $D_1'D_1 = P(I - P'\hat{P}\hat{P}'P)P' = PU(I - \Sigma^2)U'P'$ and $D_2'D_2 = \hat{P}(I - \hat{P}'PP'\hat{P})\hat{P}' = \hat{P}V(I - \Sigma^2)V'\hat{P}'$ are the compact SVD's of $D_1'D_1$ and $D_2'D_2$ respectively. Therefore, $\|D_1'D_1\| = \|D_2'D_2\|_2 = \|I - \Sigma^2\|_2$ and hence $\|(I - \hat{P}\hat{P}')PP'\|_2 = \|(I - PP')\hat{P}\hat{P}'\|_2$.
2) $\|PP' - \hat{P}\hat{P}'\|_2 = \|PP' - \hat{P}\hat{P}'PP' + \hat{P}\hat{P}'PP' - \hat{P}\hat{P}'\|_2 \le \|(I - \hat{P}\hat{P}')PP'\|_2 + \|(I - PP')\hat{P}\hat{P}'\|_2 = 2\zeta_*$.
3) Since $Q'P = 0$, then $\|Q'\hat{P}\|_2 = \|Q'(I - PP')\hat{P}\|_2 \le \|(I - PP')\hat{P}\|_2 = \zeta_*$.
4) Let $M = (I - \hat{P}\hat{P}')Q)$. Then $M'M = Q'(I - \hat{P}\hat{P}')Q$ and so $\sigma_i((I - \hat{P}\hat{P}')Q) = \sqrt{\lambda_i(Q'(I - \hat{P}\hat{P}')Q)}$. Clearly, $\lambda_{\max}(Q'(I - \hat{P}\hat{P}')Q) \le 1$. By Weyl's Theorem, $\lambda_{\min}(Q'(I - \hat{P}\hat{P}')Q) \ge 1 - \lambda_{\max}(Q'\hat{P}\hat{P}'Q) = 1 - \|Q'\hat{P}\|_2^2 \ge 1 - \zeta_*^2$. Therefore, $\sqrt{1 - \zeta_*^2} \le \sigma_i((I - \hat{P}\hat{P}')Q) \le 1$.

For the case when P and $\hat{P}$ are not the same size, the proof of 1 is used, but Σ^2 becomes $\Sigma\Sigma'$ for D_1 and $\Sigma'\Sigma$ for D_2. Since Σ is of size $r_1 \times r_2$, $\Sigma\Sigma'$ will be of size $r_1 \times r_1$ and $\Sigma'\Sigma$ will be of size $r_2 \times r_2$. Because $r_1 \le r_2$, every singular value of $D_1'D_1$ will be a singualr value of $D_2'D_2$ (using the SVD as in the proof of 1 above). Using the characterization of the matrix 2-norm as the largest singluar value, $\|D_1'D_1\|_2 \le \|D_2'D_2\|$. ■

Proof of Lemma 2.11: It is easy to see that $\mathbf{P}(\mathcal{B}^e, \mathcal{C}^e) = \mathbf{E}[\mathbb{I}_{\mathcal{B}}(X,Y)\mathbb{I}_{\mathcal{C}}(X)]$. If $\mathbf{E}[\mathbb{I}_{\mathcal{B}}(X,Y)|X] \ge p$ for all $X \in \mathcal{C}$, this means that $\mathbf{E}[\mathbb{I}_{\mathcal{B}}(X,Y)|X]\mathbb{I}_{\mathcal{C}}(X) \ge p\mathbb{I}_{\mathcal{C}}(X)$. This, in turn, implies that

$$\mathbf{P}(\mathcal{B}^e, \mathcal{C}^e) = \mathbf{E}[\mathbb{I}_{\mathcal{B}}(X,Y)\mathbb{I}_{\mathcal{C}}(X)] = \mathbf{E}[\mathbf{E}[\mathbb{I}_{\mathcal{B}}(X,Y)|X]\mathbb{I}_{\mathcal{C}}(X)] \ge p\mathbf{E}[\mathbb{I}_{\mathcal{C}}(X)].$$

Recall from Definition 2.4 that $\mathbf{P}(\mathcal{B}^e|X) = \mathbf{E}[\mathbb{I}_{\mathcal{B}}(X,Y)|X]$ and $\mathbf{P}(\mathcal{C}^e) = \mathbf{E}[\mathbb{I}_{\mathcal{C}}(X)]$. Thus, we conclude that if $\mathbf{P}(\mathcal{B}^e|X) \ge p$ for all $X \in \mathcal{C}$, then $\mathbf{P}(\mathcal{B}^e, \mathcal{C}^e) \ge p\mathbf{P}(\mathcal{C}^e)$. Using the definition of $\mathbf{P}(\mathcal{B}^e|\mathcal{C}^e)$, the claim follows. ■

Proof of Corollary 2.14:

1) Since, for any $X \in \mathcal{C}$, conditioned on X, the Z_t's are independent, the same is also true for $Z_t - g(X)$ for any function of X. Let $Y_t := Z_t - \mathbf{E}(Z_t|X)$. Thus, for any $X \in \mathcal{C}$, conditioned on X, the Y_t's are independent. Also, clearly $\mathbf{E}(Y_t|X) = 0$. Since for all $X \in \mathcal{C}$, $\mathbf{P}(b_1 I \preceq Z_t \preceq b_2 I|X) = 1$ and since $\lambda_{\max}(.)$ is a convex function, and $\lambda_{\min}(.)$ is a concave function, of a Hermitian matrix, thus $b_1 I \preceq \mathbf{E}(Z_t|X) \preceq b_2 I$ w.p. one for all $X \in \mathcal{C}$. Therefore, $\mathbf{P}(Y_t^2 \preceq (b_2 - b_1)^2 I|X) = 1$ for all $X \in \mathcal{C}$. Thus, for Theorem 2.13, $\sigma^2 = \|\sum_t (b_2 - b_1)^2 I\|_2 = \alpha(b_2 - b_1)^2$. For any $X \in \mathcal{C}$, applying Theorem 2.13 for $\{Y_t\}$'s conditioned on X, we get that, for any $\epsilon > 0$,
$$\mathbf{P}\left(\lambda_{\max}\left(\frac{1}{\alpha}\sum_t Y_t\right) \le \epsilon \Big| X\right) > 1 - n\exp\left(\frac{-\alpha\epsilon^2}{8(b_2 - b_1)^2}\right) \text{ for all } X \in \mathcal{C}$$
By Weyl's theorem, $\lambda_{\max}(\frac{1}{\alpha}\sum_t Y_t) = \lambda_{\max}(\frac{1}{\alpha}\sum_t (Z_t - \mathbf{E}(Z_t|X)) \ge \lambda_{\max}(\frac{1}{\alpha}\sum_t Z_t) + \lambda_{\min}(\frac{1}{\alpha}\sum_t -\mathbf{E}(Z_t|X))$. Since $\lambda_{\min}(\frac{1}{\alpha}\sum_t -\mathbf{E}(Z_t|X)) = -\lambda_{\max}(\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X)) \ge -b_4$, thus $\lambda_{\max}(\frac{1}{\alpha}\sum_t Y_t) \ge \lambda_{\max}(\frac{1}{\alpha}\sum_t Z_t) - b_4$. Therefore,
$$\mathbf{P}\left(\lambda_{\max}\left(\frac{1}{\alpha}\sum_t Z_t\right) \le b_4 + \epsilon \Big| X\right) > 1 - n\exp\left(\frac{-\alpha\epsilon^2}{8(b_2 - b_1)^2}\right) \text{ for all } X \in \mathcal{C}$$
2) Let $Y_t = \mathbf{E}(Z_t|X) - Z_t$. As before, $\mathbf{E}(Y_t|X) = 0$ and conditioned on any $X \in \mathcal{C}$, the Y_t's are independent and $\mathbf{P}(Y_t^2 \preceq (b_2 - b_1)^2 I|X) = 1$. As before, applying Theorem 2.13, we get that for any $\epsilon > 0$,
$$\mathbf{P}\left(\lambda_{\max}\left(\frac{1}{\alpha}\sum_t Y_t\right) \le \epsilon \Big| X\right) > 1 - n\exp\left(\frac{-\alpha\epsilon^2}{8(b_2 - b_1)^2}\right) \text{ for all } X \in \mathcal{C}$$
By Weyl's theorem, $\lambda_{\max}(\frac{1}{\alpha}\sum_t Y_t) = \lambda_{\max}(\frac{1}{\alpha}\sum_t (\mathbf{E}(Z_t|X) - Z_t)) \ge \lambda_{\min}(\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X)) +$

$\lambda_{\max}(\frac{1}{\alpha}\sum_t -Z_t) = \lambda_{\min}(\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X)) - \lambda_{\min}(\frac{1}{\alpha}\sum_t Z_t) \geq b_3 - \lambda_{\min}(\frac{1}{\alpha}\sum_t Z_t)$ Therefore, for any $\epsilon > 0$,

$$\mathbf{P}\left(\lambda_{\min}\left(\frac{1}{\alpha}\sum_t Z_t\right) \geq b_3 - \epsilon \middle| X\right) \geq 1 - n\exp\left(\frac{-\alpha\epsilon^2}{8(b_2-b_1)^2}\right) \text{ for all } X \in \mathcal{C}$$

■

Proof of Corollary 2.15: Define the dilation of an $n_1 \times n_2$ matrix M as dilation$(M) := \begin{bmatrix} 0 & M' \\ M & 0 \end{bmatrix}$. Notice that this is an $(n_1+n_2)\times(n_1+n_2)$ Hermitian matrix [32]. As shown in [32, eq. 2.12],

$$\lambda_{\max}(\text{dilation}(M)) = \|\text{dilation}(M)\|_2 = \|M\|_2 \qquad (15)$$

Thus, the corollary assumptions imply that $\mathbf{P}(\|\text{dilation}(Z_t)\|_2 \leq b_1|X) = 1$ for all $X \in \mathcal{C}$. Thus, $\mathbf{P}(-b_1 I \preceq \text{dilation}(Z_t) \preceq b_1 I|X) = 1$ for all $X \in \mathcal{C}$. Using (15), the corollary assumptions also imply that $\frac{1}{\alpha}\sum_t \mathbf{E}(\text{dilation}(Z_t)|X) = \text{dilation}(\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X)) \preceq b_2 I$ for all $X \in \mathcal{C}$. Finally, Z_t's conditionally independent given X, for any $X \in \mathcal{C}$, implies that the same thing also holds for dilation(Z_t)'s. Thus, applying Corollary 2.14 for the sequence $\{\text{dilation}(Z_t)\}$, we get that,

$$\mathbf{P}\left(\lambda_{\max}\left(\frac{1}{\alpha}\sum_t \text{dilation}(Z_t)\right) \leq b_2 + \epsilon \middle| X\right) \geq 1 - (n_1+n_2)\exp\left(\frac{-\alpha\epsilon^2}{32 b_1^2}\right) \text{ for all } X \in \mathcal{C}$$

Using (15), $\lambda_{\max}(\frac{1}{\alpha}\sum_t \text{dilation}(Z_t)) = \lambda_{\max}(\text{dilation}(\frac{1}{\alpha}\sum_t Z_t)) = \|\frac{1}{\alpha}\sum_t Z_t\|_2$ and this gives the final result. ■

Proof of Lemma 3.7: Let $A = I - PP'$. By definition, $\delta_s(A) := \max\{\max_{|T|\leq s}(\lambda_{\max}(A_T' A_T) - 1), \max_{|T|\leq s}(1 - \lambda_{\min}(A_T' A_T))\}$. Notice that $A_T' A_T = I - I_T' PP' I_T$. Since $I_T' PP' I_T$ is p.s.d., by Weyl's theorem, $\lambda_{\max}(A_T' A_T) \leq 1$. Since $\lambda_{\max}(A_T' A_T) - 1 \leq 0$ while $1 - \lambda_{\min}(A_T' A_T) \geq 0$, thus,

$$\delta_s(I - PP') = \max_{|T|\leq s}\left(1 - \lambda_{\min}(I - I_T' PP' I_T)\right) \qquad (16)$$

By Definition, $\kappa_s(P) = \max_{|T|\leq s}\frac{\|I_T' P\|_2}{\|P\|_2} = \max_{|T|\leq s}\|I_T' P\|_2$. Notice that $\|I_T' P\|_2^2 = \lambda_{\max}(I_T' PP' I_T) = 1 - \lambda_{\min}(I - I_T' PP' I_T)$ [3], and so

$$\kappa_s^2(P) = \max_{|T|\leq s}\left(1 - \lambda_{\min}(I - I_T' PP' I_T)\right) \qquad (17)$$

From (16) and (17), we get $\delta_s(I - PP') = \kappa_s^2(P)$. ■

[3]This follows because $B = I_T' PP' I_T$ is a Hermitian matrix. Let $B = U\Sigma U'$ be its EVD. Since $UU' = I$, $\lambda_{\min}(I - B) = \lambda_{\min}(U(I-\Sigma)U') = \lambda_{\min}(I - \Sigma) = 1 - \lambda_{\max}(\Sigma) = 1 - \lambda_{\max}(B)$.

Appendix B
The Need for Projection PCA

A. Projection-PCA vs Standard PCA

The reason that we cannot use standard PCA for subspace update in our work is because, in our case, the error $e_t = L_t - \hat{L}_t$ in the observed data vector $\hat{L}_t$ is correlated with the true data vector L_t; and the condition number of Cov$[L_t]$ is large (see Remark 3.4). In other works that study finite sample PCA, see [33] and references therein, the large condition number does not cause a problem because they assume that the error/noise (e_t) is uncorrelated with the true data vector (L_t). Moreover, e_t or L_t or both are zero mean (which we have too). Thus, the dominant term in the perturbation of the estimated covariance matrix, $(1/\alpha)\sum_t \hat{L}_t\hat{L}_t'$ w.r.t. the true one is $(1/\alpha)\sum_t e_t e_t'$. For α large enough, the other two terms $(1/\alpha)\sum_t L_t e_t'$ and its transpose are close to zero w.h.p. due to law or large numbers. Thus, the subspace error bound obtained using the $\sin\theta$ theorem and the matrix Hoeffding inequality, will depend, w.h.p., only on the ratio of the maximum eigenvalue of Cov$[e_t]$ to the smallest eigenvalue of Cov$[L_t]$. The probability with which this bound holds depends on f, however the probability can be made large by increasing the number of data points α. However, in our case, because e_t and L_t are correlated, this strategy does not work. We explain this below.

In this discussion, we remove the subscript j. Also, let $P_* := P_{j-1}$, $\hat{P}_* := \hat{P}_{j-1}$, $r_* = \text{rank}(P_*)$. Consider $t = t_j + k\alpha - 1$ when the k^{th} projection PCA or PCA is done. Since the error $e_t = L_t - \hat{L}_t$ is correlated with L_t, the dominant terms in the perturbation matrix seen by PCA are $(1/(t_j + k\alpha))\sum_{t=1}^{t_j+k\alpha-1} L_t e_t'$ and its transpose, while for projection PCA, they are $(1/\alpha)\Phi_0\sum_{t\in\mathcal{I}_{j,k}} L_t e_t'\Phi_0$ and its transpose. The magnitude of L_t can be large. The magnitude of e_t is smaller than a constant times that of L_t. The constant is less than one but, at $t = t_j + \alpha - 1$, it is not negligible. Thus, the norm of the perturbation seen by PCA at this time may not be small. As a result, the bound on the subspace error, $\text{SE}_{(t)}$, obtained by applying the $\sin\theta$ theorem may be more than one (and hence meaningless since by definition $\text{SE}_{(t)} \leq 1$). For projection PCA, because of Φ_0, the perturbation is much smaller and hence so is the bound on $\text{SE}_{(t)}$.

Let $\text{SE}_k := \text{SE}_{(t_j+k\alpha-1)} = \text{SE}_{(t)}$ denote the subspace error for $t \in \mathcal{I}_{j,k}$. Consider $k = 1$ first. For PCA, we can show that $\text{SE}_1 \lesssim \check{C}\kappa_s^+ g^+ + \check{C}' f\zeta_*^+$ for constants $\check{C}, \check{C}'$ that are more than one but not too large. Here g^+ is the upper bound on the condition number of Cov$(a_{t,\text{new}})$) and it is valid to assume that g^+ is small so that $\check{C}\kappa_s^+ g^+ < 1$. However, f is a bound on the maximum condition number of Cov$(a_t) =$ Cov(L_t) and this can be large. When it is, the second term may not be less than one. On the other hand, for projection PCA, we have $\text{SE}_k \leq \zeta_k + \zeta_* \leq \zeta_k^+ + \zeta_*^+$ with $\zeta_*^+ = r\zeta$, and $\zeta_k^+ \approx \check{C}\kappa_s^+ g^+ \zeta_{k-1}^+ + \check{C}' f(\zeta_*^+)^2$ and $\zeta_0^+ = 1$. Thus $\text{SE}_1 \lesssim \check{C}\kappa_s^+ g^+ + \check{C}' f(\zeta_*^+)^2 + \zeta_*^+$. The first term in this bound is similar to that of PCA, but the second term is much smaller. The third term is negligibly small. Thus, in this case, it is easier to ensure that the bound is less than one.

Moreover, our goal is to show that within a finite delay after a subspace change time, the subspace error decays down from one to a value proportional to ζ. For projection PCA, this can be done because we can separately bound the subspace error of the existing subspace, ζ_*, and of the newly added one, ζ_k, and then bound the total subspace error, $\text{SE}_{(t)}$, by $\zeta_* + \zeta_k$ for $t \in \mathcal{I}_{j,k}$. Assuming that, by $t = t_j$, ζ_* is small enough, i.e. $\zeta_* \le r_*\zeta$ with $\zeta < 0.00015/r^2 f$, we can show that within K iterations, ζ_k also becomes small enough so that $\text{SE}_{(t)} \le (r_* + c)\zeta$. However, for PCA, it is not possible to separate the subspace error in this fashion. For $k > 1$, all we can claim is that $\text{SE}_k \lesssim \check{C}\kappa_s^+ f\ \text{SE}_{k-1}$. Since f can be large (larger than $1/\kappa_s^+$), this cannot be used to show that SE_k decreases with k.

B. Why Not Use All $k\alpha$ Frames at $t = t_j + k\alpha - 1$

Another possible way to implement projection PCA is to use the past $k\alpha$ estimates $\hat{L}_t$ at the k^{th} projection PCA time, $t = t_j + k\alpha - 1$. This may actually result in an improved algorithm. We believe that it can also be analyzed using the approaches developed in this paper. However, the analysis will be more complicated. We briefly try to explain why. The perturbation seen at $t = t_j + k\alpha - 1$, $\mathcal{H}_k$, will now satisfy $\mathcal{H}_k \approx (1/(k\alpha))\sum_{k'=1}^{k}\sum_{t\in\mathcal{I}_{j,k'}}\Phi_0(-L_t e_t' - e_t L_t' + e_t e_t')\Phi_0$ instead of just being approximately equal to the last ($k' = k$) term. Bounds on each of these terms will hold with a different probability. Thus, proving a lemma similar to Lemma 6.11 will be more complicated.

APPENDIX C
PROOF OF LEMMA 6.11

For convenience, we will use $\frac{1}{\alpha}\sum_t$ to denote $\frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}}$. The proof follows using the following key facts and the Hoeffding corollaries.

Fact 10.1: Under the assumptions of Theorem 4.2 the following are true.

1) *The matrices D_{new}, R_{new}, E_{new}, D_*, $D_{\text{new},k-1}$, Φ_{k-1} are functions of the r.v. $X_{j,k-1}$. Since $X_{j,k-1}$ is independent of any a_t for $t \in \mathcal{I}_{j,k}$ the same is true for the matrices D_{new}, R_{new}, E_{new}, D_*, $D_{\text{new},k-1}$, Φ_{k-1}.*

   All terms that we bound for the first two claims of the lemma are of the form $\frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}} Z_t$ where $Z_t = f_1(X_{j,k-1})Y_t f_2(X_{j,k-1})$, Y_t is a sub-matrix of $a_t a_t'$ and $f_1(.)$ and $f_2(.)$ are functions of $X_{j,k-1}$. Thus, conditioned on $X_{j,k-1}$, the Z_t's are mutually independent. (Recall that we assume independence of the a_t's.

   All the terms that we bound for the third claim contain e_t. Using Lemma 6.4, conditioned on $X_{j,k-1}$, e_t satisfies (10) w.p. one whenever $X_{j,k-1} \in \Gamma_{j,k-1}$. Using (10), it is easy to see that all these terms are also of the above form whenever $X_{j,k-1} \in \Gamma_{j,k-1}$.

   Thus, conditioned on $X_{j,k-1}$, the Z_t's for all the above terms are mutually independent, whenever $X_{j,k-1} \in \Gamma_{j,k-1}$.

2) *It is easy to see that $\|\Phi_{k-1}P_*\|_2 \le \zeta_*$, $\zeta_0 = \|D_{\text{new}}\|_2 \le 1$, $\Phi_0 D_{\text{new}} = \Phi_0' D_{\text{new}} = D_{\text{new}}$, $\|R_{\text{new}}\| \le 1$, $\|(R_{\text{new}})^{-1}\| \le 1/\sqrt{1-\zeta_*^2}$, $E_{\text{new},\perp}{}' D_{\text{new}} = 0$, and $\|E_{\text{new}}{}'\Phi_0 e_t\| = \|(R_{\text{new}}')^{-1}D_{\text{new}}'\Phi_0 e_t\| = \|(R_{\text{new}})^{-1}D_{\text{new}}' e_t\| \le \|(R_{\text{new}}')^{-1}D_{\text{new}}' I_{T_t}\|\|e_t\| \le \frac{\kappa_s(D_{\text{new}})}{\sqrt{1-\zeta_*^2}}\|e_t\|$. The bounds on $\|R_{\text{new}}\|$ and $\|(R_{\text{new}})^{-1}\|$ follow using Lemma 2.10 and the fact that $\sigma_i(R_{\text{new}}) = \sigma_i(D_{\text{new}})$.*

3) *$X_{j,k-1} \in \Gamma_{j,k-1}$ implies that*
   a) *$\zeta_{j,*} \le \zeta_*^+$ (By definition of $\Gamma_{j,k-1}$ (Definition 5.11))*
   b) *$\zeta_{k-1} \le \zeta_{k-1}^+ \le 0.6^{k-1} + 0.4c\zeta$ (This follows by the definition of $\Gamma_{j,k-1}$ and Lemma 6.1.)*

4) *Item 3 implies that conditioned on $X_{j,k-1} \in \Gamma_{j,k-1}$*
   a) *$\kappa_s(D_{\text{new}}) \le \kappa_s^+$ (follows by Lemma 6.10),*
   b) *$\lambda_{\min}(R_{\text{new}}R_{\text{new}}{}') \ge 1 - (\zeta_*^+)^2$ (follows from Lemma 2.10 and the fact that $\sigma_{\min}(R_{\text{new}}) = \sigma_{\min}(D_{\text{new}})$),*
   c) *$\|I_{T_t}{}'\Phi_{k-1}P_*\|_2 \le \|\Phi_{k-1}P_*\|_2 \le \zeta_{j,*} \le \zeta_{j,*}^+$,*
   d) *$\|I_{T_t}{}'D_{\text{new},k-1}\|_2 \le \kappa_s(D_{\text{new},k-1})\zeta_{k-1} \le \kappa_s^+\zeta_{k-1}^+$.*

5) *By Weyl's theorem (Theorem 2.8), for a sequence of matrices B_t, $\lambda_{\min}(\sum_t B_t) \ge \sum_t \lambda_{\min}(B_t)$ and $\lambda_{\max}(\sum_t B_t) \le \sum_t \lambda_{\max}(B_t)$.*

Proof: Consider $A_k := \frac{1}{\alpha}\sum_t E_{\text{new}}{}'\Phi_0 L_t L_t{}'\Phi_0 E_{\text{new}}$. Notice that $E_{\text{new}}{}'\Phi_0 L_t = R_{\text{new}}a_{t,\text{new}} + E_{\text{new}}{}'D_* a_{t,*}$. Let $Z_t = R_{\text{new}}a_{t,\text{new}}a_{t,\text{new}}{}'R_{\text{new}}{}'$ and let $Y_t = R_{\text{new}}a_{t,\text{new}}a_{t,*}{}'D_*{}'E_{\text{new}}{}' + E_{\text{new}}{}'D_* a_{t,*}a_{t,\text{new}}{}'R_{\text{new}}{}'$, then

$$A_k \succeq \frac{1}{\alpha}\sum_t Z_t + \frac{1}{\alpha}\sum_t Y_t \tag{18}$$

Consider $\sum_t Z_t = \sum_t R_{\text{new}}a_{t,\text{new}}a_{t,\text{new}}{}'R_{\text{new}}'$.

1) Using item 1 of Fact 10.1, the Z_t's are conditionally independent given $X_{j,k-1}$.
2) Using item 1, Ostrowoski's theorem (Theorem 2.9), and item 4, for all $X_{j,k-1} \in \Gamma_{j,k-1}$, $\lambda_{\min}\left(\mathbf{E}(\frac{1}{\alpha}\sum_t Z_t|X_{j,k-1})\right) = \lambda_{\min}\left(R_{\text{new}}\frac{1}{\alpha}\sum_t \mathbf{E}(a_{t,\text{new}}a_{t,\text{new}}{}')R_{\text{new}}{}'\right) \ge \lambda_{\min}\left(R_{\text{new}}R_{\text{new}}{}'\right)\lambda_{\min}\left(\frac{1}{\alpha}\sum_t \mathbf{E}(a_{t,\text{new}}a_{t,\text{new}}{}')\right) \ge (1 - (\zeta_{j,*}^+)^2)\lambda_{\text{new},k}^-$.
3) Finally, using items 2 and the bound on $\|a_t\|_\infty$ from the model, conditioned on $X_{j,k-1}$, $0 \preceq Z_t \preceq c\gamma_{\text{new},k}^2 I \preceq c\max\left((1.2)^{2k}\gamma_{\text{new}}^2, \gamma_*^2\right) I$ holds w.p. one for all $X_{j,k-1} \in \Gamma_{j,k-1}$.

Thus, applying Corollary 2.14 with $\epsilon = \frac{c\zeta\lambda^-}{24}$, we get

$$\mathbf{P}\left(\lambda_{\min}\left(\frac{1}{\alpha}\sum_t Z_t\right) \ge (1-(\zeta_*^+)^2)\lambda_{\text{new},k} - \frac{c\zeta\lambda^-}{24}\middle| X_{j,k-1}\right)$$
$$\ge 1 - c\exp\left(\frac{-\alpha\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot\min(1.2^{4k}\gamma_{\text{new}}^4, \gamma_*^4)}\right) \tag{19}$$

for all $X_{j,k-1} \in \Gamma_{j,k-1}$.

Consider $Y_t = R_{\text{new}}a_{t,\text{new}}a_{t,*}{}'D_*{}'E_{\text{new}} + E_{\text{new}}{}'D_* a_{t,*}a_{t,\text{new}}{}' R_{\text{new}}{}'$.

1) Using item 1, the Y_t's are conditionally independent given $X_{j,k-1}$.
2) Using item 1 and the fact that $a_{t,\text{new}}$ and $a_{t,*}$ are mutually uncorrelated, $\mathbf{E}\left(\frac{1}{\alpha}\sum_t Y_t|X_{j,k-1}\right) = 0$ for all $X_{j,k-1} \in \Gamma_{j,k-1}$.

3) Using the bound on $\|a_t\|_\infty$, items 2, 4, and Fact 6.8, conditioned on $X_{j,k-1}$, $\|Y_t\| \le 2\sqrt{cr}\zeta_*^+\gamma_*\gamma_{\text{new},k} \le 2\sqrt{cr}\zeta_*^+\gamma_*^2 \le 2$ holds w.p. one for all $X_{j,k-1} \in \Gamma_{j,k-1}$.

Thus, under the same conditioning, $-bI \preceq Y_t \preceq bI$ with $b = 2$ w.p. one.

Thus, applying Corollary 2.14 with $\epsilon = \frac{c\zeta\lambda^-}{24}$, we get

$$\mathbf{P}\left(\lambda_{\min}\left(\frac{1}{\alpha}\sum_t Y_t\right) \ge \frac{-c\zeta\lambda^-}{24}\Big|X_{j,k-1}\right) \ge 1 - c\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot(2b)^2}\right) \text{ for all } X_{j,k-1} \in \Gamma_{j,k-1} \quad (20)$$

Combining (18), (19) and (20) and using the union bound, $\mathbf{P}(\lambda_{\min}(A_k) \ge \lambda_{\text{new},k}^-(1-(\zeta_*^+)^2) - \frac{c\zeta\lambda^-}{12}|X_{j,k-1}) \ge 1 - p_a(\alpha,\zeta)$ for all $X_{j,k-1} \in \Gamma_{j,k-1}$. The first claim of the lemma follows by using $\lambda_{\text{new},k}^- \ge \lambda^-$ and then applying Lemma 2.11 with $X \equiv X_{j,k-1}$ and $\mathcal{C} \equiv \Gamma_{j,k-1}$.

Now consider $A_{k,\perp} := \frac{1}{\alpha}\sum_t E_{\text{new},\perp}{}'\Phi_0 L_t L_t{}'\Phi_0 E_{\text{new},\perp}$. Using item 2, $E_{\text{new},\perp}{}'\Phi_0 L_t = E_{\text{new},\perp}{}'D_* a_{t,*}$. Thus, $A_{k,\perp} = \frac{1}{\alpha}\sum_t Z_t$ with $Z_t = E_{\text{new},\perp}{}'D_* a_{t,*} a_{t,*}{}'D_*{}'E_{\text{new},\perp}$ which is of size $(n-c)\times(n-c)$. Using the same ideas as above we can show that $0 \preceq Z_t \preceq r(\zeta_*^+)^2\gamma_*^2 I \preceq \zeta I$ and $\mathbf{E}\left(\frac{1}{\alpha}\sum_t Z_t|X_{j,k-1}\right) \preceq (\zeta_*^+)^2\lambda^+ I$. Thus by Corollary 2.14 with $\epsilon = \frac{c\zeta\lambda^-}{24}$ and Lemma 2.11 the second claim follows.

Using the expression for $\mathcal{H}_k$ given in Definition 5.7, it is easy to see that

$$\begin{aligned}\|\mathcal{H}_k\|_2 &\le \max\{\|H_k\|_2, \|H_{k,\perp}\|_2\} + \|B_k\|_2 \\ &\le \left\|\frac{1}{\alpha}\sum_t e_t e_t{}'\right\|_2 + \max(\|T2\|_2, \|T4\|_2) + \|B_k\|_2\end{aligned} \quad (21)$$

where $T2 := \frac{1}{\alpha}\sum_t E_{\text{new}}{}'\Phi_0(L_t e_t{}' + e_t L_t{}')\Phi_0 E_{\text{new}}$ and $T4 := \frac{1}{\alpha}\sum_t E_{\text{new},\perp}{}'\Phi_0(L_t e_t{}' + e_t{}'L_t)\Phi_0 E_{\text{new},\perp}$. The second inequality follows by using the facts that (i) $H_k = T1 - T2$ where $T1 := \frac{1}{\alpha}\sum_t E_{\text{new}}{}'\Phi_0 e_t e_t{}'\Phi_0 E_{\text{new}}$, (ii) $H_{k,\perp} = T3 - T4$ where $T3 := \frac{1}{\alpha}\sum_t E_{\text{new},\perp}{}'\Phi_0 e_t e_t{}'\Phi_0 E_{\text{new},\perp}$, and (iii) $\max(\|T1\|_2, \|T3\|_2) \le \|\frac{1}{\alpha}\sum_t e_t e_t{}'\|_2$. Next, we obtain high probability bounds on each of the terms on the RHS of (21) using the Hoeffding corollaries.

Consider $\|\frac{1}{\alpha}\sum_t e_t e_t{}'\|_2$. Let $Z_t = e_t e_t{}'$.

1) Using item 1, conditioned on $X_{j,k-1}$, the various Z_t's in the summation are independent, for all $X_{j,k-1} \in \Gamma_{j,k-1}$.
2) Using item 4, and the bound on $\|a_t\|_\infty$, conditioned on $X_{j,k-1}$, $0 \preceq Z_t \preceq b_1 I$ w.p. one for all $X_{j,k-1} \in \Gamma_{j,k-1}$. Here $b_1 := (\kappa_s^+\zeta_{k-1}^+\phi^+\sqrt{c}\gamma_{\text{new},k} + \zeta_*^+\phi^+\sqrt{r}\gamma_*)^2$.
3) Also using item 4, $0 \preceq \frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X_{j,k-1}) \preceq b_2 I$, with $b_2 := (\kappa_s^+)^2(\zeta_{k-1}^+)^2(\phi^+)^2\lambda_{\text{new},k}^+ + (\zeta_*^+)^2(\phi^+)^2\lambda^+$ for all $X_{j,k-1} \in \Gamma_{j,k-1}$.

Thus, applying Corollary 2.14 with $\epsilon = \frac{c\zeta\lambda^-}{24}$,

$$\mathbf{P}\left(\left\|\frac{1}{\alpha}\sum_t e_t e_t{}'\right\|_2 \le b_2 + \frac{c\zeta\lambda^-}{24}\Big|X_{j,k-1}\right) \ge 1 - n\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2 b_1^2}\right) \text{ for all } X_{j,k-1} \in \Gamma_{j,k-1}$$

Consider $T2$. Let $Z_t := E_{\text{new}}{}'\Phi_0(L_t e_t{}' + e_t L_t{}')\Phi_0 E_{\text{new}}$ which is of size $c \times c$. Then $T2 = \frac{1}{\alpha}\sum_t Z_t$.

1) Using item 1, conditioned on $X_{j,k-1}$, the various Z_t's used in the summation are mutually independent, for all $X_{j,k-1} \in \Gamma_{j,k-1}$. Using item 2, $E_{\text{new}}{}'\Phi_0 L_t = R_{\text{new}} a_{t,\text{new}} + E_{\text{new}}{}'D_* a_{t,*}$ and $E_{\text{new}}{}'\Phi_0 e_t = (R_{\text{new}}{}')^{-1} D_{\text{new}}{}'e_t$.
2) Thus, using items 2, 4, and the bound on $\|a_t\|_\infty$, it follows that conditioned on $X_{j,k-1}$, $\|Z_t\|_2 \le 2\tilde{b}_3 \le 2b_3$ w.p. one for all $X_{j,k-1} \in \Gamma_{j,k-1}$. Here, $\tilde{b}_3 := \frac{\kappa_s^+}{\sqrt{1-(\zeta_*^+)^2}}\phi^+(\kappa_s^+\zeta_{k-1}^+\sqrt{c}\gamma_{\text{new},k} + \sqrt{r}\zeta_*^+\gamma_*)(\sqrt{c}\gamma_{\text{new},k} + \sqrt{r}\zeta_*^+\gamma_*)$ and $b_3 := \frac{1}{\sqrt{1-(\zeta_*^+)^2}}(\phi^+ c\kappa_s^{+2}\zeta_{k-1}^+\gamma_{\text{new},k}^2 + \phi^+\sqrt{rc}\kappa_s^{+2}\zeta_{k-1}^+\zeta_*^+\gamma_{\text{new},k}\gamma_* + \phi^+\sqrt{rc}\kappa_s^+\zeta_*^+\gamma_*\gamma_{\text{new},k} + \phi^+ r\zeta_*^{+2}\gamma_*^2)$.
3) Also, $\|\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X_{j,k-1})\|_2 \le 2\tilde{b}_4 \le 2b_4$ where $\tilde{b}_4 := \frac{\kappa_s^+}{\sqrt{1-(\zeta_*^+)^2}}\phi^+\kappa_s^+\zeta_{k-1}^+\lambda_{\text{new},k}^+ + \frac{\kappa_s^+}{\sqrt{1-(\zeta_*^+)^2}}\phi^+(\zeta_*^+)^2\lambda^+$ and $b_4 := \frac{\kappa_s^+}{\sqrt{1-(\zeta_*^+)^2}}\phi^+\kappa_s^+\zeta_{k-1}^+\lambda_{\text{new},k}^+ + \frac{1}{\sqrt{1-(\zeta_*^+)^2}}\phi^+(\zeta_*^+)^2\lambda^+$.

Thus, applying Corollary 2.15 with $\epsilon = \frac{c\zeta\lambda^-}{24}$,

$$\mathbf{P}\left(\|T2\|_2 \le 2b_4 + \frac{c\zeta\lambda^-}{24}\Big|X_{j,k-1}\right) \ge 1 - c\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2\cdot 4b_3^2}\right) \text{ for all } X_{j,k-1} \in \Gamma_{j,k-1}$$

Consider $T4$. Let $Z_t := E_{\text{new},\perp}{}'\Phi_0(L_t e_t{}' + e_t L_t{}')\Phi_0 E_{\text{new},\perp}$ which is of size $(n-c)\times(n-c)$. Then $T4 = \frac{1}{\alpha}\sum_t Z_t$.

1) Using item 1, conditioned on $X_{j,k-1}$, the various Z_t's used in the summation are mutually independent, for all $X_{j,k-1} \in \Gamma_{j,k-1}$. Using item 2, $E_{\text{new},\perp}{}'\Phi_0 L_t = E_{\text{new},\perp}{}'D_* a_{t,*}$.
2) Thus, conditioned on $X_{j,k-1}$, $\|Z_t\|_2 \le 2b_5$ w.p. one for all $X_{j,k-1} \in \Gamma_{j,k-1}$. Here $b_5 := \phi^+ r(\zeta_*^+)^2\gamma_*^2 + \phi^+\sqrt{rc}\kappa_s^+\zeta_*^+\zeta_{k-1}^+\gamma_*\gamma_{\text{new},k}$ This follows using items 4 and the bound on $\|a_t\|_\infty$.
3) Also, $\|\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X_{j,k-1})\|_2 \le 2b_6$, $b_6 := \phi^+(\zeta_*^+)^2\lambda^+$.

Applying Corollary 2.15 with $\epsilon = \frac{c\zeta\lambda^-}{24}$,

$$\mathbf{P}\left(\|T4\|_2 \le 2b_6 + \frac{c\zeta\lambda^-}{24}\Big|X_{j,k-1}\right) \ge 1 - (n-c)\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2\cdot 4b_5^2}\right) \text{ for all } X_{j,k-1} \in \Gamma_{j,k-1} \quad (22)$$

Consider $\max(\|T2\|_2, \|T4\|_2)$. Since $b_3 > b_5$ (follows because $\zeta_{k-1}^+ \le 1$) and $b_4 > b_6$, so $2b_6 + \frac{c\zeta\lambda^-}{24} < 2b_4 + \frac{c\zeta\lambda^-}{24}$ and $1 - (n-c)\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot 4b_5^2}\right) > 1 - (n-c)\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot 4b_3^2}\right)$. Therefore, for all $X_{j,k-1} \in \Gamma_{j,k-1}$, $\mathbf{P}\left(\|T4\|_2 \le 2b_4 + \frac{c\zeta\lambda^-}{24}\Big|X_{j,k-1}\right) \ge 1 - (n-c)\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2\cdot 4b_3^2}\right)$.

By the union bound, for all $X_{j,k-1} \in \Gamma_{j,k-1}$,

$$\mathbf{P}\left(\max(\|T2\|_2, \|T4\|_2) \le 2b_4 + \frac{c\zeta\lambda^-}{24}\Big|X_{j,k-1}\right) \ge 1 - n\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2\cdot 4b_3^2}\right) \quad (23)$$

Consider $\|B_k\|_2$. Let $Z_t := E_{\text{new},\perp}{}'\Phi_0(L_t - e_t)(L_t{}' - e_t{}')\Phi_0 E_{\text{new}}$ which is of size $(n-c)\times c$. Then $B_k = \frac{1}{\alpha}\sum_t Z_t$. Using item 2, $E_{\text{new},\perp}{}'\Phi_0(L_t - e_t) = E_{\text{new},\perp}{}'(D_* a_{t,*} - \Phi_0 e_t)$, $E_{\text{new}}{}'\Phi_0(L_t - e_t) = R_{\text{new}} a_{t,\text{new}} + E_{\text{new}}{}' D_* a_{t,*} + (R'_{\text{new}})^{-1} D'_{\text{new}} e_t$. Also, $\|Z_t\|_2 \le b_7$ w.p. one for all $X_{j,k-1} \in \Gamma_{j,k-1}$ and $\|\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X_{j,k-1})\|_2 \le b_8$ for all $X_{j,k-1} \in \Gamma_{j,k-1}$. Here

$$b_7 := (\sqrt{r}\zeta_*^+(1+\phi^+)\gamma_* + (\kappa_s^+)\zeta_{k-1}^+\phi^+\sqrt{c}\gamma_{\text{new},k}) \cdot \times\left(\sqrt{c}\gamma_{\text{new},k} + \sqrt{r}\zeta_*^+\left(1 + \frac{1}{\sqrt{1-(\zeta_*^+)^2}}\kappa_s^+\phi^+\right)\gamma_* + \frac{1}{\sqrt{1-(\zeta_*^+)^2}}\kappa_s^{+2}\zeta_{k-1}^+\phi^+\sqrt{c}\gamma_{\text{new},k}\right)$$

and

$$b_8 := \left(\kappa_s^+\zeta_{k-1}^+\phi^+ + \frac{1}{\sqrt{1-(\zeta_*^+)^2}}(\kappa_s^+)^3(\zeta_{k-1}^+)^2(\phi^+)^2\right)\lambda_{\text{new},k}^+ + (\zeta_*^+)^2\left(1 + \phi^+ + \frac{1}{\sqrt{1-(\zeta_*^+)^2}}\kappa_s^+\phi^+ + \frac{1}{\sqrt{1-(\zeta_*^+)^2}}\kappa_s^+(\phi^+)^2\right)\lambda^+$$

Thus, applying Corollary 2.15 with $\epsilon = \frac{c\zeta\lambda^-}{24}$,

$$\mathbf{P}\left(\|B_k\|_2 \le b_8 + \frac{c\zeta\lambda^-}{24}\Big|X_{j,k-1}\right) \ge 1 - n\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2 b_7^2}\right) \text{ for all } X_{j,k-1} \in \Gamma_{j,k-1} \quad (24)$$

Using (21), (22), (23) and (24) and the union bound, for any $X_{j,k-1} \in \Gamma_{j,k-1}$,

$$\mathbf{P}\left(\|\mathcal{H}_k\|_2 \le b_9 + \frac{c\zeta\lambda^-}{8}\Big|X_{j,k-1}\right) \ge 1 - n\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2 b_1^2}\right) - n\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2\cdot 4b_3^2}\right) - n\exp\left(\frac{-\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2 b_7^2}\right)$$

where

$$\begin{aligned} b_9 :&= b_2 + 2b_4 + b_8 \\ &= \Bigg(\Big(\frac{2(\kappa_s^+)^2\phi^+}{\sqrt{1-(\zeta_*^+)^2}} + \kappa_s^+\phi^+\Big)\zeta_{k-1}^+ + \Big((\kappa_s^+)^2(\phi^+)^2 + \frac{(\kappa_s^+)^3(\phi^+)^2}{\sqrt{1-(\zeta_*^+)^2}}\Big)(\zeta_{k-1}^+)^2\Bigg)\lambda_{\text{new},k}^+ \\ &\quad + \Bigg((\phi^+)^2 + \frac{2\phi^+}{\sqrt{1-(\zeta_*^+)^2}} + 1 + \phi^+ + \frac{\kappa_s^+\phi^+}{\sqrt{1-(\zeta_*^+)^2}} + \frac{\kappa_s^+(\phi^+)^2}{\sqrt{1-(\zeta_*^+)^2}}\Bigg)(\zeta_*^+)^2\lambda^+ \end{aligned}$$

Using $\lambda_{\text{new},k}^- \ge \lambda^-$ and $f := \lambda^+/\lambda^-$, $b_9 + \frac{c\zeta\lambda^-}{8} \le \lambda_{\text{new},k}^-(b + 0.125c\zeta)$ where b is defined in Definition 5.2. Using Fact 6.8 and substituting $\kappa_s^+ = 0.15$, $\phi^+ = 1.2$, one can upper bound b_1, b_3 and b_7 and show that the above probability is lower bounded by $1 - p_c(\alpha,\zeta)$. Finally, applying Lemma 2.11, the third claim of the lemma follows. ■

APPENDIX D
PROOF OF LEMMA 8.15

Proof of Lemma 8.15:

1) The first claim follows because $\|D_{\text{det},k}\|_2 = \|\Psi_{k-1}G_{\text{det},k}\|_2 = \|\Psi_{k-1}[G_1 G_2 \cdots G_{k-1}]\|_2 \le \sum_{k_1=1}^{k-1}\|\Psi_{k-1}G_{k_1}\|_2 \le \sum_{k_1=1}^{k-1}\|\Psi_{k_1}G_{k_1}\|_2 = \sum_{k_1=1}^{k-1}\tilde{\zeta}_{k_1} \le \sum_{k_1=1}^{k-1}\tilde{c}_{k_1}\zeta \le r\zeta$. The first inequality follows by triangle inequality. The second one follows because $\hat{G}_1, \cdots, \hat{G}_{k-1}$ are mutually orthonormal and so $\Psi_{k-1} = \prod_{k_2=1}^{k-1}(I - \hat{G}_{k_2}\hat{G}'_{k_2})$.
2) By the first claim, $\|(I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_{\text{det},k}\|_2 = \|\Psi_{k-1}G_{\text{det},k}\|_2 \le r\zeta$. By item 2) of Lemma 2.10 with $P = G_{\text{det},k}$ and $\hat{P} = \hat{G}_{\text{det},k}$, the result $\|G_{\text{det},k}G_{\text{det},k}{}' - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k}\|_2 \le 2r\zeta$ follows.
3) Recall that $D_k \overset{QR}{=} E_k R_k$ is a QR decomposition where E_k is orthonormal and R_k is upper triangular. Therefore, $\sigma_i(D_k) = \sigma_i(R_k)$. Since $\|(I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_{\text{det},k}\|_2 = \|\Psi_{k-1}G_{\text{det},k}\|_2 \le r\zeta$ and $G'_k G_{\text{det},k} = 0$, by item 4) of Lemma 2.10 with $P = G_{\text{det},k}$, $\hat{P} = \hat{G}_{\text{det},k}$ and $Q = G_k$, we have $\sqrt{1 - r^2\zeta^2} \le \sigma_i((I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_k) = \sigma_i(D_k) \le 1$.
4) Since $D_k \overset{QR}{=} E_k R_k$, so $\|D_{\text{undet},k}{}'E_k\|_2 = \|D_{\text{undet},k}{}'D_k R_k^{-1}\|_2 = \|G_{\text{undet},k}{}'\Psi'_{k-1}\Psi_{k-1}G_k R_k^{-1}\|_2 = \|G_{\text{undet},k}{}'\Psi_{k-1}G_k R_k^{-1}\|_2 = \|G_{\text{undet},k}{}'D_k R_k^{-1}\|_2 = \|G_{\text{undet},k}{}'E_k\|_2$. Since $E_k = D_k R_k^{-1} = (I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_k R_k^{-1}$,

$$\begin{aligned} \|G_{\text{undet},k}{}'E_k\|_2 &= \|G_{\text{undet},k}{}'(I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_k R_k^{-1}\|_2 \\ &\le \frac{\|G_{\text{undet},k}{}'(I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_k\|_2}{\sqrt{1 - r^2\zeta^2)}} \\ &= \frac{\|G_{\text{undet},k}{}'\hat{G}_{\text{det},k}\hat{G}'_{\text{det},k}G_k\|_2}{\sqrt{1 - r^2\zeta^2)}} \end{aligned}$$

By item 3) of Lemma 2.10 with $P = G_{\text{det},k}$, $\hat{P} = \hat{G}_{\text{det},k}$ and $Q = G_{\text{undet},k}$, we get $\|G_{\text{undet},k}{}'\hat{G}_{\text{det},k}\|_2 \le r\zeta$. By item 3) of Lemma 2.10 with $\hat{P} = \hat{G}_{\text{det},k}$ and $Q = G_k$, we get $\|\hat{G}'_{\text{det},k} G_k\|_2 \le r\zeta$. Therefore, $\|G_{\text{undet},k}{}' E_k\|_2 = \|E_k{}' G_{\text{undet},k}\|_2 \le \frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}$. ■

APPENDIX E
PROOF OF LEMMA 8.19

Proof: We use $\frac{1}{\tilde{\alpha}}\sum_t$ to denote $\frac{1}{\tilde{\alpha}}\sum_{t\in\tilde{\mathcal{I}}_{j,k}}$.

For $t \in \tilde{\mathcal{I}}_{j,k}$, let $a_{t,k} := G_{j,k}{}'L_t$, $a_{t,\text{det}} := G_{\text{det},k}{}'L_t = [G_{j,1},\cdots G_{j,k-1}]'L_t$ and $a_{t,\text{undet}} := G_{\text{undet},k}{}'L_t = [G_{j,k+1}\cdots G_{j,\vartheta_j}]'L_t$. Then $a_t := P_j' L_t$ can be split as $a_t = [a'_{t,\text{det}}\ a'_{t,k}\ a'_{t,\text{undet}}]'$.

This lemma follows using the following facts and the Hoeffding corollaries, Corollary 2.14 and 2.15.

1) The matrices D_k, R_k, E_k, $D_{\text{det},k}$, $D_{\text{undet},k}$, Ψ_{k-1}, Φ_K are functions of the r.v. $\tilde{X}_{j,k-1}$. All terms that we bound for the first two claims of the lemma are of the form $\frac{1}{\alpha}\sum_{t\in\tilde{\mathcal{I}}_{j,k}} Z_t$ where $Z_t = f_1(\tilde{X}_{j,k-1})Y_t f_2(\tilde{X}_{j,k-1})$, Y_t is a sub-matrix of $a_t a_t'$ and $f_1(.)$ and $f_2(.)$ are functions of $\tilde{X}_{j,k-1}$. For instance, one of the terms while bounding $\lambda_{\min}(\mathcal{A}_k)$ is $\frac{1}{\tilde{\alpha}}\sum_t R_k a_{t,k} a_{t,k}{}' R_k{}'$. $\tilde{X}_{j,k-1}$ is independent of any a_t for $t \in \tilde{\mathcal{I}}_{j,k}$, and hence the same is true for the matrices D_k, R_k, E_k, $D_{\text{det},k}$, $D_{\text{undet},k}$, Ψ_{k-1}, Φ_K. Also, a_t's for different $t \in \tilde{\mathcal{I}}_{j,k}$ are mutually independent. Thus, conditioned on $\tilde{X}_{j,k-1}$, the Z_t's defined above are mutually independent.
2) All the terms that we bound for the third claim contain e_t. Using Lemma 6.4, conditioned on $\tilde{X}_{j,k-1}$, e_t satisfies (10) w.p. one whenever $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. Conditioned on $\tilde{X}_{j,k-1}$, all these terms are also of the form $\frac{1}{\alpha}\sum_{t\in\tilde{\mathcal{I}}_{j,k}} Z_t$ with Z_t as defined above, whenever $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. Thus, conditioned on $\tilde{X}_{j,k-1}$, the Z_t's for these terms are mutually independent, whenever $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$.
3) By Remark 8.14 and the definition of $\tilde{\Gamma}_{j,k-1}$, $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$ implies that $\zeta_* \le r\zeta$, $\tilde{\zeta}_{k'} \le c_{k'}\zeta$, for all $k' = 1, 2, \ldots k-1$, $\zeta_K \le \zeta_K^+ \le c\zeta$, (iv) $\phi_K \le \phi^+$ (by Lemma 6.4); (v) $\|\Phi_K P_j\|_2 \le (r+c)\zeta$; and (vi) all conclusions of Lemma 8.15 hold.
4) By the clustering assumption, $\lambda_k^- \le \lambda_{\min}(\mathbf{E}(a_{t,k}a_{t,k}{}')) \le \lambda_{\max}(\mathbf{E}(a_{t,k}a_{t,k}{}')) \le \lambda_k^+$; $\lambda_{\max}(\mathbf{E}(a_{t,\text{det}}a_{t,\text{det}}{}')) \le \lambda_1^+ = \lambda^+$; and $\lambda_{\max}(\mathbf{E}(a_{t,\text{undet}}a_{t,\text{undet}}{}')) \le \lambda_{k+1}^+$. Also, $\lambda_{\max}(\mathbf{E}(a_t a_t')) \le \lambda^+$.
5) By Weyl's theorem, for a sequence of matrices B_t, $\lambda_{\min}(\sum_t B_t) \ge \sum_t \lambda_{\min}(B_t)$ and $\lambda_{\max}(\sum_t B_t) \le \sum_t \lambda_{\max}(B_t)$.

Consider $\tilde{A}_k = \frac{1}{\tilde{\alpha}}\sum_t E_k{}'\Psi_{k-1}L_t L_t{}'\Psi_{k-1}E_k$. Notice that $E_k{}'\Psi_{k-1}L_t = R_k a_{t,k} + E_k{}'(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}})$. Let $Z_t = R_k a_{t,k}a_{t,k}{}'R_k{}'$ and let $Y_t = R_k a_{t,k}(a_{t,\text{det}}{}'D_{\text{det},k}{}' + a_{t,\text{undet}}{}'D_{\text{undet},k}{}')E_k + E_k'(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}})a_{t,k}{}'R_k{}'$. Then

$$\tilde{A}_k \succeq \frac{1}{\tilde{\alpha}}\sum_t Z_t + \frac{1}{\tilde{\alpha}}\sum_t Y_t \tag{25}$$

Consider $\frac{1}{\tilde{\alpha}}\sum_t Z_t = \frac{1}{\tilde{\alpha}}\sum_t R_k a_{t,k}a_{t,k}{}'R_k{}'$. (a) As explained above, the Z_t's are conditionally independent given $\tilde{X}_{j,k-1}$. (b) Using Ostrowoski's theorem and Lemma 8.15, for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$, $\lambda_{\min}(\mathbf{E}(\frac{1}{\tilde{\alpha}}\sum_t Z_t|\tilde{X}_{j,k-1})) = \lambda_{\min}(R_k\frac{1}{\tilde{\alpha}}\sum_t \mathbf{E}(a_{t,k}a_{t,k}{}')R_k{}') \ge \lambda_{\min}(R_k R_k{}')\lambda_{\min}(\frac{1}{\tilde{\alpha}}\sum_t \mathbf{E}(a_{t,k}a_{t,k}{}')) \ge (1 - r^2\zeta^2)\lambda_k^-$. (c) Finally, using $\|R_k\|_2 \le 1$ and $\|a_{t,k}\|_2 \le \sqrt{\tilde{c}_k}\gamma_*$, conditioned on $\tilde{X}_{j,k-1}$, $0 \preceq Z_t \preceq \tilde{c}_k\gamma_*^2 I$ holds w.p. one for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$.

Thus, applying Corollary 2.14 with $\epsilon = 0.1\zeta\lambda^-$, and using $\tilde{c}_k \le r$, for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$,

$$\begin{aligned}&\mathbf{P}\left(\lambda_{\min}\Big(\frac{1}{\tilde{\alpha}}\sum_t Z_t\Big) \ge (1 - r^2\zeta^2)\lambda_k^- - 0.1\zeta\lambda^- \Big| \tilde{X}_{j,k-1}\right)\\ &\ge 1 - \tilde{c}_k \exp\left(\frac{-\tilde{\alpha}\epsilon^2}{8(\tilde{c}_k\gamma_*^2)^2}\right) \ge 1 - r\exp\left(\frac{-\tilde{\alpha}\cdot(0.1\zeta\lambda^-)^2}{8r^2\gamma_*^4}\right)\end{aligned} \tag{26}$$

Consider $Y_t = R_k a_{t,k}(a_{t,\text{det}}{}'D_{\text{det},k}{}' + a_{t,\text{undet}}{}'D_{\text{undet},k}{}')E_k + E_k'(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}})a_{t,k}{}'R_k{}'$. (a) As before, the Y_t's are conditionally independent given $\tilde{X}_{j,k-1}$. (b) Since $\mathbf{E}[a_t] = 0$ and $\text{Cov}[a_t] = \Lambda_t$ is diagonal, $\mathbf{E}(\frac{1}{\tilde{\alpha}}\sum_t Y_t|\tilde{X}_{j,k-1}) = 0$ whenever $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. (c) Conditioned on $\tilde{X}_{j,k-1}$, $\|Y_t\|_2 \le 2\sqrt{\tilde{c}_k r}\gamma_*^2 r\zeta(1 + \frac{r\zeta}{\sqrt{1-r^2\zeta^2}}) \le 2r^2\zeta\gamma_*^2(1 + \frac{10^{-4}}{\sqrt{1-10^{-4}}}) \le \frac{2}{r}(1 + \frac{10^{-4}}{\sqrt{1-10^{-4}}}) < 2.1$ holds w.p. one for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. This follows because $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$ implies that $\|D_{\text{det},k}\|_2 \le r\zeta$, $\|E_k{}'D_{\text{undet},k}\|_2 = \|E_k{}'G_{\text{undet},k}\|_2 \le \frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}$. Thus, under the same conditioning, $-bI \preceq Y_t \preceq bI$ with $b = 2.1$ w.p. one. Thus, applying Corollary 2.14 with $\epsilon = 0.1\zeta\lambda^-$, we get

$$\begin{aligned}&\mathbf{P}\left(\lambda_{\min}\Big(\frac{1}{\tilde{\alpha}}\sum_t Y_t\Big) \ge -0.1\zeta\lambda^- \Big| \tilde{X}_{j,k-1}\right)\\ &\ge 1 - r\exp\left(\frac{-\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8(4.2)^2}\right) \text{ for all } \tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}\end{aligned} \tag{27}$$

Combining (25), (26) and (27) and using the union bound, $\mathbf{P}(\lambda_{\min}(\tilde{A}_k) \ge \lambda_k^-(1 - r^2\zeta^2) - 0.2\zeta\lambda^-|\tilde{X}_{j,k-1}) \ge 1 - \tilde{p}_1(\tilde{\alpha},\zeta)$ for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$ where

$$\tilde{p}_1(\tilde{\alpha},\zeta) := r\exp\left(\frac{-\tilde{\alpha}\cdot(0.1\zeta\lambda^-)^2}{8r^2\gamma_*^4}\right) + r\exp\left(\frac{-\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8(4.2)^2}\right) \tag{28}$$

The first claim of the lemma follows by using $\lambda_k^- \ge \lambda^-$ and applying Lemma 2.11 with $X \equiv \tilde{X}_{j,k-1}$ and $\mathcal{C} \equiv \tilde{\Gamma}_{j,k-1}$.

Consider $\tilde{A}_{k,\perp} := \frac{1}{\alpha}\sum_t E_{k,\perp}{}'\Psi_{k-1}L_t L_t{}'\Psi_{k-1}E_{k,\perp}$. Notice that $E_{k,\perp}{}'\Psi_{k-1}L_t = E_{k,\perp}{}'(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}})$. Thus, $\tilde{A}_{k,\perp} = \frac{1}{\tilde{\alpha}}\sum_t Z_t$ with $Z_t = E_{k,\perp}{}'(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}})(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}})'E_{k,\perp}$ which is of size $(n - \tilde{c}_k) \times (n - \tilde{c}_k)$. (a) As before, given $\tilde{X}_{j,k-1}$, the Z_t's are independent. (b) Conditioned on $\tilde{X}_{j,k-1}$, $0 \preceq Z_t \preceq r\gamma_*^2 I$ w.p. one for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. (c) $\mathbf{E}(\frac{1}{\alpha}\sum_t Z_t|\tilde{X}_{j,k-1}) \preceq (\lambda_{k+1}^+ + r^2\zeta^2\lambda^+)I$ for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$.

Thus applying Corollary 2.14 with $\epsilon = 0.1\zeta\lambda^-$ and using $\tilde{c}_k \geq \tilde{c}_{\min}$, we get

$$\mathbf{P}(\lambda_{\max}(\tilde{A}_{k,\perp}) \leq \lambda_{k+1}^+ + r^2\zeta^2\lambda^+ + 0.1\zeta\lambda^- | \tilde{X}_{j,k-1}) \geq 1 - \tilde{p}_2(\tilde{\alpha},\zeta) \text{ for all } \tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1} \quad (29)$$

where

$$\tilde{p}_2(\tilde{\alpha},\zeta) := (n - \tilde{c}_{\min}) \exp\left(\frac{-\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8r^2\gamma_*^4}\right) \quad (30)$$

The second claim follows using $\lambda_k^- \geq \lambda^-$, $f := \lambda^+/\lambda^-$, $\tilde{h}_k := \lambda_{k+1}{}^+/\lambda_k{}^-$ in the above expression and applying Lemma 2.11.

Consider the third claim. Using the expression for $\tilde{\mathcal{H}}_k$ given in Definition 8.3, it is easy to see that

$$\|\tilde{\mathcal{H}}_k\|_2 \leq \max\{\|\tilde{H}_k\|_2, \|\tilde{H}_{k,\perp}\|_2\} + \|\tilde{B}_k\|_2 \leq \left\|\frac{1}{\tilde{\alpha}}\sum_t e_t e_t{}'\right\|_2 + \max(\|T2\|_2, \|T4\|_2) + \|\tilde{B}_k\|_2 \quad (31)$$

where $T2 := \frac{1}{\tilde{\alpha}}\sum_t E_k{}'\Psi_{k-1}(L_t e_t{}' + e_t L_t{}')\Psi_{k-1}E_k$ and $T4 := \frac{1}{\tilde{\alpha}}\sum_t E_{k,\perp}{}'\Psi_{k-1}(L_t e_t{}' + e_t{}' L_t)\Psi_{k-1}E_{k,\perp}$. The second inequality follows by using the facts that (i) $\tilde{H}_k = T1 - T2$ where $T1 := \frac{1}{\tilde{\alpha}}\sum_t E_k{}'\Psi_{k-1}e_t e_t{}'\Psi_{k-1}E_k$, (ii) $\tilde{H}_{k,\perp} = T3 - T4$ where $T3 := \frac{1}{\tilde{\alpha}}\sum_t E_{k,\perp}{}'\Psi_{k-1}e_t e_t{}'\Psi_{k-1}E_{k,\perp}$, and (iii) $\max(\|T1\|_2, \|T3\|_2) \leq \|\frac{1}{\tilde{\alpha}}\sum_t e_t e_t{}'\|_2$.

Next, we obtain high probability bounds on each of the terms on the RHS of (21) using the Hoeffding corollaries.

Consider $\|\frac{1}{\tilde{\alpha}}\sum_t e_t e_t{}'\|_2$. Let $Z_t = e_t e_t{}'$. (a) As explained in the beginning of the proof, conditioned on $\tilde{X}_{j,k-1}$, the various Z_t's in the summation are independent whenever $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. Also, by Lemma 6.4, under this conditioning, $\hat{T}_t = T_t$ for all $t \in \tilde{I}_{j,k}$ and hence e_t satisfies (10) in this interval. Recall also that in this interval, $\Phi_{(t)} = \Phi_K$. Thus, using $\|\Phi_K P_j\|_2 \leq (r+c)\zeta$,

$$\|e_t\|_2 \leq \phi^+\sqrt{\zeta}$$

(b) Conditioned on $\tilde{X}_{j,k-1}$, $0 \preceq Z_t \preceq b_1 I$ w.p. one for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. Here $b_1 := \phi^{+2}\zeta$. (c) Using $\|\Phi_K P_j\|_2 \leq (r+c)\zeta$, $0 \preceq \frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|\tilde{X}_{j,k-1}) \preceq b_2 I$, $b_2 := (r+c)^2\zeta^2\phi^{+2}\lambda^+$ for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$.

Thus, applying Corollary 2.14 with $\epsilon = 0.1\zeta\lambda^-$,

$$\mathbf{P}\left(\left\|\frac{1}{\tilde{\alpha}}\sum_t e_t e_t{}'\right\|_2 \leq b_2 + 0.1\zeta\lambda^- \middle| \tilde{X}_{j,k-1}\right) \geq 1 - n\exp\left(\frac{-\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8\cdot b_1^2}\right) \text{ for all } \tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1} \quad (32)$$

Consider $T2$. Let $Z_t := E_k{}'\Psi_{k-1}(L_t e_t{}' + e_t L_t{}')\Psi_{k-1}E_k$ which is of size $\tilde{c}_k \times \tilde{c}_k$. Then $T2 = \frac{1}{\tilde{\alpha}}\sum_t Z_t$. (a) Conditioned on $\tilde{X}_{j,k-1}$, the various Z_t's used in the summation are mutually independent whenever $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. (b) Notice that $E_k{}'\Psi_{k-1}L_t = R_k a_{t,k} + E_k{}'(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}})$ and $E_k{}'\Psi_{k-1}e_t = (R_k^{-1})'D_k' e_t = (R_k^{-1})'D_k' I_{T_t}[(\Phi_K)_{T_t}'(\Phi_K)_{T_t}]^{-1}I_{T_t}{}'\Phi_K P_j a_t$. Thus conditioned on $\tilde{X}_{j,k-1}$, $\|Z_t\|_2 \leq 2b_3$ w.p. one for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. Here, $b_3 := \frac{\sqrt{r\zeta}}{\sqrt{1-r^2\zeta^2}}\phi^+\gamma_*$. This follows using $\|(R_k^{-1})'\|_2 \leq 1/\sqrt{1-r^2\zeta^2}$, $\|e_t\|_2 \leq \phi^+\sqrt{\zeta}$ and $\|E_k'\Psi_{k-1}L_t\|_2 \leq \|L_t\|_2 \leq \sqrt{r}\gamma_*$. (c) Also, $\|\frac{1}{\tilde{\alpha}}\sum_t \mathbf{E}(Z_t|\tilde{X}_{j,k-1})\|_2 \leq 2b_4$ where $b_4 := \kappa_{s,D}^+\kappa_{s,e}^+(r+c)\zeta\phi^+(\lambda_k^+ + r\zeta\lambda^+ + \frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}\lambda_{k+1}^+)$. Here $\kappa_{s,D}^+ = \kappa_{s,*}^+ + r\zeta$ defined in Remark 8.10 is the bound on $\max_j \max_k \kappa_s(D_{j,k})$.

Thus, applying Corollary 2.15 with $\epsilon = 0.1\zeta\lambda^-$, for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$,

$$\mathbf{P}(\|T2\|_2 \leq 2b_4 + 0.1\zeta\lambda^-|\tilde{X}_{j,k-1}) \geq 1 - \tilde{c}_k\exp\left(\frac{-\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32\cdot 4b_3^2}\right)$$

Consider $T4$. Let $Z_t := E_{k,\perp}{}'\Psi_{k-1}(L_t e_t{}' + e_t L_t{}')\Psi_{k-1}E_{k,\perp}$ which is of size $(n-\tilde{c}_k)\times(n-\tilde{c}_k)$. Then $T4 = \frac{1}{\tilde{\alpha}}\sum_t Z_t$. (a) conditioned on $\tilde{X}_{j,k-1}$, the various Z_t's used in the summation are mutually independent whenever $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. (b) Notice that $E_{k,\perp}{}'\Psi_{k-1}L_t = E_{k,\perp}{}'(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}})$. Thus, conditioned on $\tilde{X}_{j,k-1}$, $\|Z_t\|_2 \leq 2b_5$ w.p. one for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. Here $b_5 := \sqrt{r\zeta}\phi^+\gamma_*$. (c) Also, for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$, $\|\frac{1}{\tilde{\alpha}}\sum_t \mathbf{E}(Z_t|\tilde{X}_{j,k-1})\|_2 \leq 2b_6$, $b_6 := \kappa_{s,e}^+(r+c)\zeta\phi^+(\lambda_{k+1}^+ + r\zeta\lambda^+)$. Applying Corollary 2.15 with $\epsilon = 0.1\zeta\lambda^-$, for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$,

$$\mathbf{P}(\|T4\|_2 \leq 2b_6 + 0.1\zeta\lambda^-|\tilde{X}_{j,k-1}) \geq 1 - (n-\tilde{c}_k)\exp\left(\frac{-\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32\cdot 4b_5^2}\right) \geq 1 - (n-\tilde{c}_{\min})\exp\left(\frac{-\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32\cdot 4b_5^2}\right).$$

Consider $\max(\|T2\|_2, \|T4\|_2)$. By union bound and using $b_3 > b_5$, for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$,

$$\mathbf{P}(\max(\|T2\|_2, \|T4\|_2) \leq 2\max(b_4, b_6) + 0.1\zeta\lambda^-|\tilde{X}_{j,k-1}) \geq 1 - n\exp\left(\frac{-\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32\cdot 4b_3^2}\right) \quad (33)$$

Consider $\|\tilde{B}_k\|_2$. Let $Z_t := E_{k,\perp}{}'\Psi_{k-1}(L_t - e_t)(L_t{}' - e_t{}')\Psi_{k-1}E_k$ which is of size $(n-\tilde{c}_k)\times\tilde{c}_k$. Then $\tilde{B}_k = \frac{1}{\tilde{\alpha}}\sum_t Z_t$. (a) conditioned on $\tilde{X}_{j,k-1}$, the various Z_t's used in the summation are mutually independent whenever $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$. (b) Notice that $E_{k,\perp}{}'\Psi_{k-1}(L_t - e_t) = E_{k,\perp}{}'(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}} - \Psi_{k-1}e_t)$ and $E_k{}'\Psi_{k-1}(L_t - e_t) = R_k a_{t,k} + E_k{}'(D_{\text{det},k}a_{t,\text{det}} + D_{\text{undet},k}a_{t,\text{undet}} - \Psi_{k-1}e_t)$. Thus, conditioned on $\tilde{X}_{j,k-1}$, $\|Z_t\|_2 \leq b_7$ w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Here $b_7 := (\sqrt{r}\gamma_* + \phi^+\sqrt{\zeta})^2$. (c) $\|\frac{1}{\tilde{\alpha}}\sum_t \mathbf{E}(Z_t|\tilde{X}_{j,k-1})\|_2 \leq b_8$ for all $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$ where

$$b_8 := (r+c)\zeta\kappa_{s,e}^+\phi^+\lambda_k^+ + \left[(r+c)\zeta\kappa_{s,e}^+\phi^+ + (r+c)\zeta\kappa_{s,e}^+\frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}\right]\lambda_{k+1}^+ + [r^2\zeta^2 + 2(r+c)r\zeta^2\kappa_{s,e}^+\phi^+ + (r+c)^2\zeta^2\kappa_{s,e}^{+2}\phi^{+2}]\lambda^+$$

Thus, applying Corollary 2.15 with $\epsilon = 0.1\zeta\lambda^-$,

$$\mathbf{P}(\|\tilde{B}_k\|_2 \leq b_8 + 0.1\zeta\lambda^- | \tilde{X}_{j,k-1}) \geq 1 - n \exp \times \left(\frac{-\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot b_7^2} \right) \text{ for all } \tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1} \quad (34)$$

Using (31), (32), (33) and (34) and the union bound, for any $\tilde{X}_{j,k-1} \in \tilde{\Gamma}_{j,k-1}$,

$$\mathbf{P}(\|\tilde{\mathcal{H}}_k\|_2 \leq b_9 + 0.2\zeta\lambda^- | \tilde{X}_{j,k-1}) \geq 1 - \tilde{p}_3(\tilde{\alpha}, \zeta)$$

where $b_9 := b_2 + 2b_4 + b_8$ and

$$\tilde{p}_3(\tilde{\alpha}, \zeta) : = n \exp\left(\frac{-\tilde{\alpha}\epsilon^2}{8 \cdot b_1^2} \right) + n \exp\left(\frac{-\tilde{\alpha}\epsilon^2}{32 \cdot 4b_3^2} \right) + n \exp\left(\frac{-\tilde{\alpha}\epsilon^2}{32 \cdot b_7^2} \right) \quad (35)$$

with $b_1 = {\phi^+}^2\zeta$, $b_3 := \sqrt{r\zeta}\phi^+\gamma_*$, $b_7 := (\sqrt{r}\gamma_* + \phi^+\sqrt{\zeta})^2$. Using $\lambda_k^- \geq \lambda^-$, $f := \lambda^+/\lambda^-$, $\tilde{g}_k := \lambda_k^+/\lambda_k^-$ and $\tilde{h}_k := \lambda_{k+1}^+/\lambda_k^-$, and then applying Lemma 2.11, the third claim of the lemma follows. ∎

REFERENCES

[1] C. Qiu and N. Vaswani, "Real-time robust principal components' pursuit," in *Proc. Allerton Conf. Commun., Control, Comput.*, 2010.

[2] C. Qiu, N. Vaswani, and L. Hogben, "Recursive robust PCA or recursive sparse recovery in large but structured noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 5954–5958.

[3] C. Qiu and N. Vaswani, "Recursive sparse recovery in large but structured noise—Part 2," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2013, pp. 864–868.

[4] S. Roweis, "EM algorithms for PCA and SPCA," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 1998, pp. 626–632.

[5] F. De la Torre and M. J. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, nos. 1–3, pp. 117–142, 2003.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, 2011.

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

[8] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2002, pp. 707–720.

[9] D. Skocaj and A. Leonardis, "Weighted and robust incremental method for subspace learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2. Oct. 2003, pp. 1494–1501.

[10] Y. Li, L.-Q. Xu, J. Morphett, and R. Jacobs, "An integrated algorithm of incremental and robust PCA," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2003, pp. 245–248.

[11] J. Wright and Y. Ma, "Dense error correction via l^1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3540–3560, Jul. 2010.

[12] T. Zhang and G. Lerman. (2014, Jan.). A novel M-estimator for robust PCA. *J. Mach. Learn. Res.* [Online]. *15(1)*, pp. 749–808. Available: http://dl.acm.org/citation.cfm?id=2627435.2627458

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

[14] M. McCoy and J. A. Tropp, "Two proposals for robust PCA using semidefinite programming," *Electron. J. Statist.*, vol. 5, pp. 1123–1160, Sep. 2011.

[15] M. B. McCoy and J. A. Tropp, "Sharp recovery bounds for convex deconvolution, with applications," arXiv:1205.1580.

[16] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[17] Y. Hu, S. Goud, and M. Jacob, "A fast majorize–minimize algorithm for the recovery of sparse and low-rank matrices," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 742–753, Feb. 2012.

[18] A. E. Waters, A. C. Sankaranarayanan, and R. G. Baraniuk, "SpaRCS: Recovering low-rank and sparse matrices from compressive measurements," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2011.

[19] E. Richard, P.-A. Savalle, and N. Vayatis, "Estimation of simultaneously sparse and low rank matrices," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, 2012.

[20] D. Hsu, S. M. Kakade, and T. Zhang, "Robust matrix decomposition with outliers," arXiv:1011.1518.

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

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

[23] A. Ganesh, K. Min, J. Wright, and Y. Ma, "Principal component pursuit with reduced linear measurements," arXiv:1202.6445.

[24] M. Tao and X. Yuan, "Recovering low-rank and sparse components of matrices from incomplete and noisy observations," *SIAM J. Optim.*, vol. 21, no. 1, pp. 57–81, 2011.

[25] C. Qiu and N. Vaswani, "Recursive sparse recovery in large but correlated noise," in *Proc. 49th Allerton Conf. Commun. Control Comput.*, 2011.

[26] H. Guo, C. Qiu, and N. Vaswani, "An online algorithm for separating sparse and low-dimensional signal sequences from their sum," *IEEE Trans. Signal Process.*, 2014, arXiv: 1303.4261 [cs.IT], submitted for publication.

[27] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1568–1575.

[28] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[29] K. Lee and Y. Bresler, "ADMiRA: Atomic decomposition for minimum rank approximation," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4402–4416, Sep. 2010.

[30] E. Candes, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus l'Academie Sci.*, vol. 346, no. 9, pp. 589–592, 2008.

[31] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM J. Numer. Anal.*, vol. 7, no. 1, pp. 1–46, Mar. 1970.

[32] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012.

[33] B. Nadler, "Finite sample approximation results for principal component analysis: A matrix perturbation approach," *Ann. Statist.*, vol. 36, no. 6, pp. 2791–2817, 2008.

[34] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[35] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2010, pp. 3830–3833.

[36] K. Mitra, A. Veeraraghavan, and R. Chellappa, "Robust regression using sparse learning for high dimensional parameter estimation problems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2010, pp. 3846–3849.

[37] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[38] C. Qiu and N. Vaswani, "Support-predicted modified-CS for recursive robust principal components' pursuit," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul./Aug. 2011, pp. 668–672.

[39] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. New York, NY, USA: Oxford Univ. Press, 2001.

[40] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[41] G. Li and Z. Chen, "Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo," *J. Amer. Statist. Assoc.*, vol. 80, no. 391, pp. 759–766, 1985.

[42] P. Feng and Y. Bresler, "Spectrum-blind minimum-rate sampling and reconstruction of multiband signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 3. May 1996, pp. 1688–1691.

[43] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[44] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[45] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[46] E. Candes and T. Tao, “The Dantzig selector: Statistical estimation when p is much larger than n,” *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[47] D. Hsu, S. M. Kakade, and T. Zhang, “Robust matrix decomposition with sparse corruptions,” *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7221–7234, Nov. 2011.

[48] J. Zhan and N. Vaswani, “Performance guarantees for ReProCS—Correlated low-rank matrix entries case,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, May 2014.

[49] B. Lois, N. Vaswani, and C. Qiu, “Performance guarantees for undersampled recursive sparse recovery in large but structured noise,” in *Proc. GlobalSIP*, 2013.

Chenlu Qiu received a B.S. from Southeast University in China in 2006 in Information Engineering and a Ph.D. from Iowa State University in 2013 in Electrical Engineering. She is currently with the Traffic Management Research Institute of the Ministry of Public Security, China. Her research interests include robust PCA and video analysis.

Namrata Vaswani received a B.Tech. from Indian Institute of Technology (IIT), Delhi, in 1999 and a Ph.D. from University of Maryland, College Park, in 2004, both in Electrical Engineering. During 2004-05, she was a research scientist at Georgia Tech. Since Fall 2005, she has been with the Iowa State University where she is currently an associate professor of electrical and computer engineering. She held the Harpole-Pentair assistant professorship at ISU during 2008-09. From 2009 to 2013, she served as an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING. Her research interests are in signal and information processing for problems motivated by big-data and bio-imaging applications. Her current work focuses on recursive sparse recovery, robust PCA, matrix completion and applications in video and medical imaging.

Brian Lois received a B.S. from Marquette University, Milwaukee, WI in Mathematics in 2010. He is currently a Ph.D. candidate in Mathematics and Electrical Engineering at Iowa State University. His research interests are in linear algebra and signal processing.

Leslie Hogben is Dio Lewis Holl Chair in Applied Mathematics and Professor of Mathematics at Iowa State University, and Associate Director for Diversity of the American Institute of Mathematics. She is the author of more than 70 research papers and advisor to numerous PhD students, the Secretary/Treasurer of the International Linear Algebra Society, and an associate editor of the journals *Linear Algebra and its Applications* and *Electronic Journal of Linear Algebra*. Her research is in linear algebra, especially combinatorial matrix theory, spectral graph theory, and applications of linear algebra and graph theory.