2013

Recursive robust PCA or recursive sparse recovery in large but structured noise

Chenlu Qiu
Iowa State University

Recursive robust PCA or recursive sparse recovery in large but structured noise

by

Chenlu Qiu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Electrical Engineering

Program of Study Committee:
Namrata Vaswani, Major Professor
Zhengdao Wang
Aleksandar Dogandzic
Nicola Elia
Leslie Hogben
Dan Nordman

Iowa State University

Ames, Iowa

2013

DEDICATION

I would like to dedicate this dissertation to my family without whose support I would not have been able to complete this work.

TABLE OF CONTENTS

LIST OF TABLES vi

LIST OF FIGURES vii

ABSTRACT viii

CHAPTER 1. Introduction 1

1.1 Notation 5

1.2 Dissertation Organization 7

CHAPTER 2. Mathematical Preliminaries 8

2.1 Compressive Sensing result 8

2.2 Results from linear algebra 9

2.3 Simple probability facts and matrix Hoeffding inequalities 10

CHAPTER 3. Problem Definition and Model Assumptions 13

3.1 Problem Definition 13

3.2 Slow Subspace Change 15

3.3 Denseness assumption and its relation with RIC 15

3.4 Model Verification 17

CHAPTER 4. Recursive Projected CS (ReProCS) and its Performance Guarantees 20

4.1 The Projection-PCA algorithm 20

4.2 The Recursive Projected CS (ReProCS) Algorithm 21

4.3 Performance Guarantees 24

4.4 Proof of Theorem 4.3.1 28

4.4.1 Key Lemmas – 1: Bounding the RIC, sparse recovery and LS error and subspace estimation error 33

4.4.2 Key Lemmas – 2: Showing high probability exponential decay of the subspace error 34

4.4.3 Proof Outline for Theorem 4.3.1 38

4.4.4 Proof of Theorem 4.3.1 40

4.5 ReProCS with practical parameters setting 41

4.6 Experimental Results 42

CHAPTER 5. ReProCS with cluster-PCA (ReProCS-cPCA) an its performance Guarantee 50

5.1 Clustering assumption 50

5.2 The ReProCS-cPCA algorithm 53

5.3 Performance Guarantees 55

5.4 Proof of Theorem 5.3.1 60

5.4.1 Proof Outline of Theorem 5.3.1 67

5.4.2 Key Lemmas 69

5.4.3 Proof of Theorem 5.3.1 70

5.5 Experimental Results 71

CHAPTER 6. Conclusions and Future Work 78

APPENDIX A. Proof of the Lemmas and Corollaries in Chapter 2 79

A.1 Proof of Lemma 2.2.4 79

A.2 Proof of Lemma 2.3.1 80

A.3 Proof of Corollary 2.3.4 80

A.4 Proof of Corollary 2.3.5 81

APPENDIX B. Proof of Lemma 3.3.2 82

APPENDIX C. Proof of the Lemmas in Chapter 4 83

C.1 Proof of Lemma 4.4.10 83

C.2 Simple facts 83
C.3 Proof of Lemma 4.4.11 84
C.4 Proof of Lemma 4.4.12 85
C.5 Key facts for proving Lemmas 4.4.14 and 4.4.15 86
C.6 Proof of Lemma 4.4.14 88
C.7 Proof of Lemma 4.4.15 89
C.8 Proof of Lemma 4.4.18 92
C.9 Proof of Lemma 4.4.21 93

APPENDIX D. Proof of the Lemmas in Chapter 5 94

D.1 Proof of Lemma 5.4.15 94
D.2 Lemmas used to prove Lemma 5.4.16 95
D.2.1 A lemma needed for bounding the subspace error, $\tilde{\zeta}_k$ 97
D.2.2 Bounding on the subspace error, $\tilde{\zeta}_k$ 97
D.2.3 Proof of Lemma D.2.7 99
D.3 Proof of Lemma D.2.3 104

BIBLIOGRAPHY 106

LIST OF TABLES

Table 5.1 Comparing and contrasting the addition proj-PCA step and proj-PCA used in the deletion step (cluster-PCA) 68

LIST OF FIGURES

Figure 3.1 The subspace change model. 16

Figure 3.2 Verification of Slow Subspace Change. 19

Figure 4.1 The K times projection PCA algorithm 24

Figure 4.2 ReProCS with $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 10$. 47

Figure 4.3 ReProCS with $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 50$. 48

Figure 4.4 ReProCS with $r_0 = 36$, $s = \max_t |T_t| = 100$ and $\Delta = 10$. 49

Figure 5.1 Illustration of the clustering assumption (assume $\Lambda_t = \Lambda_{\tilde{t}_j}$). 51

Figure 5.2 A diagram illustrating subspace estimation by ReProCS-cPCA 55

Figure 5.3 ReProCS-cPCA with $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 10$. 74

Figure 5.4 ReProCS-cPCA with $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 50$ 75

ABSTRACT

In this work, we study the problem of recursively recovering a time sequence of sparse vectors, S_t, from measurements $M_t := S_t + L_t$ that are corrupted by structured noise L_t which is dense and can have large magnitude. The structure that we require is that L_t should lie in a low dimensional subspace that is either fixed or changes "slowly enough"; and the eigenvalues of its covariance matrix are "clustered". We do not assume anything about the sequence of sparse vectors, except a bound on their support size. Their support sets and their nonzero element values may be either independent or correlated over time (usually in many applications they are correlated). 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, we introduce a novel solution called Recursive Projected Compressive Sensing (ReProCS). Under mild assumption, we show that ReProCS can exactly recover the support set of S_t at all times; and the reconstruction errors of both S_t and L_t are upper bounded by a time-invariant and small value at all times. ReProCS is designed under the assumption that the subspace in which the most recent several L_t's lie can only grow over time. Therefore, it needs to assume a bound on the total number of subspace changes, J. To address this limitation, we introduce a novel subspace estimation scheme called cluster-PCA and we refer to the resulting algorithm as ReProCS with cluster-PCA (ReProCS-cPCA). ReProCS-cPCA does not need a bound on J as long as the delay between subspace change times increases in proportion to $\log J$. An extra assumption that is needed though is that the eigenvalues of the covariance matrix of L_t are sufficiently clustered. As a by-product, at certain times, the basis vectors for the subspace in which the most recent several L_t's lies is also recovered.

CHAPTER 1. Introduction

In this work, we study the problem of recovering a time sequence of sparse vectors, S_t, from measurements $M_t := S_t + L_t$ that are corrupted by large magnitude but dense and structured noise, L_t. The structure that we require is that L_t should lie in a low dimensional subspace that is either fixed or changes "slowly enough"; and the eigenvalues of its covariance matrix are "clustered". As a by-product, at certain times, we are also able to recover a basis matrix for the subspace in which the recent several L_t's lies. Thus, at these times, we also solve the recursive robust principal components' analysis (PCA) problem. For recursive robust PCA, L_t is the signal of interest while S_t can be interpreted as the outlier (sparse noise).

A key application where the above problem occurs is in video analysis where the goal is to separate a slowly changing background from moving foreground objects [1, 2]. If we stack each frame as a column vector, the background is well modeled as lying in a low dimensional subspace that may gradually change over time, while the moving foreground objects constitute the sparse vectors [2,3] which change in a correlated fashion over time. Another key application is online detection of brain activation patterns from functional MRI (fMRI) sequences. In this case, the "active" region of the brain is the the correlated sparse vector.

Many of the older works on sparse recovery with structured noise study the case of sparse recovery from large but sparse noise (outliers), e.g., [3–5]. However, here we are interested in sparse recovery in large but low dimensional noise. On the other hand, most older works on robust PCA cannot recover the outlier (S_t) when its nonzero entries have magnitude much smaller than that of the low dimensional part (L_t) [1, 6, 7]. The main goal of this work is to study sparse recovery and hence we do not discuss these older works here. Some recent works on robust PCA such as [8, 9] assume that an entire measurement vector M_t is either an inlier

(S_t is a zero vector) or an outlier (all entries of S_t can be nonzero), and a certain number of M_t's are inliers. These works also cannot be used when all S_t's are nonzero but sparse.

In a series of recent works [2, 10], a new and elegant solution, which is referred to as Principal Components' Pursuit (PCP) in [2], has been proposed. It redefines batch robust PCA as a problem of separating a low rank matrix, $\mathcal{L}_t := [L_1, \ldots, L_t]$, from a sparse matrix, $\mathcal{S}_t := [S_1, \ldots, S_t]$, using the measurement matrix, $\mathcal{M}_t := [M_1, \ldots, M_t] = \mathcal{L}_t + \mathcal{S}_t$. Thus these works can be interpreted as batch solutions to sparse recovery in large but low dimensional noise. 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 [11–20].

It was shown in [2] that, with high probability (w.h.p.), one can recover $\mathcal{L}_t$ and $\mathcal{S}_t$ exactly by solving

$$\min_{\mathcal{L},\mathcal{S}} \|\mathcal{L}\|_* + \lambda \|\mathcal{S}\|_{1,\text{vec}} \text{ subject to } \mathcal{L} + \mathcal{S} = \mathcal{M}_t \tag{1.1}$$

provided that (a) $\mathcal{L}_t$ is dense (its left and right singular vectors satisfy certain conditions); (b) any element of the matrix $\mathcal{S}_t$ is nonzero w.p. ϱ, and zero w.p. $1 - \varrho$, independent of all others (in particular, this means that the support sets of the different S_t's are independent over time); and (c) the rank of $\mathcal{L}_t$ and the support size of $\mathcal{S}_t$ are small enough. Here $\|B\|_*$ is the nuclear norm of B (sum of singular values of B) while $\|B\|_{1,\text{vec}}$ is the ℓ_1 norm of B seen as a long vector. In most applications, it is fair to assume that the low dimensional part, L_t (background in case of video) is dense. However, the assumption that the support of the sparse part (foreground in case of video) is independent over time is often not valid. Foreground objects typically move in a correlated fashion, and may even not move for a few frames. This results in $\mathcal{S}_t$ being sparse and low rank.

The question then is, what can we do if $\mathcal{L}_t$ is low rank and dense, but $\mathcal{S}_t$ is sparse and may also be low rank? In this case, without any extra information, in general, it is not possible to separate $\mathcal{S}_t$ and $\mathcal{L}_t$. Suppose that an initial short sequence of L_t's is available. For example, in the video application, it is often realistic to assume that an initial background-only training sequence is available. Can we use this to do anything better?

One possible solution is as follows. We can compute the matrix containing the left singular vectors of the initial short training sequence, $\hat{P}_0$. This can be used to modify PCP as follows. We solve

$$\min_{\mathcal{S}} \|\mathcal{S}\|_1, \text{ subject to } \|(I - \hat{P}_0\hat{P}_0')(\mathcal{M}_t - \mathcal{S})\|_F \leq \epsilon, \tag{1.2}$$

where $\|.\|_F$ is the Frobenius norm. This then becomes the standard ℓ_1 minimization solution for a batch sparse recovery problem in noise. As we show later in Lemma 3.3.2, denseness of $\hat{P}_0$ ensures that the restricted isometry constant of $(I - \hat{P}_0\hat{P}_0')$ is small and hence $\mathcal{S}_t$ can be recovered accurately by solving (1.2) as long as the "noise" it sees is small. Here the "noise" is $(I - \hat{P}_0\hat{P}_0')\mathcal{L}_t$. This is small only if span$(\hat{P}_0)$ approximately contains span$(\mathcal{L}_t)$, i.e. the subspace spanned by the future background frames is an approximate subset of that of the initial training dataset. This is unreasonable to expect in a long sequence. Even though the change of subspace from one time instant to the next is usually "slow", the net change over a long sequence can be significant.

We introduced the Recursive Projected Compressive Sensing (ReProCS) algorithm that provided one possible solution to this problem by using the extra piece of information that an initial short sequence of L_t's, or L_t's in small noise, is available (which can be used to get an accurate estimate of the subspace in which the initial L_t's lie) and assuming slow subspace change (as explained in Sec. 3.2). The key idea of ReProCS is as follows. At time t, assume that a $n \times r$ matrix with orthonormal columns, $\hat{P}_{(t-1)}$, is available with span$(\hat{P}_{(t-1)}) \approx$ span$(\mathcal{L}_{t-1})$. We project M_t perpendicular to span$(\hat{P}_{(t-1)})$. Because of slow subspace change, this cancels out most of the contribution of L_t. Recovering S_t from the projected measurements then becomes a classical sparse recovery / compressive sensing (CS) problem in small noise [21]. Under a denseness assumption on span$(\mathcal{L}_{t-1})$, one can show that S_t can be accurately recovered via ℓ_1 minimization. Thus, $L_t = M_t - S_t$ can also be recovered accurately. We use the estimates of L_t in a projection-PCA based subspace estimation algorithm to update $\hat{P}_{(t)}$.

ReProCS assumes that the subspace in which the most recent several L_t's lie can only grow over time. It assumes a model in which at every subspace change time, t_j, some new directions get added to this subspace. After every subspace change, it uses projection-PCA to estimate the newly added subspace. As a result the rank of $\hat{P}_{(t)}$ keeps increasing with every subspace change. Therefore, the number of effective measurements available for the CS step, $(n - \text{rank}(\hat{P}_{(t-1)}))$, keeps reducing. To keep this number large enough at all times, ReProCS needs to assume a bound on the total number of subspace changes, J.

In practice, usually, the dimension of the subspace in which the most recent several L_t's lie 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 existing directions can get deleted from this subspace and to assume an upper bound on the difference between the total number of added and deleted directions (the earlier model is a special case of this). We introduce a novel approach called *cluster-PCA* that re-estimates the current subspace after the newly added directions have been accurately estimated. This re-estimation step ensures that the deleted directions have been "removed" from the new $\hat{P}_{(t)}$. We refer to the resulting algorithm as *ReProCS-cPCA*. We will see that ReProCS-cPCA does not need a bound on J as long as the delay between subspace change times increases in proportion to $\log J$. An extra assumption that is needed though is that the eigenvalues of the covariance matrix of L_t are sufficiently clustered at certain times as explained in Sec 5.1.

Under the clustering assumption and some other mild assumptions, we show that, w.h.p, at all times, ReProCS-cPCA can exactly recover the support of S_t, and the reconstruction errors of both S_t and L_t are upper bounded by a time invariant and small value. Moreover, we show that the subspace recovery error decays roughly exponentially with every projection-PCA step. The proof techniques developed in this work are very different from those used to obtain performance guarantees in recent batch robust PCA works such as [2, 8–12, 16–20, 22]. Our proof utilizes sparse recovery results [21]; results from matrix perturbation theory (sin θ theorem [23] and Weyl's theorem [24]) and the matrix Hoeffding inequality [25].

Our result for ReProCS and ReProCS-cPCA do not assume any model on the sparse vectors, S_t's. In particular, it allows the support sets of the S_t's to be either independent, e.g. generated via the model of [2] (resulting in $\mathcal{S}_t$ being full rank w.h.p.), or correlated over time (can result in $\mathcal{S}_t$ being low rank). The only thing that is required is that there be *some* support changes every so often. We should point out that some of the other works that study the batch problem, e.g. [16], also allow $\mathcal{S}_t$ to be low rank.

A key difference of our work compared with most existing work analyzing finite sample PCA, e.g. [26], and references therein, is that in these works, the noise/error in the observed data is independent of the true (noise-free) data. However, in our case, because of how $\hat{L}_t$ 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 the main reason we need to develop and analyze projection-PCA based approaches for both subspace addition and deletion.

ReProCS and ReProCS-cPCA approaches are related to that of [27–29] 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. However, the big difference is that in all of these works the basis for the subspace of the low dimensional signal is *perfectly known.* We study *the case where the subspace is not known and can change over time.*

1.1 Notation

For a set $T \subseteq \{1, 2, \dots n\}$, 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. $T^c := \{i \in \{1, 2, \dots n\} : i \notin T\}$. The notations $T_1 \subseteq T_2$ and $T_2 \supseteq T_1$ both mean that T_1 is a subset of T_2.

We use the notation $[t_1, t_2]$ to denote the interval that contains t_1 and t_2, as well as all integers between them, i.e. $[t_1, t_2] := \{t_1, t_1 + 1, \cdots, t_2\}$. The notation $[L_t; t \in [t_1, t_2]]$ is used to denote the matrix $[L_{t_1}, L_{t_1+1}, \cdots, L_{t_2}]$.

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, $\text{supp}(v)$, 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 $|\text{supp}(v)| \leq s$.

For a tall matrix P, $\text{span}(P)$ denotes the subspace spanned by the column vectors of P.

For a matrix B, B' denotes its transpose, and $B^\dagger$ denotes 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 and $\|B\|_{\max}$ denotes the maximum over the absolute values of all its entries. We let $\sigma_i(B)$ denote 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 (EVD) of B. Here U is an orthonormal matrix and Λ is a diagonal matrix with entries arranged in non-increasing 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, we have $\|B\|_2 = \sqrt{\max(\lambda_{\max}^2(B), \lambda_{\min}^2(B))}$. Thus, for a $b \geq 0$, $\|B\|_2 \leq b$ implies that $-b \leq \lambda_{\min}(B) \leq \lambda_{\max}(B) \leq b$. If B is a Hermitian p.s.d. matrix, then $\|B\|_2 = \lambda_{\max}(B)$.

The notation [.] denotes an empty matrix. We use I to denote an identity matrix. For an $m \times n$ matrix B and an index set $T \subseteq \{1, 2, \dots n\}$, B_T is the sub-matrix of B containing columns with indices in the set T. Notice that $B_T = BI_T$. We use $B \setminus B_T$ to denote B_{T^c} (here $T^c := \{i \in \{1, 2, \cdots, n\} : i \notin T\}$). Given another matrix B_2 of size $m \times n_2$, $[B \; B_2]$ constructs a new matrix by concatenating matrices B and B_2 in horizontal direction. Thus, $[(B \setminus B_T) \; B_2] = [B_{T^c} \; B_2]$. For any matrix B and sets T_1, T_2, $(B)_{T_1,T_2}$ denotes the sub-matrix containing the rows with indices in T_1 and columns with indices in T_2.

Definition 1.1.1 *We refer to a tall matrix P as a* basis matrix *if it satisfies $P'P = I$.*

Definition 1.1.2 *The s-*restricted isometry constant (RIC) *[27], δ_s, for an $n \times m$ matrix Ψ is the smallest real number satisfying $(1 - \delta_s)\|x\|_2^2 \leq \|\Psi_T x\|_2^2 \leq (1 + \delta_s)\|x\|_2^2$ for all sets $T \subseteq \{1, 2, \dots n\}$ with $|T| \leq s$ and all real vectors x of length $|T|$.*

It is easy to see that $\max_{T:|T|\leq s} \|({\Psi_T}'\Psi_T)^{-1}\|_2 \leq \frac{1}{1-\delta_s(\Psi)}$ [27].

Definition 1.1.3 *Let X and Z be two random variables (r.v.) and let $\mathcal{B}$ be a set of values that Z can take.*

1. *We use $\mathcal{B}^e$ to denote the* event $Z \in \mathcal{B}$*, i.e. $\mathcal{B}^e := \{Z \in \mathcal{B}\}$.*

2. *The probability of event $\mathcal{B}^e$ can be expressed as [30],*

$$\mathbf{P}(\mathcal{B}^e) := \mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)].$$

where

$$\mathbb{I}_{\mathcal{B}}(Z) := \begin{cases} 1 & \text{if } Z \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases}$$

is an indicator function of Z on the set $\mathcal{B}$ and $\mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)]$ is the expectation of $\mathbb{I}_{\mathcal{B}}(Z)$.

3. *Define $\mathbf{P}(\mathcal{B}^e|X) := \mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)|X]$ where $\mathbf{E}[\mathbb{I}_{\mathcal{B}}(Z)|X]$ is the conditional expectation of $\mathbb{I}_{\mathcal{B}}(Z)$ given X.*

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".

1.2 Dissertation Organization

The dissertation is organized as follows. In Chapter 2, we give the mathematical preliminaries. In Chapter 3, we give the problem definition followed by the model and key assumptions. We discuss the ReProCS algorithm and its performance guarantees in Chapter 4. ReProCS with cluster-PCA and its performance grantees are presented in Chapter 5. Finally, conclusions are summarized in Chapter 6. Many parts of these chapters are taken verbatim from [31] [32] [33] [34].

CHAPTER 2. Mathematical Preliminaries

In this section, we state certain results from the literature, or certain lemmas which follow easily using these results, that will be used later. Parts of this chapter are taken verbatim from [31] [32] [33] [34].

2.1 Compressive Sensing result

Compressed sensing is a signal processing technique for efficiently acquiring and reconstructing a signal, by finding solutions to underdetermined linear systems. This takes advantage of the signal's sparseness or compressibility in some domain, allowing the entire signal to be determined from relatively few measurements. The error bound for noisy compressive sensing (CS) based on the RIC is as follows [21].

Theorem 2.1.1 ([21]) *Suppose we observe*

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

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

$$\min_x \|x\|_1 \text{ subject to } \|y - \Psi x\|_2 \leq \xi \tag{2.1}$$

Assume that x is s-sparse, $\|z\|_2 \leq \xi$, and $\delta_{2s}(\Psi) < b(\sqrt{2}-1)$ with a $0 \leq b < 1$. Then the solution of (2.1) obeys

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

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

Remark 2.1.2 *Notice that if b is small enough, C_1 is a small constant but $C_1 > 1$. For example, if $\delta_{2s}(\Psi) \leq 0.15$, then $C_1 \leq 7$. If $C_1\xi > \|x\|_2$, the normalized reconstruction error bound* *would be greater than 1, making the result useless. Hence, (2.1) gives a small reconstruction error bound only for the small noise case, i.e., the case where $\|z\|_2 \leq \xi \ll \|x\|_2$. In fact this is true for most existing literature on CS and sparse recovery, with the exception of [3–5] (focus on large but sparse noise) and [2, 10].*

2.2 Results from linear algebra

Davis and Kahan's $\sin\theta$ theorem [23] studies the rotation of eigenvectors by perturbation.

Theorem 2.2.1 ($\sin\theta$ theorem [23]) *Given two Hermitian matrices $\mathcal{A}$ and $\mathcal{H}$ satisfying*

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

where $[E \;\, E_\perp]$ is an orthonormal matrix. 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

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

The above result bounds the amount by which the two subspaces span(E) and 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 2.2.2 (Weyl [24]) *Let $\mathcal{A}$ and $\mathcal{H}$ be two $n \times n$ Hermitian matrices. For each $i = 1, 2, \ldots, n$ we have*

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

Theorem 2.2.3 (Ostrowski [24]) *Let H and W be $n \times n$ matrices, with H Hermitian and W nonsingular. For each $i = 1, 2 \dots n$, there exists a positive real number θ_i such that $\lambda_{\min}(WW') \leq \theta_i \leq \lambda_{\max}(WW')$ and $\lambda_i(WHW') = \theta_i \lambda_i(H)$. Therefore,*

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

The following lemma proves some simple linear algebra facts.

Lemma 2.2.4 *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. $\|PP' - \hat{P}\hat{P}'\|_2 \leq 2\|(I - \hat{P}\hat{P}')P\|_2 = 2\zeta_*$
3. $\|\hat{P}'Q\|_2 \leq \zeta_*$
4. $\sqrt{1 - \zeta_*^2} \leq \sigma_i((I - \hat{P}\hat{P}')Q) \leq 1$

The proof is in the Appendix A.

2.3 Simple probability facts and matrix Hoeffding inequalities

The following lemma follows easily using Definition 1.1.3.

Lemma 2.3.1 *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 $0 \leq p \leq 1$, if $\mathbf{P}(\mathcal{B}^e|X) \geq p$ for all $X \in \mathcal{C}$, then $\mathbf{P}(\mathcal{B}^e|\mathcal{C}^e) \geq p$ as long as $\mathbf{P}(\mathcal{C}^e) > 0$.*

The proof is in Appendix A.

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

Lemma 2.3.2 *For a sequence of events $E_0^e, E_1^e, \dots E_m^e$ that satisfy $E_0^e \supseteq E_1^e \supseteq E_2^e \dots \supseteq E_m^e$, 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

$$\mathbf{P}(E^e_m|E^e_0) = \mathbf{P}(E^e_m, E^e_{m-1}, \dots E^e_0|E^e_0) = \prod_{k=1}^{m} \mathbf{P}(E^e_k|E^e_{k-1}, E^e_{k-2}, \dots E^e_0)$$
$$= \prod_{k=1}^{m} \mathbf{P}(E^e_k|E^e_{k-1})$$

■

Next, we state the matrix Hoeffding inequality [25, Theorem 1.3] which gives tail bounds for sums of independent random matrices.

Theorem 2.3.3 (Matrix Hoeffding for a zero mean Hermitian matrix [25]) *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* $\epsilon > 0$*,*

$$\mathbf{P}(\lambda_{\max}(\sum_t Z_t) \leq \epsilon) \geq 1 - n\exp(-\frac{\epsilon^2}{8\sigma^2}), \text{ where } \sigma^2 = \|\sum_t A_t^2\|_2$$

The following two corollaries of Theorem 2.3.3 are easy to prove. The proofs are given in the Appendix A.

Corollary 2.3.4 (Matrix Hoeffding for a nonzero mean Hermitian 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)* $\mathbf{P}(b_1 I \preceq Z_t \preceq b_2 I|X) = 1$ *and (iii)* $b_3 I \preceq \frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X) \preceq b_4 I$*. Then for all* $\epsilon > 0$*,*

$$\mathbf{P}(\lambda_{\max}(\frac{1}{\alpha}\sum_t Z_t) \leq b_4 + \epsilon|X) \geq 1 - n\exp(-\frac{\alpha\epsilon^2}{8(b_2 - b_1)^2}) \text{ for all } X \in$$
$$\mathbf{P}(\lambda_{\min}(\frac{1}{\alpha}\sum_t Z_t) \geq b_3 - \epsilon|X) \geq 1 - n\exp(-\frac{\alpha\epsilon^2}{8(b_2 - b_1)^2}) \text{ for all } X \in \mathcal{C}$$

The proof is in the Appendix A.

Corollary 2.3.5 (Matrix Hoeffding 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 $\epsilon > 0$,

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

The proof is in the Appendix A.

CHAPTER 3. Problem Definition and Model Assumptions

In this chapter, we give the problem definition below followed by the model and key assumptions. Parts of this chapter are taken verbatim from [31] [32] [33] [34].

3.1 Problem Definition

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

$$M_t = L_t + S_t \tag{3.1}$$

Here S_t is a sparse vector with support set size at most s and minimum magnitude of nonzero values at least $S_{\min}$. L_t is a dense but low dimensional vector, i.e. $L_t = P_{(t)} a_t$ where $P_{(t)}$ is an $n \times r_{(t)}$ basis matrix with $r_{(t)} \ll n$, that changes every so often. $P_{(t)}$ and a_t change 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 either zero or very small. The goal is

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

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

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

Assumption 3.1.1 (Model on L_t) *We assume that $L_t = P_{(t)} a_t$ where $P_{(t)}$ and a_t satisfy the following.*

1. *$P_{(t)} = P_j$ for all $t_j \leq t < t_{j+1}$, $j = 0, 1, 2 \cdots J$, where P_j is an $n \times r_j$ basis matrix with $r_j \ll n$ and $r_j \ll (t_{j+1} - t_j)$. We let $t_0 = 0$ and t_{J+1} equal the sequence length. This can be infinity also. At the change times, t_j, P_j changes as $P_j = [(P_{j-1} \setminus P_{j,old})\ P_{j,new}]$. Here, $P_{j,new}$ is an $n \times c_{j,new}$ basis matrix with $P_{j,new}' P_{j-1} = 0$ and $P_{j,old}$ contains $c_{j,old}$ columns of P_{j-1}. Thus $r_j = r_{j-1} + c_{j,new} - c_{j,old}$. Also, $0 < t_{train} \leq t_1$. This model is illustrated in Fig. 3.2.*

2. *There exists a constant $c_{\max}$ such that $0 \leq c_{j,new} \leq c_{\max}$ and $\sum_{i=1}^{j} (c_{i,new} - c_{i,old}) \leq c_{\max}$ for all j. Thus, $r_j = r_0 + \sum_{i=1}^{j} (c_{i,new} - c_{i,old})$.*

3. *$a_t := P_{(t)}{}' L_t$, is a r_j length random variable (r.v.) with the following properties.*

   (a) *a_t's are mutually independent over t.*

   (b) *a_t is a zero mean bounded r.v., i.e. $\mathbf{E}(a_t) = 0$ and there exists a constant γ_* such that $\|a_t\|_\infty \leq \gamma_*$ for all t.*

   (c) *Its covariance matrix $\Lambda_t := Cov[a_t] = \mathbf{E}(a_t a_t')$ is diagonal with $\lambda^- := \min_t \lambda_{\min}(\Lambda_t) > 0$ and $\lambda^+ := \max_t \lambda_{\max}(\Lambda_t) < \infty$. Thus, the condition number of any Λ_t is bounded by $f := \frac{\lambda^+}{\lambda^-}$.*

Also, P_j and a_t satisfy the assumptions discussed in the next two subsections.

Definition 3.1.2 *The following notation will be used frequently. Let $P_{j,*} := P_{(t_j - 1)} = P_{j-1}$. For $t \in [t_j, t_{j+1} - 1]$, let $a_{t,*} := P_{j,*}{}' L_t = P_{j-1}{}' L_t$ be the projection of L_t along $P_{j,*}$ of which $a_{t,*,nz} := (P_{j-1} \setminus P_{j,old})' L_t$ is the nonzero part. Also, let $a_{t,new} := P_{j,new}' L_t$ be the projection of L_t along the newly added directions. Thus,*

$$a_{t,*} = \begin{bmatrix} a_{t,*,nz} \\ \mathbf{0} \end{bmatrix} \text{ and } a_t = \begin{bmatrix} a_{t,*,nz} \\ a_{t,new} \end{bmatrix}$$

where $\mathbf{0}$ *is a* $c_{j,old}$ *length zero vector (since* ${P_{j,old}}' L_t = \mathbf{0}$*). Using the above, for* $t \in [t_j, t_{j+1}-1]$*,* L_t *can be rewritten as*

$$L_t = P_j a_t = (P_{j-1} \setminus P_{j,old}) a_{t,*,nz} + P_{j,new} a_{t,new} = P_{j,*} a_{t,*} + P_{j,new} a_{t,new}$$

and Λ_t *can be split as*

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

where $(\Lambda_t)_{*,nz} := Cov(a_{t,*,nz})$ *and* $(\Lambda_t)_{new} = Cov(a_{t,new})$ *are diagonal matrices.*

3.2 Slow Subspace Change

By slow subspace change we mean all of the following.

1. First, the delay between consecutive subspace change times, $t_{j+1} - t_j$, is large enough.

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

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

   for a $v > 1$ but not too large[1].

3. Third, the number of newly added directions is small, i.e. $c_{j,\text{new}} \le c_{\max} \ll r_0$. This is verified in Sec. 3.4.

3.3 Denseness assumption and its relation with RIC

For a tall $n \times r$ matrix, B, or for a $n \times 1$ vector, B, we define the the denseness coefficient as follows [32]:

$$\kappa_s(B) := \max_{|T| \le s} \frac{\|{I_T}' B\|_2}{\|B\|_2}. \tag{3.2}$$

[1] Small γ_{new} and slowly increasing $\gamma_{\text{new},k}$ is needed for the noise seen by the sparse recovery step to be small. However, if γ_{new} is zero or very small, it will be impossible to estimate the new subspace. This will not happen in our model because $\gamma_{\text{new}} \ge \lambda^- > 0$.

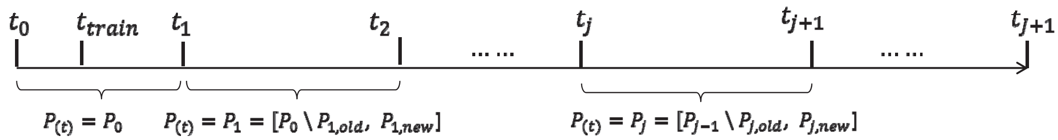


Figure 3.1 The subspace change model.

where $\|.\|_2$ is the matrix or vector 2-norm respectively. Clearly, $\kappa_s(B) \leq 1$. The denseness coefficient measures the denseness (non-compressibility) of a vector B or of the columns of a matrix B. For a vector, a small value indicates that its entries are spread out, i.e. it is a dense vector. A large value indicates that it is compressible (approximately or exactly sparse). Similarly, for a matrix B, a small value means that most (or all) of its columns are dense vectors.

Remark 3.3.1 *The following facts should be noted about $\kappa_s(.)$.*

1. *For an $n \times r$ matrix B, $\kappa_s(B)$ is a non-decreasing function of s.*
2. *For an $n \times r$ basis matrix B, $\kappa_s(B)$ is a non-decreasing function of $r = rank(B)$.*
3. *A loose bound on $\kappa_s(B)$ obtained using triangle inequality is $\kappa_s(B) \leq s\kappa_1(B)$.*
4. *For a basis matrix P, $\|P\|_2 = 1$ and hence $\kappa_s(P) = \max_{|T| \leq s} \|I_T' P\|_2$ and $\kappa_s(PP') = \kappa_s(P)$. Thus, for any other basis matrix Q for which $span(Q) = span(P)$, $\kappa_s(P) = \kappa_s(Q)$. Thus, $\kappa_s(P)$ is a property of $span(P)$, which is the subspace spanned by the columns of P, and not of the actual entries of P.*

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

Lemma 3.3.2 *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. Thus, using Theorem 2.1.1, all s-sparse vectors, S_t can be accurately recovered from $y_t := (I - PP')S_t + \beta_t$ if β_t is small noise.

3.4 Model Verification

We now discuss 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. 3.1. 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 $\mathcal{L}_{1,p} = (I - P_0 P_0')\mathcal{L}_{d+1:2d}$; 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 $c_{mx} = 3$ and $r_0 = 8$ for the lake sequence and $c_{mx} = 5$ and $r_0 = 29$ for the curtain sequence. Thus the ratio c_{mx}/r_0 is sufficiently small in both cases. In Fig 3.2, 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 $\gamma_{\text{new}} := \max_j \max_{t_j \le t < t_j + \alpha} ||a_{t,\text{new}}||_\infty = 0.125\gamma_*$ 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 $\|a_{t,\text{new}}\|_\infty \le \max(v^{k-1}\gamma_{\text{new}}, \gamma_*)$ 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.

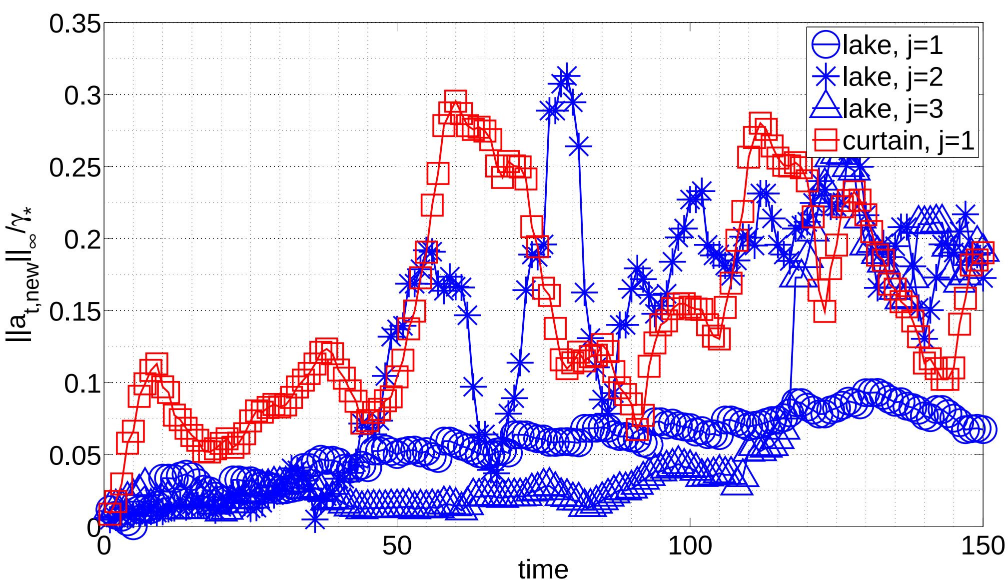


Figure 3.2 Verification of Slow Subspace Change.

CHAPTER 4. Recursive Projected CS (ReProCS) and its Performance Guarantees

ReProCS considers the case that $c_{j,\text{old}} = 0$ for all j. Therefore, $P_j = [P_{j-1}\ P_{j,\text{new}}]$ and $r_j = r_{j-1} + c_{j,new}$. In Sec. 4.1, we first explain the main idea of projection-PCA (proj-PCA). In Sec 4.2, we explain the ReProCS algorithm and why it works. We summarize the Recursive Projected CS (ReProCS) algorithm in Algorithm 2. It uses the following definition.

Definition 4.0.1 *Define the time interval* $\mathcal{I}_{j,k} := [t_j + (k-1)\alpha, t_j + k\alpha - 1]$ *for* $k = 1, \dots K$ *and* $\mathcal{I}_{j,K+1} := [t_j + K\alpha, t_{j+1} - 1]$. *Here,* K *is the algorithm parameter in Algorithm 2.*

We give the performance guarantees (Theorem 4.3.1) in Sec 4.3. The proof of Theorem 4.3.1 is given in Sec 4.4.4. In Sec 4.6, we show numerical experiments demonstrating Theorem 4.3.1, as well as the comparisons with PCP. Parts of this chapter are taken verbatim from [31] [32].

4.1 The Projection-PCA algorithm

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 $\frac{1}{\alpha_{\mathcal{D}}}\mathcal{D}_{\text{proj}}\mathcal{D}_{\text{proj}}' \overset{EVD}{=} \begin{bmatrix} Q\ Q_\perp \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda_\perp \end{bmatrix} \begin{bmatrix} Q' \\ Q_\perp{}' \end{bmatrix}$ where Q is an $n \times r$ basis matrix and $\alpha_{\mathcal{D}}$ is the number of columns in $\mathcal{D}$.

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 $\frac{1}{\alpha_{\mathcal{D}}}\mathcal{D}_{\text{proj}}\mathcal{D}_{\text{proj}}'$. Here $\alpha_{\mathcal{D}}$ 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 $\frac{1}{\alpha_{\mathcal{D}}}\mathcal{D}\mathcal{D}'$.

We should mention that the idea of projecting perpendicular to a partly estimated subspace has been used in different contexts in past work [8, 35].

Algorithm 2 Recursive Projected CS (ReProCS)

Parameters: algorithm parameters: ξ, ω, α, K, model parameters: t_j, r_0, $c_{j,\text{new}}$ (set as in Theorem 4.3.1)

Input: M_t, *Output:* $\hat{S}_t$, $\hat{L}_t$, $\hat{P}_{(t)}$

Initialization: Given training sequence $[L_t : 1 \le t \le t_{\text{train}}]$, $\hat{P}_0 \leftarrow$ proj-PCA$([L_t : 1 \le t \le t_{\text{train}}], [.], r_0)$. Let $\hat{P}_{(t)} \leftarrow \hat{P}_0$. Let $j \leftarrow 1$, $k \leftarrow 1$. For $t > t_{\text{train}}$, do the following:

1. Estimate T_t and S_t via Projected CS:
   (a) Nullify most of L_t: compute $\Phi_{(t)} \leftarrow I - \hat{P}_{(t-1)}\hat{P}'_{(t-1)}$, compute $y_t \leftarrow \Phi_{(t)} M_t$
   (b) Sparse Recovery: compute $\hat{S}_{t,\text{cs}}$ as the solution of $\min_x \|x\|_1 \ s.t. \ \|y_t - \Phi_{(t)} x\|_2 \le \xi$
   (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 $\hat{P}_{(t)}$ by Projection PCA
   (a) If $t = t_j + k\alpha - 1$,
      i. $\hat{P}_{j,\text{new},k} \leftarrow$ proj-PCA$([\hat{L}_t : t \in \mathcal{I}_{j,k}], \hat{P}_{j-1}, c_{j,\text{new}})$
      ii. set $\hat{P}_{(t)} \leftarrow [\hat{P}_{j-1} \ \hat{P}_{j,\text{new},k}]$; increment $k \leftarrow k + 1$.
   Else
      i. set $\hat{P}_{(t)} \leftarrow \hat{P}_{(t-1)}$.
   (b) If $t = t_j + K\alpha - 1$, then set $\hat{P}_j \leftarrow [\hat{P}_{j-1} \ \hat{P}_{j,\text{new},K}]$. Increment $j \leftarrow j + 1$. Reset $k \leftarrow 1$.
4. Increment $t \leftarrow t + 1$ and go to step 1.

4.2 The Recursive Projected CS (ReProCS) Algorithm

The key idea of ReProCS is as follows. Assume that the current basis matrix $P_{(t)}$ has been accurately predicted using past estimates of L_t, i.e. we have $\hat{P}_{(t-1)}$ with $\|(I - \hat{P}_{(t-1)}\hat{P}'_{(t-1)})P_{(t)}\|_2$ small. We project 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 $r_* = \operatorname{rank}(\hat{P}_{(t-1)})$, therefore y_t has only $n - r_*$ "effective" measurements[1], even though its 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 $\|(I - \hat{P}_{(t-1)} \hat{P}'_{(t-1)}) P_{(t)}\|_2$ 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 [36–38]. 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 3.3.2, the RIC of $\Phi_{(t)}$ is small enough. Using Theorem 2.1.1, this ensures that S_t can be accurately recovered from y_t.

By thresholding on the recovered S_t, one gets an estimate of its support (step 1c). By computing a least squares (LS) estimate of S_t on the estimated support and setting it to zero everywhere else (step 1d), we can get a more accurate final estimate, $\hat{S}_t$, as first suggested in [39]. 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 4.4.11, if 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:

$$e_t = I_{T_t} (\Phi_{(t)})_{T_t}{}^{\dagger} \beta_t = I_{T_t} [(\Phi_{(t)})'_{T_t} (\Phi_{(t)})_{T_t}]^{-1} I_{T_t}{}' \Phi_{(t)} L_t \tag{4.1}$$

The second equality follows because $(\Phi_{(t)})_T{}' \Phi_{(t)} = (\Phi_{(t)} I_T)' \Phi_{(t)} = I_T{}' \Phi_{(t)}$ for any set T. Consider a $t \in \mathcal{I}_{j,1}$. At this time, L_t satisfies $L_t = P_{j-1} a_{t,*} + P_{j,\text{new}} a_{t,\text{new}}$, $P_{(t)} = P_j = [P_{j-1}, P_{j,\text{new}}]$, $\hat{P}_{(t-1)} = \hat{P}_{j-1}$ and so $\Phi_{(t)} = \Phi_{j,0} := I - \hat{P}_{j-1} \hat{P}'_{j-1}$. Let $\Phi_{j,k} := I - \hat{P}_{j-1} \hat{P}'_{j-1} - \hat{P}_{j,\text{new},k} \hat{P}'_{j,\text{new},k}$ (with $\hat{P}_{j,\text{new},0} = [.]$), $\zeta_{j,k} := \|\Phi_{j,k} P_{j,\text{new}}\|_2$, $\kappa_{s,k} := \max_j \kappa_s(\Phi_{j,k} P_{j,\text{new}})$, $\phi_k := \max_j \max_{|T| \le s} \|[(\Phi_{j,k})'_T (\Phi_{j,k})_T]^{-1}\|_2$, $r_* := r_0 + (j-1) c_{\max}$, and $c := c_{\max}$. We assume that the delay between change times is large enough so that by $t = t_j$, $\hat{P}_{(t-1)} = \hat{P}_{j-1}$ is an accurate enough estimate of P_{j-1}, i.e. $\|\Phi_{j,0} P_{j-1}\|_2 \le r_* \zeta$ for a ζ small enough. Using $\|I_{T_t}{}' \Phi_{j,0} P_{j-1}\|_2 \le \|\Phi_{j,0} P_{j-1}\|_2 \le r_* \zeta$, $\|I_{T_t}{}' \Phi_{j,0} P_{\text{new}}\|_2 \le \kappa_{s,0} \|\Phi_{j,0} P_{j,\text{new}}\|_2$ and $\zeta_{j,0} = \|\Phi_{j,0} P_{\text{new}}\|_2 \le 1$, we get that $\|e_t\|_2 \le \phi_0 r_* \zeta \sqrt{r_*} \gamma_* + \phi_0 \kappa_{s,0} \sqrt{c} \gamma_{\text{new}}$. The denseness assumption on P_{j-1}; $\|\Phi_{j,0} P_{j-1}\|_2 \le r_* \zeta$

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

and $\phi_0 \leq 1/(1-\delta_s(\Phi_{j,0}))$ ensure that ϕ_0 is only slightly more than one (see Lemma 4.4.10). If $\sqrt{\zeta} < 1/\gamma_*$, the first term in the bound on $\|e_t\|_2$ is of the order of $\sqrt{\zeta}$ and hence negligible. The denseness assumption on $\Phi_{j,0}P_{j,\text{new}}$, whose columns span the currently unestimated part of span$(P_{j,\text{new}})$, ensures that $\kappa_{s,0}$ is significantly less than one. As a result, $\phi_0\kappa_{s,0} < 1$ and so the error $\|e_t\|_2$ is of the order of $\sqrt{c}\gamma_{\text{new}}$. 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 projection PCA (step 3). We illustrate the$K$ times projection PCA algorithm in Fig 4.2. 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 $\Phi_{(t)} = \Phi_{j,1}$. 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 4.3.1), under mild assumptions, 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 $\hat{P}_j = [\hat{P}_{j-1}, \hat{P}_{j,\text{new},K}]$.

The reason standard PCA cannot be used and we need proj-PCA is because $e_t = \hat{L}_t - L_t$ is correlated with L_t. The discussion here also applies to recursive or online PCA which is just a fast algorithm for computing standard PCA. In most existing works that analyze finite sample PCA, e.g. see [26] and references therein, the noise or error in the "data" used for PCA (here $\hat{L}_t$'s) is uncorrelated with the true values of the data (here L_t's) and is zero mean. Thus, when computing the eigenvectors of $(1/\alpha)\sum_t \hat{L}_t\hat{L}_t'$, the dominant term of the perturbation, $(1/\alpha)\sum_t \hat{L}_t\hat{L}_t' - (1/\alpha)\sum_t L_tL_t'$, is $(1/\alpha)\sum_t e_te_t'$ (the terms $(1/\alpha)\sum_t L_te_t'$ and its transpose are close to zero w.h.p. due to law of large numbers). By assuming that the error/noise e_t is small enough, the perturbation can be made small enough.

However, for our problem, because e_t and L_t are correlated, the dominant terms in the

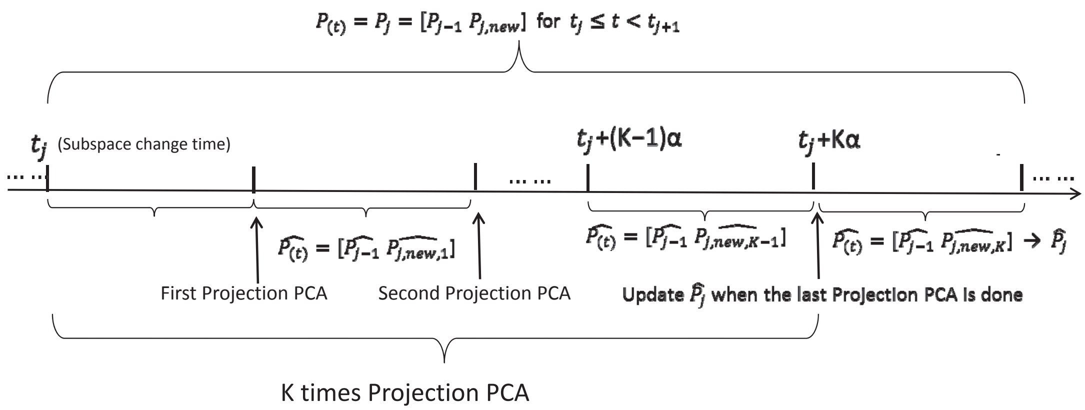


Figure 4.1 The K times projection PCA algorithm

perturbation seen by standard PCA will be $(1/\alpha)\sum_t L_t e_t{}'$ and its transpose. Since L_t can have large magnitude, the bound on the perturbation will be large and this will create problems when applying the $\sin\theta$ theorem (Theorem 2.2.1) to bound the subspace error. On the other hand, when using proj-PCA, L_t gets replaced by $(I - \hat{P}_{j-1}\hat{P}_{j-1}')L_t$ and this results in significantly smaller perturbation.

4.3 Performance Guarantees

We state the performance guarantees of ReProCS in Theorem 4.3.1. The proof outline is given in Sec. 4.4.3 and the actual proof is given in Sec. 4.4.4 the subsequent sections.

Theorem 4.3.1 *Consider Algorithm 2. Let* $c := c_{\max}$ *and* $r := r_0 + (J-1)c$*. Assume that* L_t *obeys the model given in Sec. 3.1 with* $c_{j,old} = 0$ *and there are a total of* J *change times. Assume also that the initial subspace estimate is accurate enough, i.e.* $\|(I - \hat{P}_0\hat{P}_0')P_0\| \le r_0\zeta$*, for a* ζ *that satisfies*

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

If the following conditions hold:

1. *the algorithm parameters are set as* $\xi = \xi_0(\zeta)$, $7\rho\xi \le \omega \le S_{\min} - 7\rho\xi$, $K = K(\zeta)$, $\alpha \ge \alpha_{add}(\zeta)$*, where* $\xi_0(\zeta), \rho, K(\zeta), \alpha_{add}(\zeta)$ *are defined in Definition 4.4.1.*

2. P_{j-1}, $P_{j,new}$, $D_{j,new,k} := (I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,new,k}\hat{P}_{j,new,k}')P_{j,new}$ *and* $Q_{j,new,k} := (I - P_{j,new}{P_{j,new}}')\hat{P}_{j,new,k}$ *have dense enough columns, i.e.*

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

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

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

$$\max_j (t_{j+1} - t_j) > K\alpha,$$
$$\max_j \max_{t_j+(k-1)\alpha \leq t < t_j + k\alpha} \|a_{t,new}\|_\infty \leq \gamma_{new,k} := \min(1.2^{k-1}\gamma_{new}, \gamma_*), \ \textit{for all } k = 1, 2, \dots K,$$
$$14\rho\xi_0(\zeta) \leq S_{\min},$$

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

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

where $g_{j,k}$ *is defined in Definition 4.4.1.*

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

1. *at all times, t,*

$$\hat{T}_t = T_t \quad \textit{and}$$

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

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

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

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

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

$$\leq \begin{cases} 0.18\sqrt{c}0.72^{k-1}\gamma_{new} + 1.2(\sqrt{r}+0.06\sqrt{c})\sqrt{\zeta} & \text{if } t \in \mathcal{I}_{j,k},\ k = 1,\cdots K \\ 1.2\sqrt{r}\sqrt{\zeta} & \text{if } t \in \mathcal{I}_{j,K+1} \end{cases}$$

This result says the following. Consider Algorithm 2. Assume that the initial subspace error is small enough. If (a) the algorithm parameters are set appropriately; (b) the matrices defining the previous subspace, the newly added subspace, and the currently unestimated part of the newly added subspace are dense enough; (c) the subspace change is slow enough; and (d) the condition number of the average covariance matrix of $a_{t,\text{new}}$ is small enough, then, w.h.p., we will get exact support recovery at all times. Moreover, the sparse recovery error will always be bounded by $0.18\sqrt{c}\gamma_{\text{new}}$ plus a constant times $\sqrt{\zeta}$. Since ζ is very small, $\gamma_{\text{new}} \ll S_{\min}$, and c is also small, the normalized reconstruction error for recovering S_t will be small at all times.

In the second conclusion, we bound the subspace estimation error, $\text{SE}_{(t)}$. 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.

We discuss the assumptions used by our result. 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.

Now consider the assumptions made on the model. 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 $14\rho\xi_0 \leq S_{\min}$ 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 Sec. 3.4. As explained there, one cannot estimate the delay between change times with just one video sequence of a particular type (need an ensemble) and hence the first assumption cannot be verified.

We also assume that condition number of the average covariance matrix of $a_{t,\text{new}}$, is not too large. This is an assumption made for simplicity. It can be removed if the newly added eigenvalues can be separated into clusters so that the condition number of each cluster is small (even though the overall condition number is large). This latter assumption is usually true for real data. Under this assumption, we can use the cluster projection PCA approach described in [34] for ReProCS with deletion. The idea is to use projection PCA to first only recover the eigenvectors corresponding to the cluster with the largest eigenvalues; then project perpendicular to these and $\hat{P}_{j-1}$ to recover the eigenvectors for the next cluster and so on.

Other than these, we assume the independence of a_t's over time. This is done so that we can use the matrix Hoeffding inequality [25, Theorem 1.3] to obtain high probability bounds on the terms in the subspace error bound. In simulations, and in experiments with real data, we are able to also deal with correlated a_t's. In future work, it should be possible to replace independence by a milder assumption, e.g. a random walk model on the a_t's. In that case, at $t_j + k\alpha - 1$, one would compute the eigenvectors of $(1/\alpha)\sum_{t\in\mathcal{I}_{j,k}} \Phi_{j,0}(\hat{L}_t - \hat{L}_{t-1})(\hat{L}_t - \hat{L}_{t-1})'\Phi_{j,0}'$. Moreover, one may need to use the matrix Azuma inequality [25, Theorem 7.1] instead of Hoeffding to bound the terms in the subspace error bound.

Finally, we assume denseness of P_{j-1} and $P_{j,\text{new}}$ as well as of $D_{j,\text{new},k}$ and $Q_{j,\text{new},k}$ in condition 2. The denseness assumption of P_{j-1} and $P_{j,\text{new}}$ is a subset of the assumptions made in earlier works [2]. It 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. Thus, most columns of the matrix $\mathcal{L}_t$ are dense and consequently the same is true for any basis matrix for $\text{span}(\mathcal{L}_t)$. Now consider denseness of $D_{j,\text{new},k}$ whose columns span the currently unestimated part of the newly added subspace. Our proof actually only needs $\|{I_{T_t}}' D_{j,\text{new},k}\|_2 / \|D_{j,\text{new},k}\|_2$ to be small at every projection PCA time, $t = t_j + k\alpha - 1$. We attempted to verify this in simulations done with a dense P_j and $P_{j,\text{new}}$. Except for the case of exactly constant support of S_t, in all other cases (including the case of very gradual support change, e.g. the models considered in Sec 4.6), this ratio was small for most projection PCA times. We also saw that even if at a few projection PCA times, this ratio was close to one, that just meant that, at those times, the subspace error remained roughly equal to that at the previous time. As a result, a larger K was required for the subspace error to become small enough. It did not mean that the algorithm became unstable. It should be possible to use a similar idea to modify our result as well. An analogous discussion applies also to $Q_{j,\text{new},k}$. In fact denseness of $Q_{j,\text{new},k}$ is not essential, it is possible to prove a slightly more complicated version of Theorem 4.3.1 without assuming denseness of $Q_{j,\text{new},k}$.

4.4 Proof of Theorem 4.3.1

We first define the various quantities that will be used in the lemmas leading to the proof of Theorem 4.3.1.

Definition 4.4.1 *We define here the parameters used in Theorem 4.3.1.*

1. *Define* $K(\zeta) := \left\lceil \frac{\log(0.6c\zeta)}{\log 0.6} \right\rceil$
2. *Define* $\xi_0(\zeta) := \sqrt{c}\gamma_{new} + \sqrt{\zeta}(\sqrt{r} + \sqrt{c})$
3. *Define* $\rho := \max_t\{\kappa_1(\hat{S}_{t,cs} - S_t)\}$. *Notice that* $\rho \leq 1$.
4. *Let* $K = K(\zeta)$. *We define* $\alpha_{add}(\zeta)$ *as the smallest value of* α *so that* $(p_K(\alpha,\zeta))^{KJ} \geq 1 - n^{-10}$, *where* $p_K(\alpha,\zeta)$ *is defined in Lemma 4.4.16. We can compute an explicit value for* α_{add} *by using the fact that for any* $x \leq 1$ *and* $r \geq 1$, $(1-x)^r \geq 1 - rx$. *This gives us*

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

In words, α_{add} *is the smallest value of the number of data points,* α*, needed for one projection PCA step to ensure that Theorem 4.3.1 holds w.p. at least* $(1-n^{-10})$.

5. *Define the condition number of* $Cov(a_{t,new})$ *averaged over* $t \in \mathcal{I}_{j,k}$ *as*

$$g_{j,k} := \frac{{\lambda_{j,new,k}}^+}{{\lambda_{j,new,k}}^-} \text{ where}$$
$${\lambda_{j,new,k}}^+ := \lambda_{\max}(\frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}}(\Lambda_t)_{new}), \quad {\lambda_{j,new,k}}^- := \lambda_{\min}(\frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}}(\Lambda_t)_{new}),$$

Notice that $\lambda^- \le {\lambda_{j,new,k}}^- \le {\lambda_{j,new,k}}^+ \le \lambda^+$ *and thus* $g_{j,k} \le f = \lambda^+/\lambda^-$. *Recall that* $\Lambda_t = Cov[a_t] = \mathbf{E}(a_t a_t')$, $(\Lambda_t)_{new} = \mathbf{E}(a_{t,new}a'_{t,new})$, $\lambda^- = \min_t \lambda_{\min}(\Lambda_t)$ *and* $\lambda^- = \max_t \lambda_{\max}(\Lambda_t)$.

Definition 4.4.2 *We define the noise seen by the sparse recovery step at time t as*

$$\beta_t := \|(I - \hat{P}_{(t-1)}\hat{P}'_{(t-1)})L_t\|_2.$$

Also the reconstruction error of S_t *is*

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

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

Definition 4.4.3 *We define the subspace estimation errors as follows. Recall that* $\hat{P}_{j,new,0} = [.]$ *(empty matrix).*

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

Remark 4.4.4 *Recall from the model given in Sec 3.1 and from Algorithm 2 that*

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

From Definition 4.4.3 and the above, it is easy to see that

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

Definition 4.4.5 *Define the following*

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

Definition 4.4.6

1. *Let* $D_{j,new} \overset{QR}{=} E_{j,new}R_{j,new}$ *denote its QR decomposition. Here* $E_{j,new}$ *is a basis matrix while* $R_{j,new}$ *is upper triangular.*

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

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

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

4. *Define*

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

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

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

6. *Recall from Algorithm 2 that $\mathcal{A}_{j,k} + \mathcal{H}_{j,k} \overset{EVD}{=} \begin{bmatrix} \hat{P}_{j,new,k} \, \hat{P}_{j,new,k,\perp} \end{bmatrix} \begin{bmatrix} \Lambda_k & 0 \\ 0 & \Lambda_{k,\perp} \end{bmatrix} \begin{bmatrix} \hat{P}'_{j,new,k} \\ \hat{P}'_{j,new,k,\perp} \end{bmatrix}$ is the EVD of $\mathcal{A}_{j,k} + \mathcal{H}_{j,k}$. Here $\hat{P}_{j,new,k}$ is a $n \times c_{j,new}$ basis matrix.*

7. Using the above, $\mathcal{A}_{j,k}+\mathcal{H}_{j,k}$ can be decomposed in two ways as follows.

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

$$= \begin{bmatrix} E_{j,new} \, E_{j,new,\perp} \end{bmatrix} \begin{bmatrix} A_{j,k}+H_{j,k} & B'_{j,k} \\ B_{j,k} & A_{j,k,\perp}+H_{j,k,\perp} \end{bmatrix} \begin{bmatrix} E_{j,new}{}' \\ E_{j,new,\perp}{}' \end{bmatrix}$$

Remark 4.4.7 *Thus, from the above definition,* $\mathcal{H}_{j,k} = \frac{1}{\alpha}[\Phi_0 \sum_t (-L_t e_t' - e_t L_t' + e_t e_t')\Phi_0 + F + F']$ *where* $F := E_{new,\perp} E'_{new,\perp} \Phi_0 \sum_t L_t L_t' \Phi_0 E_{new} E'_{new} = E_{new,\perp} E'_{new,\perp} (D_{*,k-1} a_{t,*})(D_{*,k-1} a_{t,*} + D_{new,k-1} a_{t,new})' E_{new} E'_{new}$. *Since* $\mathbf{E}[a_{t,*} a'_{t,new}] = 0$, $\|\frac{1}{\alpha} F\|_2 \lesssim r^2 \zeta^2 \lambda^+$ *w.h.p.*

Definition 4.4.8 *In the sequel, we let*

1. $r := r_0 + (J-1)c_{\max}$ *and* $c := c_{\max} = \max_j c_{j,new}$,
2. $\kappa_{s,*} := \max_j \kappa_s(P_{j-1})$, $\kappa_{s,new} := \max_j \kappa_s(P_{j,new})$, $\kappa_{s,k} := \max_j \kappa_s(D_{j,new,k})$, $\tilde{\kappa}_{s,k} := \max_j \kappa_s((I - P_{j,new} P_{j,new}{}')\hat{P}_{j,new,k})$, $g_k := \max_j g_{j,k}$,
3. $\kappa^+_{2s,*} := 0.3$, $\kappa^+_{2s,new} := 0.15$, $\kappa^+_s := 0.15$, $\tilde{\kappa}^+_{2s} := 0.15$ *and* $g^+ := \sqrt{2}$ *are the upper bounds assumed in Theorem 4.3.1 on* $\max_j \kappa_{2s}(P_j)$, $\max_j \kappa_{2s}(P_{j,new})$, $\max_j \max_k \kappa_s(D_{j,new,k})$, $\max_j \kappa_{2s}(Q_{j,new,k})$ *and* $\max_j \max_k g_{j,k}$ *respectively,*
4. $\phi^+ := 1.1735$ *is the upper bound on* ϕ_k *that follows using the above bounds (see Fact C.2.1),*
5. $\zeta^+_{j,*} := r_0 \zeta + (j-1)c\zeta$,
6. $\gamma_{new,k} := \min(1.2^{k-1}\gamma_{new}, \gamma_*)$,
7. $P_{j,*} := P_{j-1}$ *and* $\hat{P}_{j,*} := \hat{P}_{j-1}$ *(the point of doing this becomes clear in the next remark).*

Remark 4.4.9 *Notice that the subscript* j *always appears as the first subscript, while* k *is the last one. At many places in this paper, we remove the subscript* j *for simplicity. Whenever there is only one subscript, it refers to the value of* k*, e.g.,* Φ_0 *refers to* $\Phi_{j,0}$*,* $\hat{P}_{new,k}$ *refers to* $\hat{P}_{j,new,k}$*. Also,* $P_* := P_{j-1}$ *and* $\hat{P}_* := \hat{P}_{j-1}$.

4.4.1 Key Lemmas – 1: Bounding the RIC, sparse recovery and LS error and subspace estimation error

At most places in this and the next section, we remove the subscript j for simplicity. Whenever this is done, the convention stated in Remark 4.4.9 applies. Also recall that $P_* := P_{j-1}$ and $\hat{P}_* := \hat{P}_{j-1}$.

We first bound the RIC of Φ_k in terms of the denseness coefficients of P_* and P_{new} and their estimation errors. Next, we use these to bound the sparse recovery and LS error. Finally, we obtain a bound on the subspace estimation error at the k^{th} projection PCA step in terms of the various matrices used in the decomposition of the $\mathcal{A}_k$ and $\mathcal{H}_k$ given in Definition 4.4.6.

4.4.1.1 Bounding the RIC of Φ_k

Lemma 4.4.10 (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. $\kappa_s(\hat{P}_{new,k}) \leq \kappa_{s,new} + \tilde{\kappa}_{s,k}\zeta_k + \zeta_*$
4. $\delta_s(\Phi_0) = \kappa_s^2(\hat{P}_*) \leq \kappa_{s,*}^2 + 2\zeta_*$
5. $\delta_s(\Phi_k) = \kappa_s^2([\hat{P}_* \ \hat{P}_{new,k}]) \leq \kappa_s^2(\hat{P}_*) + \kappa_s^2(\hat{P}_{new,k}) \leq \kappa_{s,*}^2 + 2\zeta_* + (\kappa_{s,new} + \tilde{\kappa}_{s,k}\zeta_k + \zeta_*)^2$ *for* $k \geq 1$

The proof is in Appendix C.1.

4.4.1.2 Bounding the Sparse Recovery and LS Error

Lemma 4.4.11 (Sparse Recovery and LS Error) *Pick ζ as given in Theorem 4.3.1 and let $\zeta_*^+ := (r_0 + (j-1)c)\zeta$. Let ξ_0, ρ be as defined in Theorem 4.3.1. If*

1. *the first three conditions of Theorem 4.3.1 hold,*

2. $\zeta_* \le \zeta_*^+ := (r_0 + (j-1)c)\zeta$ *and*

3. $\zeta_{k-1} \le \zeta_{k-1}^+ \le 0.6^{k-1} + 0.4c\zeta$

then for all $t \in \mathcal{I}_{j,k}$, *for any* $1 \le k \le K+1$,

1. the projection noise β_t *satisfies* $\|\beta_t\|_2 \le \sqrt{c}0.72^{k-1}\gamma_{new} + \sqrt{\zeta}(\sqrt{r} + 0.4\sqrt{c}) \le \xi_0$.

2. the CS error satisfies $\|\hat{S}_{t,cs} - S_t\|_2 \le 7\xi_0$.

3. $\hat{T}_t = T_t$

4. e_t *satisfies*

$$e_t = I_{T_t}[(\Phi_{k-1})_{T_t}{}'(\Phi_{k-1})_{T_t}]^{-1} I_{T_t}{}'[(\Phi_{k-1}P_*)a_{t,*} + D_{new,k-1}a_{t,new}] \tag{4.2}$$

and $\|e_t\|_2 \le 0.18\sqrt{c}0.72^{k-1}\gamma_{new} + 1.2\sqrt{\zeta}(\sqrt{r} + 0.06\sqrt{c})$.

The proof is given in Appendix C.

4.4.1.3 Bounding the subspace estimation error

The following lemma is a consequence of Weyl's theorem (Theorem 2.2.2) and the $\sin\theta$ theorem (Theorem 2.2.1)

Lemma 4.4.12 *If* $\lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2 > 0$, *then*

$$\zeta_k \le \frac{\|\mathcal{R}_k\|_2}{\lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2} \le \frac{\|\mathcal{H}_k\|_2}{\lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2} \tag{4.3}$$

where $\mathcal{R}_k := \mathcal{H}_k E_{new}$ *and* A_k, $A_{k,\perp}$, $\mathcal{H}_k$ *are defined in Definition 4.4.6.*

The proof is given in Appendix C.4.

4.4.2 Key Lemmas – 2: Showing high probability exponential decay of the subspace error

At most places in this section, we remove the subscript j for ease of notation. We retain it where needed, e.g. in defining the r.v. $X_{j,k}$ and in defining and using the set $\Gamma_{j,k}$ or for the time interval $\mathcal{I}_{j,k}$. Also, recall that $P_* := P_{j-1}$ and $\hat{P}_* := \hat{P}_{j-1}$.

In this section, in Lemmas 4.4.14 and 4.4.15, under the assumption that $\zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$ and the four conditions of Theorem 4.3.1 hold, we obtain high probability bounds on each of the terms of (4.3), conditioned on $\Gamma_{j,k-1}^e$. Under the same assumptions, Lemma 4.4.16 combines the result of these two lemmas with (4.3) to obtain a high probability upper bound on ζ_k conditioned on $\Gamma_{j,k-1}^e$. We use this upper bound to define ζ_k^+ in Definition 4.4.17. In Lemma 4.4.18, we show that, under the assumptions of Theorem 4.3.1 this ζ_k^+ indeed satisfies $\zeta_k^+ \leq 0.6^k + 0.4c\zeta$. Lemma 4.4.21 then combines the results of Lemmas 4.4.16 and 4.4.18 to finally conclude that just under the assumptions of Theorem 4.3.1, $\zeta_k \leq 0.6^k + 0.4c\zeta$ w.h.p. This, along with $\zeta_* \leq \zeta_*^+$, implies that the subspace error decays exponentially towards a constant times ζ w.h.p.

4.4.2.1 Obtaining high probability bounds on $\zeta_{j,k}$

Recall that $\kappa_{2s,*}^+ := 0.3$ and $\kappa_{2s,\text{new}}^+ = 0.15$, $\tilde{\kappa}_{2s}^+ = 0.15$, $\kappa_s^+ = 0.15$ and $g^+ = \sqrt{2}$ and $\phi^+ = 1.1735 < 1.2$.

Definition 4.4.13 *Define the following functions (we will see their utility in the lemmas that follow):*

$$\begin{aligned}
C(x;u) &:= (1+\frac{2\kappa_s^+}{\sqrt{1-u^2}})\kappa_s^+\phi^+ x + (1+\frac{\kappa_s^+}{\sqrt{1-u^2}})(\kappa_s^+)^2(\phi^+)^2x^2 \\
O(u,v) &:= \frac{uv}{f}(1+\phi^+ + \frac{2\phi^+}{\sqrt{1-u^2}} + (\phi^+)^2 + \kappa_s^+\frac{\phi^+(1+\phi^+)}{\sqrt{1-u^2}}) \\
g_{inc}(x;u,v,w) &:= C(x;u)g^+ + O(u,v)f + 0.125w \\
g_{dec}(x;u,v,w) &:= 1-u^2-uv-0.125w-g_{inc}(x;u,v,w) \\
f_{inc}(x;u,v,w) &:= \frac{g_{inc}(x;u,v,w)}{g_{dec}(x;u,v,w)}
\end{aligned}$$

As we will see in the lemmas below, $\lambda_{new,k}^- g_{inc}(\zeta_{k-1}^+;\zeta_*^+,\zeta_{j,*}^+ f, c\zeta)$ is a high probability upper bound on $\|\mathcal{H}_k\|_2$, $\lambda_{new,k}^- g_{dec}(\zeta_{k-1}^+;\zeta_*^+,\zeta_{j,*}^+ f, c\zeta)$ is a high probability lower bound for $\lambda_{\min}(A_k) - \lambda_{\max}(A_{k,\perp}) - \|\mathcal{H}_k\|_2$ and $f_{inc}(\zeta_{k-1}^+;\zeta_*^+,\zeta_{j,*}^+ f, c\zeta)$ is a high probability upper bound for ζ_k.

Lemma 4.4.14 *Consider $t \in \mathcal{I}_{j,k}$. Pick ζ as given in Theorem 4.3.1 and let $\zeta_*^+ := (r_0 + (j-1)c)\zeta$. Assume that the four conditions of Theorem 4.3.1 hold. Also, assume that we are given* *a series of constants* ζ_k^+, *with* $\zeta_0^+ = 1$ *and* $\zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$. *Define the random variable*

$$X_{j,k} := [a_1, a_2, \dots a_{t_j+k\alpha-1}].$$

Define the set $\Gamma_{j,k}$ *as follows.*

$$\begin{aligned}\Gamma_{j,k} := \{ & X_{j,k} : \ \zeta_{1,*} \leq r_0\zeta;\ \zeta_{j',k'} \leq \zeta_{k'}^+, \text{ for all } j' = 1,2,\dots j-1, k' = 0,1,\dots K; \\ & \zeta_{j,k'} \leq \zeta_{k'}^+, \text{ for all } k' = 0,1,\dots k \ \} \\ & \cap \{X_{j,k} : \hat{T}_t = T_t \text{ and } e_t \text{ satisfies (4.2) for all } t \leq t_j + k\alpha - 1\}\end{aligned}$$

Recall that $\Gamma_{j,k}^e$ *denotes the event* $X_{j,k} \in \Gamma_{j,k}$. *Assume that* $\mathbf{P}(\Gamma_{j,k-1}^e) > 0$ *for all* $1 \leq k \leq K+1$. *Then,*

1. *for all* $1 \leq k \leq K$, $\mathbf{P}(\lambda_{\min}(A_k) \geq \lambda_{new,k}^-(1 - (\zeta_*^+)^2 - \frac{c\zeta}{12}) | \Gamma_{j,k-1}^e) > 1 - p_{a,k}(\alpha,\zeta)$.
2. *for all* $1 \leq k \leq K$, $\mathbf{P}(\lambda_{\max}(A_{k,\perp}) \leq \lambda_{new,k}^-((\zeta_*^+)^2 f + \frac{c\zeta}{24}) | \Gamma_{j,k-1}^e) > 1 - p_b(\alpha,\zeta)$ *where*

$$\begin{aligned}p_{a,k}(\alpha,\zeta) &:= c\exp\Big(-\frac{\alpha\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot \min(1.2^{4k}\gamma_{new}^4, \gamma_*^4)}\Big) + c\exp\Big(-\frac{\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot 4^2}\Big) \text{ and} \\ p_b(\alpha,\zeta) &:= (n-c)\exp\Big(-\frac{\alpha c^2\zeta(\lambda^-)^2}{8\cdot 24^2}\Big).\end{aligned} \tag{4.4}$$

Lemma 4.4.15 *Under the same settings as Lemma 4.4.14, for all* $k \geq 1$,

1. $\mathbf{P}(\{\hat{T}_t = T_t$ *and* e_t *satisfies (4.2) for all* $t \in \mathcal{I}_{j,k}\} | \Gamma_{j,k-1}^e) = 1$.
2. $\mathbf{P}(\|\mathcal{H}_k\|_2 \leq \lambda_{new,k}^- g_{inc}(\zeta_{k-1}^+; \zeta_*^+, \zeta_*^+ f, c\zeta)\ |\Gamma_{j,k-1}^e) \geq 1 - p_c(\alpha,\zeta)$ *where*

$$\begin{aligned}p_c(\alpha,\zeta) := & \, n\exp\Big(-\frac{\alpha\zeta^2(\lambda^-)^2}{8\cdot 24^2(0.0324\gamma_{new}^2 + 0.0072\gamma_{new} + 0.0004)^2}\Big) \\ & + n\exp\Big(-\frac{\alpha\zeta^2(\lambda^-)^2}{32\cdot 24^2(0.06\gamma_{new}^2 + 0.0006\gamma_{new} + 0.4)^2}\Big) \\ & + n\exp\Big(-\frac{\alpha\zeta^2(\lambda^-)^2\epsilon^2}{32\cdot 24^2(0.186\gamma_{new}^2 + 0.00034\gamma_{new} + 2.3)^2}\Big)\end{aligned}$$

The proofs of Lemma 4.4.14 and Lemma 4.4.15 are in Appendix C.

Lemma 4.4.16 *Under the same settings as in Lemma 4.4.14, for all* $k \geq 1$,

1. *If* $g_{dec}(\zeta_{k-1}^+;\zeta_*^+,\zeta_*^+ f;c\zeta) > 0$, *then* $\mathbf{P}(\zeta_k \le f_{inc}(\zeta_{k-1}^+;\zeta_*^+,\zeta_*^+ f;c\zeta)|\Gamma_{j,k-1}^e) \ge p_k(\alpha,\zeta)$ *where*

$$p_k(\alpha,\zeta) := 1 - p_{a,k}(\alpha,\zeta) - p_b(\alpha,\zeta) - p_c(\alpha,\zeta) \tag{4.5}$$

This lemma is an easy consequence of Lemmas 4.4.12, 4.4.14 and 4.4.15.

Definition 4.4.17 *Define the series* $\{{\zeta_k}^+\}_{k=0,1,2,\cdots}$ *as follows*

$$\zeta_0^+ := 1,\ \zeta_k^+ := f_{inc}({\zeta_{k-1}}^+;\zeta_*^+,\zeta_*^+ f,c\zeta),\ \text{for } k \ge 1. \tag{4.6}$$

Using Definition 4.4.13, an explicit expression for ζ_k^+ *is*

$$\zeta_k^+ = \frac{b + 0.125c\zeta}{1 - (\zeta_*^+)^2 - (\zeta_*^+)^2 f - 0.25c\zeta - b} \text{ where } b := C\kappa_s^+ g^+ \zeta_{k-1}^+ + \tilde{C}(\kappa_s^+)^2 g^+ (\zeta_{k-1}^+)^2 + C' f (\zeta_*^+)^2,$$

$C := (\frac{2\kappa_s^+\phi^+}{\sqrt{1-(\zeta_*^+)^2}} + \phi^+)$, $C' := ((\phi^+)^2 + \frac{2\phi^+}{\sqrt{1-(\zeta_*^+)^2}} + 1 + \phi^+ + \frac{\kappa_s^+\phi^+}{\sqrt{1-(\zeta_*^+)^2}} + \frac{\kappa_s^+(\phi^+)^2}{\sqrt{1-(\zeta_*^+)^2}})$, $\tilde{C} := ((\phi^+)^2 + \frac{\kappa_s^+(\phi^+)^2}{\sqrt{1-(\zeta_*^+)^2}})$,

4.4.2.2 Exponential decay of the bounds on $\zeta_{j,k}$

Lemma 4.4.18 (Exponential decay of ζ_k^+) *Pick ζ as given in Theorem 4.3.1. Assume that the four conditions of Theorem 4.3.1 hold. Define the series ${\zeta_k}^+$ as in Definition 4.4.17. Then,*

1. $\zeta_0^+ = 1$, $\zeta_k^+ \le \zeta_{k-1}^+ \le 0.5985$ *for all* $k \ge 1$.
2. $\zeta_k^+ \le 0.6^k + 0.4c\zeta$ *for all* $k \ge 0$
3. $g_{dec}(\zeta_k^+;\zeta_*^+,\zeta_*^+ f,c\zeta) \ge g_{dec}(0.596;10^{-4},1.5\times 10^{-4},10^{-4}) > 0$ *for all* $k \ge 1$.

The proof is in Sec. C.8.

4.4.2.3 High probability exponential decay of $\zeta_{j,k}$

Definition 4.4.19 *Define the event* $\tilde{\Gamma}_{j,k}^e$ *for* $k = 1,2\dots K+1$ *as*

$$\tilde{\Gamma}_{j,k}^e := \begin{cases} \{\zeta_{j,k} \le \zeta_k^+, \hat{T}_t = T_t \text{ and } e_t \text{ satisfies (4.2) for all } t \in \mathcal{I}_{j,k}\} & \text{if } 1 \le k \le K \\ \{\hat{T}_t = T_t \text{ and } e_t \text{ satisfies (4.2) for all } t \in \mathcal{I}_{j,k}\} & \text{if } k = K+1 \end{cases}$$

Remark 4.4.20 *Recall that the event $\Gamma^e_{j,k}$ is defined in Lemma 4.4.14 as follows.*

$$\Gamma^e_{j,k} := \{\zeta_{1,*} \le r_0\zeta;\ \zeta_{j',k'} \le \zeta^+_{k'},\ \text{for all } j' = 1,2,\dots j-1, k' = 0,1,\dots K;$$
$$\zeta_{j,k'} \le \zeta^+_{k'},\ \text{for all } k' = 0,1,\dots k\} \cap$$
$$\{\hat{T}_t = T_t \text{ and } e_t \text{ satisfies (4.2) for all } t \le t_j + k\alpha - 1\}$$

It is easy to see that $\Gamma^e_{j,k} = \Gamma^e_{j,k-1} \cap \tilde{\Gamma}^e_{j,k}$ for all $1 \le k \le K$ and $\Gamma^e_{j+1,0} = \Gamma^e_{j,K} \cap \tilde{\Gamma}^e_{j,K+1}$. Thus, $\Gamma^e_{j,k} = \Gamma^e_{j,0} \cap \tilde{\Gamma}^e_{j,1} \cdots \cap \tilde{\Gamma}^e_{j,k}$ and $\Gamma^e_{j+1,0} = \Gamma^e_{j,0} \cap (\cap_{k=1}^{K+1} \tilde{\Gamma}^e_{j,k}) = \Gamma^e_{1,0} \cap \cap_{j'=1}^{j} (\cap_{k=1}^{K+1} \tilde{\Gamma}^e_{j',k})$.

Lemma 4.4.21 *Pick ζ as given in Theorem 4.3.1. Let $\zeta^+_{j,*} := (r_0 + (j-1)c)\zeta$ and let ζ^+_k be as defined in Definition 4.4.17. Also, let $p_k(\alpha, \zeta)$ be as defined in Lemma 4.4.16 and let the events $\tilde{\Gamma}^e_{j,k}$ and $\Gamma^e_{j,k}$ be as defined above in Definition 4.4.19 and Remark 4.4.20. Assume that the four conditions of Theorem 4.3.1 hold. Also, assume that $\mathbf{P}(\Gamma^e_{j,k-1}) > 0$ for all $1 \le k \le K+1$. Then,*

1. *$\zeta^+_k \le 0.6^k + 0.4c\zeta$ for all $0 \le k \le K$,*
2. *$\mathbf{P}(\tilde{\Gamma}^e_{j,k} | \Gamma^e_{j,k-1}) \ge p_k(\alpha, \zeta)$ for all $1 \le k \le K$ and*
3. *$\mathbf{P}(\tilde{\Gamma}^e_{j,K+1} | \Gamma^e_{j,K}) = 1$.*

The proof is in Appendix C.

4.4.3 Proof Outline for Theorem 4.3.1

The proof of the theorem is an easy consequence of the following lemmas.

1. In Lemma 4.4.10, we use Lemma 3.3.2 to bound the RIC for the CS measurement matrices, i.e. we bound $\delta_s(\Phi_{j,0})$ and $\delta_s(\Phi_{j,k})$, in terms of the denseness coefficients $\kappa_s(P_{j-1})$ and $\kappa_s(P_{j,\text{new}})$ and the subspace errors $\zeta_{j,*}$ and $\zeta_{j,k}$.

2. Let the bound on $\zeta_{j,*}$ be $\zeta^+_{j,*} = (r_0 + (j-1)c)\zeta$ and that on $\zeta_{j,k-1}$ be ζ^+_{k-1} for all j.

3. In Lemma 4.4.11, assuming that $\zeta_{j,*} \le \zeta^+_{j,*}$, $\zeta_{j,k-1} \le \zeta^+_{k-1}$, $\zeta^+_{k-1} \le 0.6^{k-1} + 0.4c\zeta$ and the first three conditions of the theorem hold, we show the following for all $t \in \mathcal{I}_{j,k}$, $k = 1, \dots (K+1)$.

(a) We bound $\|\beta_t\|_2$ in terms of $\zeta_{j,k-1}$ and $\zeta_{j,*}$.

(b) Next, we show that $\|\beta_t\|_2 \leq \xi$ (with ξ chosen as given in the theorem). We use this, Lemma 4.4.10 and Theorem 2.1.1 (CS result) to bound the CS error $\|\hat{S}_{t,\text{cs}} - S_t\|_2$.

(c) Next, we show that if the support estimation threshold ω is chosen as given in the theorem, then $\hat{T}_t = T_t$.

(d) With $\hat{T}_t = T_t$, we are able to give an exact expression for the LS step error, $e_t := \hat{S}_t - S_t$ and also bound it. Recall that e_t is also equal to $L_t - \hat{L}_t$.

4. In Lemma 4.4.12, we use the $\sin\theta$ theorem and Weyl's theorem (Theorems 2.2.1 and 2.2.2) to bound the subspace error $\zeta_{j,k}$ for projection PCA done at $t = t_j + k\alpha - 1$ in terms of the perturbation matrix, $\mathcal{H}_{j,k}$, and the various components of the decomposition of $\mathcal{A}_{j,k}$ given in Definition 4.4.6.

5. Let $\Gamma^e_{j,k}$ denote the event that (i) $\zeta_{1,*} \leq r_0\zeta$, $\zeta_{j',k'} \leq \zeta^+_{k'}$ for all $1 \leq j' \leq j-1$, $0 \leq k' \leq K$, and $\zeta_{j,k'} \leq \zeta^+_{k'}$, for all $0 \leq k' \leq k$, and (ii) $\hat{T}_t = T_t$ and e_t satisfies (4.2) for all $t \leq t_j + k\alpha - 1$. Under the assumption that $\zeta^+_k \leq 0.6^k + 0.4c\zeta$, with K defined as in the theorem, it is clear that $\zeta_{j',K} \leq \zeta^+_K \leq c\zeta$. Thus, $\Gamma^e_{j,k}$ implies that $\zeta_{j,*} \leq \zeta^+_{j,*} = (r_0 + (j-1)c)\zeta$ (this is easy to see using Remark 4.4.4).

6. In Lemmas 4.4.14 and 4.4.15, under the assumption that $\zeta^+_{k-1} \leq 0.6^{k-1} + 0.4c\zeta$ and the conditions of the theorem hold, we obtain high probability bounds on the various terms in the bound on $\zeta_{j,k}$ from Lemma 4.4.12, conditioned on $\Gamma^e_{j,k-1}$.

(a) These lemmas first use Lemma 4.4.11 to show that $\hat{T}_t = T_t$ and thus e_t has an exact expression given by (4.2) and then apply the matrix Hoeffding inequality (Corollary 2.3.4 or Corollary 2.3.5). Lemma 2.2.4 and Fact C.2.1 are used to obtain the final expressions for the bounds and the probabilities with which they hold.

(b) A by-product is the following conclusion. Conditioned on $\Gamma^e_{j,k-1}$, the event that $\hat{T}_t = T_t$ and e_t satisfies (4.2) for all $t \in \mathcal{I}_{j,k}$ holds with probability one.

7. In Lemma 4.4.16, under the assumption that $\zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$ and the four conditions of the theorem hold, we combine the bound of Lemma 4.4.12 with the bounds on its individual terms from Lemmas 4.4.14 and 4.4.15 to obtain a high probability upper bound on $\zeta_{j,k}$, conditioned on $\Gamma_{j,k-1}^e$. The obtained bound is a function of ζ_{k-1}^+, $\zeta_{j,*}^+$ and of the bounds on $\kappa_s(D_{j,\text{new},k})$ and on $g_{j,k}$. We use this upper bound to define ζ_k^+ in Definition 4.4.17.

8. In Lemma 4.4.18, assuming that the four conditions of the theorem hold, we show that ζ_k^+ as defined in Definition 4.4.17 decreases with k and that it indeed satisfies $\zeta_k^+ \leq 0.6^k + 0.4c\zeta$ for all $k \leq K$.

9. Lemma 4.4.21 combines the results of Lemma 4.4.16 and Lemma 4.4.18. It shows that just under the assumptions of the theorem, given $\Gamma_{j,k-1}^e$, the event that $\zeta_{j,k} \leq \zeta_k^+ \leq 0.6^k + 0.4c\zeta$ and that $\hat{T}_t = T_t$ and e_t satisfies (4.2) for all $t \in \mathcal{I}_{j,k}$ holds with a certain probability that depends on α and ζ.

The proof of the theorem follows easily by applying Lemma 4.4.21 for each j and k and finally using Lemma 4.4.18 and the definition of K. In the end, we use the definition of α_{add} and $\alpha \geq \alpha_{\text{add}}$ to show that the the result holds w.p. at least $1 - n^{-10}$. Thus, for large enough n, the result holds w.h.p.

4.4.4 Proof of Theorem 4.3.1

1. By the assumption that $\|(I - \hat{P}_0\hat{P}_0')P_0\| \leq r_0\zeta$, $\mathbf{P}(\{\zeta_{1,*} \leq \zeta_{1,*}^+\}) = 1$. By Lemma 4.4.11, $\zeta_{1,*} \leq \zeta_{1,*}^+$ implies that $\hat{T}_t = T_t$ for all $t_{\text{train}} \leq t \leq t_1 - 1$. Thus, $\mathbf{P}(\Gamma_{1,0}^e) = 1$.

2. Recall that $\Gamma_{j,k}^e = \Gamma_{j,k-1}^e \cap \tilde{\Gamma}_{j,k}^e$ for all $k \geq 1$ and $\Gamma_{j+1,0}^e = \Gamma_{j,0}^e \cap (\cap_{k=1}^{K+1} \tilde{\Gamma}_{j,k}^e)$. Thus, $\mathbf{P}(\Gamma_{j+1,0}^e) = \mathbf{P}(\Gamma_{j,0}^e) \prod_{k=1}^{K+1} \mathbf{P}(\tilde{\Gamma}_{j,k}^e | \Gamma_{j,0}^e, \tilde{\Gamma}_{j,1}^e, \ldots \tilde{\Gamma}_{j,k-1}^e) = \mathbf{P}(\Gamma_{j,0}^e) \prod_{k=1}^{K+1} \mathbf{P}(\tilde{\Gamma}_{j,k}^e | \Gamma_{j,k-1}^e)$. Thus, $\mathbf{P}(\Gamma_{j+1,0}^e) = \mathbf{P}(\Gamma_{1,0}^e) \prod_{j'=1}^{j} \prod_{k=1}^{K+1} \mathbf{P}(\tilde{\Gamma}_{j',k}^e | \Gamma_{j',k-1}^e)$.

3. Since $\mathbf{P}(\Gamma_{1,0}^e) = 1 > 0$ and $p_k(\alpha, \zeta) > 0$ for all k, we can apply Lemma 4.4.21 for every k and j' starting with $k = 1, j' = 1$. Thus, by Lemma 4.4.21 $\mathbf{P}(\Gamma_{J+1,0}^e) \geq (\prod_{k=1}^{K} p_k(\alpha, \zeta))^J \geq (p_K(\alpha, \zeta))^{KJ}$. The last inequality follows because $p_k \geq p_K$.

4. Now,

   (a) $\Gamma^e_{J+1,0}$ implies that (i) $\hat{T}_t = T_t$ and e_t satisfies (4.2) for all $t < t_{J+1}$; (ii) $\zeta_{j,k} \leq \zeta_k^+$ for all $k \leq K$, $j \leq J$.

   (b) By Lemma 4.4.18, $\zeta_k^+ \leq 0.6^k + 0.4c\zeta$. Thus, $\Gamma^e_{J+1,0}$ implies that $\zeta_{1,*} \leq r_0\zeta$ and $\zeta_{j,k} \leq 0.6^k + 0.4c\zeta$ for all $j \leq J$, $k \leq K$. Using the definition of K, this means that $\zeta_{j,K} \leq c\zeta$ for all j. By Remark 4.4.4, all this implies that for $t \in \mathcal{I}_{j,k}$, $\text{SE}_t \leq \zeta_{j,*} + \zeta_{j,k-1} \leq (r_0+(j-1)c)\zeta + 0.4c\zeta + 0.6^{k-1}$, and for $t \in I_{j,K+1}$, $\text{SE}_t \leq \text{SE}_{j,K} \leq (r_0+jc)\zeta$.

   (c) Combining the previous two conclusions and using Fact C.2.1, $\Gamma^e_{J+1,0}$ implies that the bounds on $\|e_t\|_2$ hold.

5. Since $\mathbf{P}(\Gamma^e_{J+1,0}) \geq (p_K(\alpha,\zeta))^{KJ}$, all of the above hold w.p. at least $(p_K(\alpha,\zeta))^{KJ}$. Using the definition of α_{add}, $(p_K(\alpha,\zeta))^{KJ} \geq 1 - n^{-10}$ whenever $\alpha \geq \alpha_{\text{add}}$. Thus the above conclusions hold w.p. at least $1 - n^{-10}$.

4.5 ReProCS with practical parameters setting

The ReProCS algorithm given in Algorithm 2 uses knowledge of $t_j, r_0, c_{j,\text{new}}$ from the model and it has four parameters ξ, ω, α, K that can be set in terms of the model parameters as given in Theorem 4.3.1. However, it is unreasonable to expect that, in practice, the model parameters are known. We provide here reasonable heuristics for setting both the model and the algorithm parameters automatically.

For a vector v, we define the 99%-energy set of v as $T_{0.99}(v) := \{i : |v_i| \geq v_{0.99}\}$ where the threshold $v_{0.99}$ is the largest value of $|v_i|$ so that $\|v_{T_{0.99}}\|_2^2 \geq 0.99\|v\|_2^2$. It is computed by sorting $|v_i|$ in non-increasing order of magnitude. One keeps adding elements to $T_{0.99}$ until $\|v_{T_{0.99}}\|_2^2 \geq 0.99\|v\|_2^2$.

We pick $\alpha = 100$ arbitrarily. We let $\xi = \xi_t$ and $\omega = \omega_t$ vary with time. Recall that ξ_t is the upper bound on $\|\beta_t\|_2$. We do not know β_t. All we have is an estimate of β_t from $t-1$, $\hat{\beta}_{t-1} = (I - \hat{P}_{(t-1)}\hat{P}_{(t-1)}')\hat{L}_{t-1}$. We used a value a little larger than $\|\hat{\beta}_{t-1}\|_2$ for ξ_t: we let $\xi_t = 2\|\hat{\beta}_{t-1}\|_2$. The parameter ω_t is the support estimation threshold. One reasonable way to pick this is to use a percentage energy threshold of $\hat{S}_{t,\text{cs}}$ [40]. In this work, we used $\omega_t = 0.5(\hat{S}_{t,\text{cs}})_{0.99}$.

Let $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_{t_{\text{train}}}$ denote the eigenvalues of $\frac{1}{t_{\text{train}}} \sum_{t=1}^{t_{\text{train}}} L_t L_t'$. We estimate r_0 and λ^- as

$$\hat{r}_0 = \max_{i=1,2,\cdots,t_{\text{train}}-1} (\frac{\hat{\lambda}_i - \hat{\lambda}_{i+1}}{\hat{\lambda}_i}), \ \hat{\lambda}^- = \hat{\lambda}_{\hat{r}_0} \tag{4.7}$$

This heuristic relies on the fact that the maximum normalized difference between consecutive eigenvalues is from λ^- to zero.

We split projection PCA into two phases: "detect" and "estimate". In the "detect" phase, we estimate the change time t_j and the number of new added directions $c_{j,\text{new}}$ as follows. We keep doing projection PCA every α frames and looking for eigenvalues above $\hat{\lambda}^-$. If there are any eigenvalues above $\hat{\lambda}^-$, we let $\hat{t}_j = t - \alpha + 1$ and we let $\hat{c}_{j,\text{new}}$ be the number of these eigenvalues. Also, we increment j and we reset k to one. At this time, the algorithm enters the "estimate" phase. In this phase, we keep doing projection PCA every α frames until the stopping criterion given in step 3(a)iiB of Algorithm 3 is satisfied (this estimates K). The idea is to stop when k exceeds $K_{\min}$ and $\hat{P}_{j,\text{new},k}' P_{j,\text{new}}$ is approximately equal to $\hat{P}_{j,\text{new},k-1}' P_{j,\text{new}}$ three times in a row; or when $k = K_{\max}$. We pick $K_{\min} = 5, K_{\max} = 20$ arbitrarily. When the stopping criterion is satisfied, we let $K_j = k$ and $\hat{P}_j = [\hat{P}_{j-1}, \hat{P}_{j,\text{new},K_j}]$, and the algorithm enters the "detect" phase.

4.6 Experimental Results

The simulated data is generated as follows.

The measurement matrix $\mathcal{M}_t := [M_1, M_2, \cdots, M_t]$ is of size 2048×5200. 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 $\mathcal{S}_t := [S_1, S_2, \cdots, S_t]$ is generated as follows.

1. For $1 \leq t \leq t_{\text{train}} = 200$, $S_t = 0$.

2. For $t_{\text{train}} < t \leq 5200$, 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 $P'_{1\to 2}1$ with equal probability and the magnitudes are uniformly distributed between 2 and 3. Thus, $S_{\min} = 2$.

The low rank matrix $\mathcal{L}_t := [L_1, L_2, \cdots, L_t]$ 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 $t_2 = 2501$. Let U be an $2048 \times (r_0 + c_{1,\text{new}} + c_{2,\text{new}})$ orthonormalized random Gaussian matrix.

   (a) For $1 \leq t \leq t_1 - 1$, $P_{(t)} = P_0$ has rank r_0 with $P_0 = U_{[1,2,\cdots,r_0]}$.

   (b) For $t_1 \leq t \leq t_2 - 1$, $P_{(t)} = P_1 = [P_0 \ P_{1,\text{new}}]$ has rank $r_1 = r_0 + c_{1,\text{new}}$ with $P_{1,\text{new}} = U_{[r_0+1,\cdots,r_0+c_{1,\text{new}}]}$.

   (c) For $t \geq t_2$, $P_{(t)} = P_2 = [P_1 \ P_{2,\text{new}}]$ has rank $r_2 = r_1 + c_{2,\text{new}}$ with $P_{2,\text{new}} = U_{[r_0+c_{1,\text{new}}+1,\cdots,r_0+c_{1,\text{new}}+c_{2,\text{new}}]}$.

2. a_t is independent over t. The various $(a_t)_i$'s are also mutually independent for different i.

   (a) For $1 \leq t < t_1$, we let $(a_t)_i$ be uniformly distributed between $-\gamma_{i,t}$ and $\gamma_{i,t}$, where

$$\gamma_{i,t} = \begin{cases} 400 & \text{if } i = 1, 2, \cdots, r_0/4, \forall t, \\ 30 & \text{if } i = r_0/4+1, r_0/4+2, \cdots, r_0/2, \forall t. \\ 2 & \text{if } i = r_0/2+1, r_0/2+2, \cdots, 3r_0/4, \forall t. \\ 1 & \text{if } i = 3r_0/4+1, 3r_0/4+2, \cdots, r_0, \forall t. \end{cases} \tag{4.8}$$

   (b) For $t_1 \leq t < t_2$, $a_{t,*}$ 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

$$\gamma_{r_1,t} = \begin{cases} 1.1^{k-1} & \text{if } t_1 + (k-1)\alpha \le t \le t_1 + k\alpha - 1, k = 1,2,3,4, \\ 1.1^{4-1} = 1.331 & \text{if } t \ge t_1 + 4\alpha. \end{cases} \tag{4.9}$$

(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

$$\gamma_{r_2,t} = \begin{cases} 1.1^{k-1} & \text{if } t_2 + (k-1)\alpha \le t \le t_2 + k\alpha - 1, k = 1,2,\cdots,7, \\ 1.1^{7-1} = 1.7716 & \text{if } t \ge t_2 + 7\alpha. \end{cases} \tag{4.10}$$

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.

For Fig. 4.2 and Fig. 4.3, we used $s = 20$, $r_0 = 36$ and $c_{1,\text{new}} = c_{2,\text{new}} = 1$. We let $\Delta = 10$ for Fig. 4.2 and $\Delta = 50$ for Fig. 4.3. 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. 4.2, $\mathcal{S}_t$ has rank $29, 39, 49, 259$ at $t = 300, 400, 500, 2600$; for Fig. 4.3, $\mathcal{S}_t$ has rank $21, 23, 25, 67$ 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 $\frac{\|I_{T_t}{}' D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$ at the projection PCA times. 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 at every $t = t_j + k\alpha - 1$.

We compared against PCP [2]. At every $t = t_j + 4k\alpha$, we solved (1.1) with $\lambda = 1/\sqrt{\max(n,t)}$ 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. In Fig. 4.2, Fig. 4.3 and Fig. 4.4, the times at which PCP is done are marked by red triangles.

As can be seen from Fig. 4.2, the subspace error $\text{SE}_{(t)}$ of ReProCS decreased exponentially and stabilized after about 4 projection PCA update steps. The averaged normalized error for S_t followed a similar trend. ReProCS(practical) performed similar to ReProCS but stabilized in about 6 projection PCA update steps. In Fig. 4.3 where $\Delta = 50$, the subspace error $\text{SE}_{(t)}$ also decreased but the decrease was a bit slower as compared to Fig. 4.2 where $\Delta = 10$. Also, the ratio $\frac{\|{I_{T_t}}' D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$ was now larger. Because of the correlated support change, the error of PCP was larger in both cases. The difference in performance between ReProCS and PCP is larger when $\Delta = 50$.

For Fig. 4.4, we increased s to 100 and we used $\Delta = 10$. A larger s results in a larger $\frac{\|{I_{T_t}}' D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$ (and larger $\kappa_s(D_{j,\text{new},k})$). Thus, the rate of decrease of $\text{SE}_{(t)}$ is smaller than that for the previous two figures. The error of S_t followed a similar trend.

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. For ReProCS, the average error $\frac{1}{5200}\sum_{t=201}^{5200} \frac{\|\hat{S}_t - S_t\|_2}{\|S_t\|_2} = 8.4 \times 10^{-3}$. The error of PCP was also very high: $\frac{1}{5200}\sum_{t=201}^{5200} \frac{\|\hat{S}_t - S_t\|_2}{\|S_t\|_2} = 0.43$.

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.

Algorithm 3 ReProCS(practical)

Input: M_t, *Output:* $\hat{S}_t$, $\hat{L}_t$, $\hat{P}_{(t)}$.

Initialization: Given training sequence $[L_1, L_2, \cdots, L_{t_{train}}]$, compute the EVD of $\frac{1}{t_{\text{train}}} \sum_{t=1}^{t_{\text{train}}} L_t {L_t}' \overset{EVD}{=} E \Lambda E'$ and then estimate $\hat{r}_0$ and $\hat{\lambda}^-$ using (4.7). Let $\hat{P}_0$ retain the eigenvectors with the $\hat{r}_0$ largest eigenvalues.

At $t = t_{\text{train}}$, let $\hat{P}_{(t)} \leftarrow \hat{P}_0$. Let $j \leftarrow 0$, $k \leftarrow 1$, $\hat{t}_j = t_{\text{train}} + 1$ and $flag \leftarrow detect$. For $t > t_{\text{train}}$, do the following:

1. Do step 1) of Algorithm 2 but with ξ and ω replaced by ξ_t and ω_t computed as explained in Sec. 4.5.

2. Do step 2) of Algorithm 2.

3. Projection PCA: Update $\hat{P}_{(t)}$ as follows.

   (a) If $t = \hat{t}_j + k\alpha - 1$, compute EVD of $\frac{1}{\alpha} \sum_{t=\hat{t}_j+(k-1)\alpha}^{\hat{t}_j+k\alpha-1} (I - \hat{P}_{j-1}\hat{P}_{j-1}') \hat{L}_t \hat{L}_t' (I - \hat{P}_{j-1}\hat{P}_{j-1}')$

      i. If $flag = detect$,

         A. If no eigenvalues are above $\hat{\lambda}^-$, then $\hat{P}_{(t)} \leftarrow \hat{P}_{(t-1)}$. Increment $k \leftarrow k + 1$.

         B. If there are eigenvalues above $\hat{\lambda}^-$, then $\hat{t}_j \leftarrow t - \alpha + 1$, $j \leftarrow j + 1$, $k \leftarrow 1$, $flag \leftarrow estimate$.

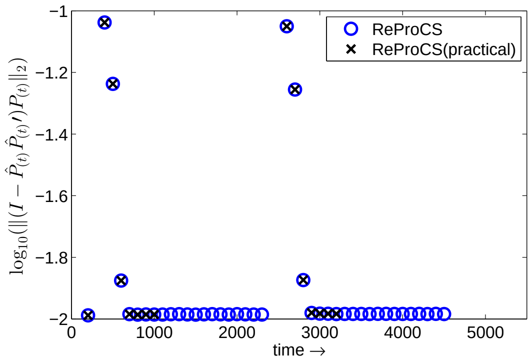
      ii. Else if $flag = estimate$,

         A. Let $\hat{P}_{j,\text{new},k}$ retain the eigenvectors with eigenvalues above $\hat{\lambda}^-$, $\hat{P}_{(t)} \leftarrow [\hat{P}_{j-1} \ \hat{P}_{j,\text{new},k}]$ and $k \leftarrow k + 1$.

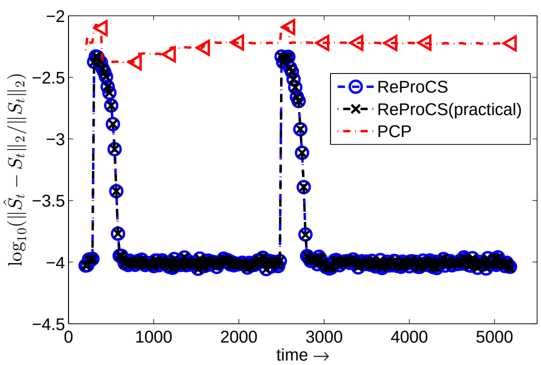
         B. If if $k \geq K_{\min}$ and $\frac{\| \sum_{t-\alpha+1}^{t} (\hat{P}_{j,\text{new},i-1}\hat{P}_{j,\text{new},i-1}' - \hat{P}_{j,\text{new},i}\hat{P}_{j,\text{new},i}') L_t \|_2}{\| \sum_{t-\alpha+1}^{t} \hat{P}_{j,\text{new},i-1}\hat{P}_{j,\text{new},i-1}' L_t \|_2} < 0.01$ for $i = k-2, k-1, k$; or $k = K_{\max}$, then $\hat{K}_j \leftarrow k$, $\hat{P}_j \leftarrow [\hat{P}_{j-1} \ \hat{P}_{j,\text{new},\hat{K}_j}]$ and reset $flag \leftarrow detect$.

   Else ($t \neq \hat{t}_j + k\alpha - 1$) set $\hat{P}_{(t)} \leftarrow \hat{P}_{(t-1)}$.

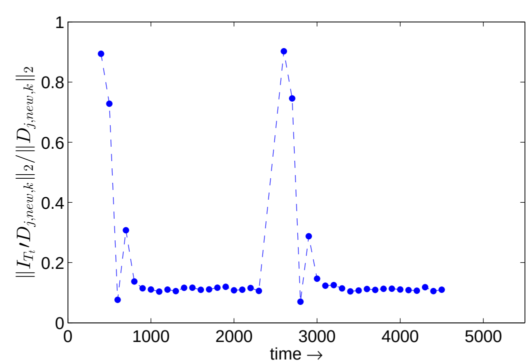
4. Increment $t \leftarrow t + 1$ and go to step 1.



(a) subspace error, $\mathrm{SE}_{(t)}$

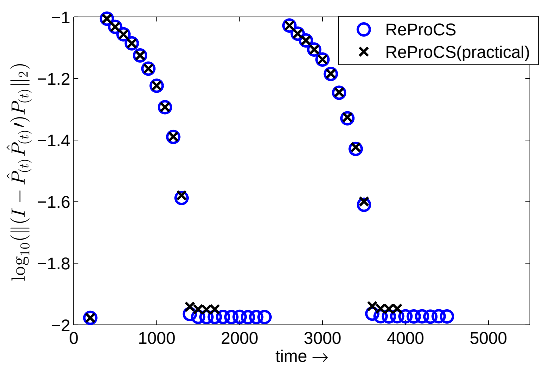


(b) recon error of S_t

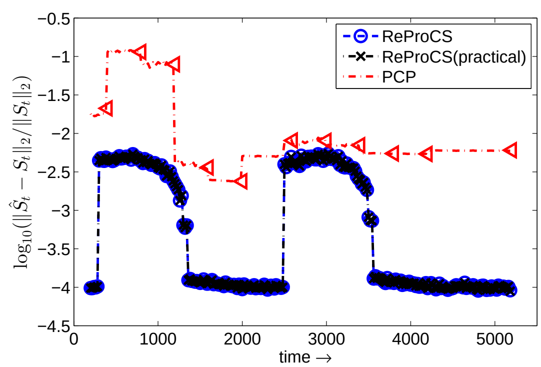


(c) plot of $\frac{\|I_{T_t}{}'D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$

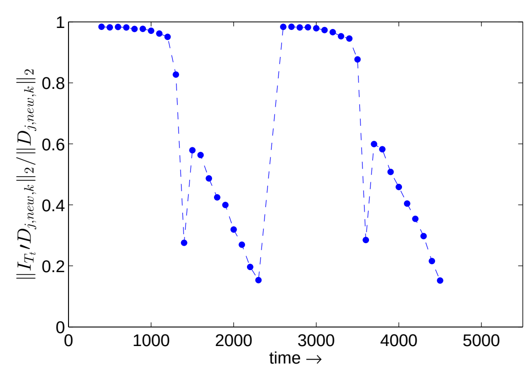
Figure 4.2 ReProCS with $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 10$.



(a) subspace error, $\mathrm{SE}_{(t)}$

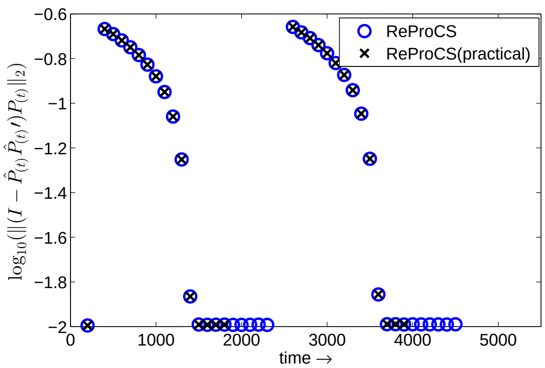


(b) recon error of S_t

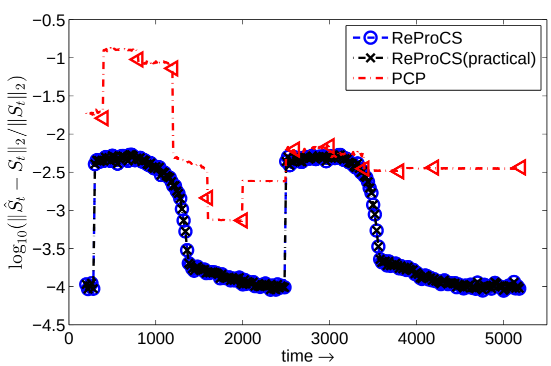


(c) plot of $\frac{\|I_{T_t}{}' D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$

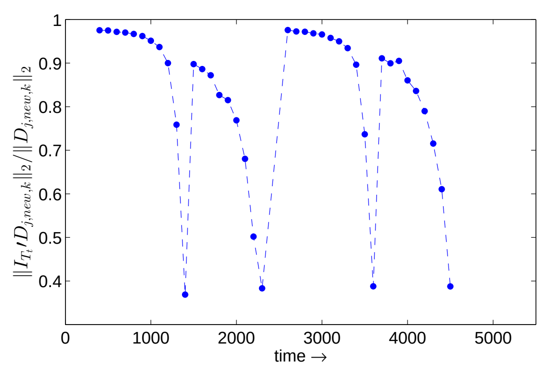
Figure 4.3 ReProCS with $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 50$.



(a) subspace error, $\mathrm{SE}_{(t)}$



(b) recon error of S_t



(c) plot of $\frac{\|I_{T_t}{}'D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$

Figure 4.4 ReProCS with $r_0 = 36$, $s = \max_t |T_t| = 100$ and $\Delta = 10$.

CHAPTER 5. ReProCS with cluster-PCA (ReProCS-cPCA) an its performance Guarantee

ReProCS-cPCA needs an extra assumption that the eigenvalues of the covariance matrix of L_t are sufficiently clustered as explained in Sec. 5.1. We develop the ReProCS-cPCA algorithm in Sec 5.2. We summarize the ReProCS-cPCA algorithm in Algorithm 4. We give the performance guarantees (Theorem 5.3.1) in Sec 5.3. Here we also provide a discussion of the result and the assumptions it makes. The proof of Theorem 5.3.1 is given in Sec 5.4. The key lemmas needed for it are given and proved in Appendix D.2. In Sec 5.5, we show numerical experiments demonstrating Theorem 5.3.1, as well as comparisons with ReProCS and PCP. Parts of this chapter are taken verbatim from [33] [34].

5.1 Clustering assumption

For positive integers K and α, let $\tilde{t}_j := t_j + K\alpha$. 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 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.

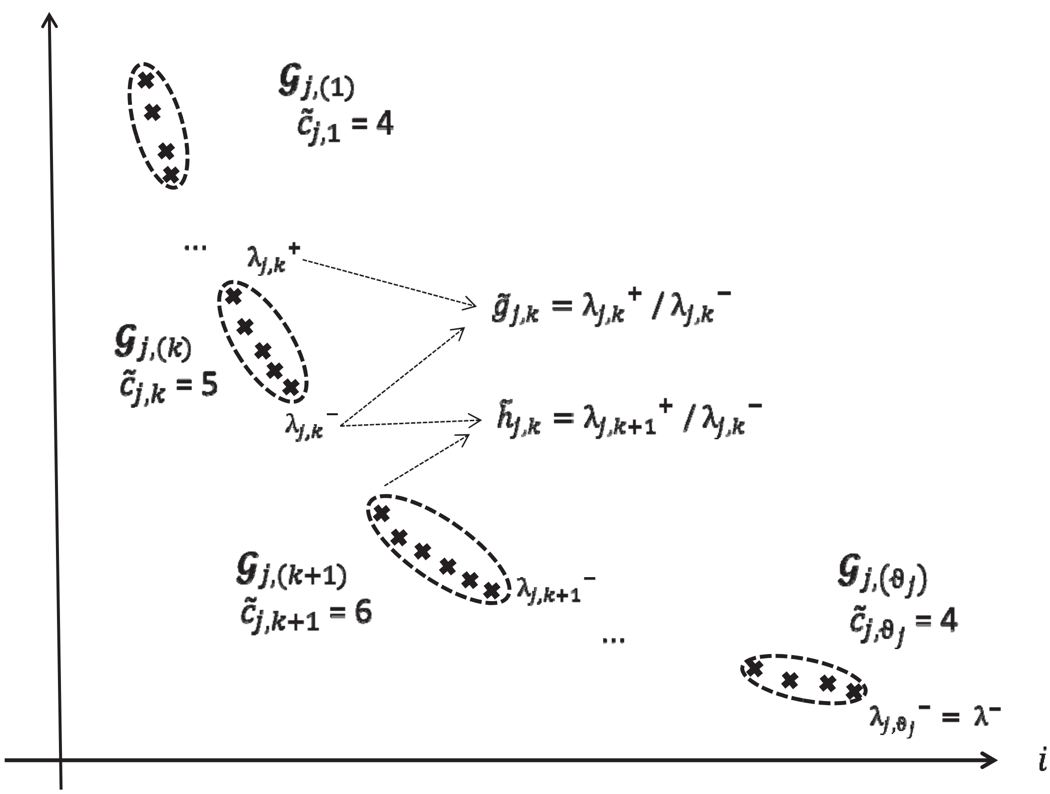


Figure 5.1 Illustration of the clustering assumption (assume $\Lambda_t = \Lambda_{\tilde{t}_j}$).

Assumption 5.1.1 *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* $i = 1, 2, \cdots, r_j$, $\min_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_i(\Lambda_t) \geq \max_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_{i+1}(\Lambda_t)$.

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 cluster contains the largest set of eigenvalues, the second one the next smallest set and so on (see Fig 5). Let*

   (a) $G_{j,k} := (P_j)_{\mathcal{G}_{j,(k)}}$ *be the corresponding cluster of eigenvectors, then* $P_j = [G_{j,1}, G_{j,2}, \cdots, G_{j,\vartheta_j}]$;

   (b) $\tilde{c}_{j,k} := |\mathcal{G}_{j,(k)}|$ *be the number of elements in* $\mathcal{G}_{j,(k)}$, *then* $\sum_{k=1}^{\vartheta_j} \tilde{c}_{j,k} = r_j$;

   (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* $\tilde{g}_{j,k} \geq 1$*)*;

   (e) $\tilde{h}_{j,k} := {\lambda_{j,k+1}}^+ / {\lambda_{j,k}}^-$ *(notice that* $\tilde{h}_{j,k} < 1$*)*;

   (f) $\tilde{g}_{\max} := \max_j \max_{k=1,2,\cdots,\vartheta_j} \tilde{g}_{j,k}$, $\tilde{h}_{\max} := \max_j \max_{k=1,2,\cdots,\vartheta_j} \tilde{h}_{j,k}$, $\tilde{c}_{\min} := \min_j \min_{k=1,2,\cdots,\vartheta_j} \tilde{c}_{j,k}$

   (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 5.3.1.*

Remark 5.1.2 *The assumption above can, in fact, be relaxed to only require the following. The matrices* Λ_t *are such that there exists a partition,* $\mathcal{G}_{j,(1)}, \mathcal{G}_{j,(2)}, \cdots, \mathcal{G}_{j,(\vartheta_j)}$, *of the index set* $\{1, 2, \ldots r_j\}$ *so that* $\min_{i \in \mathcal{G}_{j,(k)}} \min_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_i(\Lambda_t) > \max_{i \in \mathcal{G}_{j,(k+1)}} \max_{t \in [\tilde{t}_j, t_{j+1}-1]} \lambda_i(\Lambda_t)$. *Define all quantities as above. We assume that* $\tilde{g}_{\max}$ *and* $\tilde{h}_{\max}$ *are small enough.*

5.2 The ReProCS-cPCA algorithm

ReProCS-cPCA is summarized in Algorithm 4. It uses the following definition.

Definition 5.2.1 *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. $\tilde{\mathcal{I}}_{j,\vartheta_j+1} := [\tilde{t}_j + \vartheta_j\tilde{\alpha}, t_{j+1} - 1]$.

Notice that $[t_j, t_{j+1} - 1] = (\cup_{k=1}^{K} \mathcal{I}_{j,k}) \cup (\cup_{k=1}^{\vartheta_j} \tilde{\mathcal{I}}_{j,k}) \cup \tilde{\mathcal{I}}_{j,\vartheta_j+1}$. *Also, K, α and $\tilde{\alpha}$ are parameters given in Algorithm 4.*

ReProCS-cPCA proceeds as follows. The algorithms begins with the knowledge of $\hat{P}_0$ and initializes $\hat{P}_{(t_{\text{train}})} \leftarrow \hat{P}_0$. $\hat{P}_0$ can be computed as the top r_0 left singular vectors of $\mathcal{M}_{t_{\text{train}}}$ (since, by assumption, $\mathcal{S}_{t_{\text{train}}}$ is either zero or very small). For $t > t_{\text{train}}$, the following is done. Step 1 projects M_t perpendicular to $\hat{P}_{(t-1)}$, solves the ℓ_1 minimization problem, followed by support recovery and finally computes a least squares (LS) estimate of S_t on its estimated support. This final estimate $\hat{S}_t$ is used to estimate L_t as $\hat{L}_t = M_t - \hat{S}_t$ in step 2. The sparse recovery error, $e_t := \hat{S}_t - S_t$. Since $\hat{L}_t = M_t - \hat{S}_t$, e_t also satisfies $e_t = L_t - \hat{L}_t$. Thus, a small e_t (accurate recovery of S_t) means that L_t is also recovered accurately. Step 3a is used at times when no subspace update is done. In step 3b, the estimated $\hat{L}_t$'s are used to obtain improved estimates of $\text{span}(P_{j,\text{new}})$ every α frames for a total of $K\alpha$ frames using the proj-PCA procedure given in Algorithm 1. Within K proj-PCA updates (K chosen as given in Theorem 5.3.1), it can be shown that both $\|e_t\|_2$ and the subspace error, $\text{SE}_{(t)} := \|(I - \hat{P}_{(t)}\hat{P}'_{(t)})P_{(t)}\|_2$, drop down to a constant times ζ. In particular, if at $t = t_j - 1$, $\text{SE}_{(t)} \le r\zeta$, then at $t = \tilde{t}_j := t_j + K\alpha$, we can show that $\text{SE}_{(t)} \le (r + c_{\max})\zeta$. Here $r := r_{\max} = r_0 + c_{\max}$.

To bring $\text{SE}_{(t)}$ down to $r\zeta$ before t_{j+1}, we need a step so that by $t = t_{j+1} - 1$ we have an estimate of only $\text{span}(P_j)$, i.e. we have "deleted" $\text{span}(P_{j,\text{old}})$. One simple way to do this is by standard PCA: at $t = \tilde{t}_j + \tilde{\alpha} - 1$, compute $\hat{P}_j \leftarrow \text{proj-PCA}([\hat{L}_t; t \in \tilde{\mathcal{I}}_{j,1}], [.], r_j)$ and let $\hat{P}_{(t)} \leftarrow \hat{P}_j$. Using the $\sin\theta$ theorem and the Hoeffding corollaries, it can be shown that, as long as f is small enough, doing this is guaranteed to give an accurate estimate of $\text{span}(P_j)$. However f being small is not compatible with the slow subspace change assumption. Notice from Sec 3 that $\lambda^- \leq \gamma_{\text{new}}$ and $\mathbf{E}[||L_t||_2^2] \leq r\lambda^+$. Slow subspace change implies that γ_{new} is small. Thus, λ^- is small. However, to allow L_t to have large magnitude, λ^+ needs to be large. Thus, $f = \lambda^+/\lambda^-$ cannot be small unless we require that L_t has small magnitude for all times t.

In step 3c, we introduce a generalization of the above strategy called cluster-PCA, that removes the bound on f, but instead only requires that the eigenvalues of $\text{Cov}(L_t)$ be sufficiently clustered as explained in Sec 5.1. The main idea is to recover one cluster of entries of P_j at a time. In the k^{th} iteration, we apply proj-PCA on $[\hat{L}_t; t \in \tilde{I}_{j,k}]$ with $P \leftarrow [\hat{G}_{j,1}, \hat{G}_{j,2}, \dots \hat{G}_{j,k-1}])$ to estimate $\text{span}(G_{j,k})$. The first iteration uses $P \leftarrow [.]$, i.e. it computes standard PCA to estimate $\text{span}(G_{j,1})$. By modifying the approach used for ReProCS for analyzing the addition step, we can show that since $\tilde{g}_{j,k}$ and $\tilde{h}_{j,k}$ are small enough (by Assumption 5.1.1), $\text{span}(G_{j,k})$ 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 $\hat{P}_j \leftarrow [\hat{G}_{j,1}, \hat{G}_{j,2} \dots \hat{G}_{j,\vartheta_j}]$ and $\hat{P}_{(t)} \leftarrow \hat{P}_j$. All of this is done at $t = \tilde{t}_j + \vartheta_j\tilde{\alpha} - 1$. Thus, at this time, $\text{SE}_{(t)} = \|(I - \hat{P}_j\hat{P}_j')P_j\|_2 \leq \sum_{k=1}^{\vartheta_j} \|(I - \sum_{i=1}^{k} \hat{G}_{j,i}\hat{G}'_{j,i})G_{j,k}\|_2 \leq \sum_{k=1}^{\vartheta_j} \tilde{c}_{j,k}\zeta = r_j\zeta \leq r\zeta$. 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. 5.2.

The ReProCS-cPCA algorithm has parameters ξ, ω, α, $\tilde{\alpha}$, K and it uses knowledge of model parameters t_j, r_0, $c_{j,\text{new}}$, ϑ_j and $\tilde{c}_{j,i}$. If the model is known the algorithm parameters can be set as in Theorem 5.3.1. In practice, typically the model is unknown. In this case, the parameters t_j, r_0, $c_{j,\text{new}}$, ξ, ω, K can be set as explained for the ReProCS algorithm. The parameters ϑ_j and $\tilde{c}_{j,i}$ for $i = 1, 2 \dots \vartheta_j$, can be set by computing the eigenvalues of $\frac{1}{\tilde{\alpha}} \sum_{t \in \tilde{I}_{j,1}} \hat{L}_t\hat{L}_t'$ and clustering them using any standard clustering algorithm, e.g. k-means clustering or split-and-merge[1]. We pick α and $\tilde{\alpha}$ somewhat arbitrarily. A thumb rule is that α

[1] One simple split-and-merge approach is as follows. Start with a single cluster. Split into two clusters: select the split so that $\tilde{g}_{\max}$ is minimized. Split each of these clusters into two parts again while ensuring $\tilde{g}_{\max}$ is

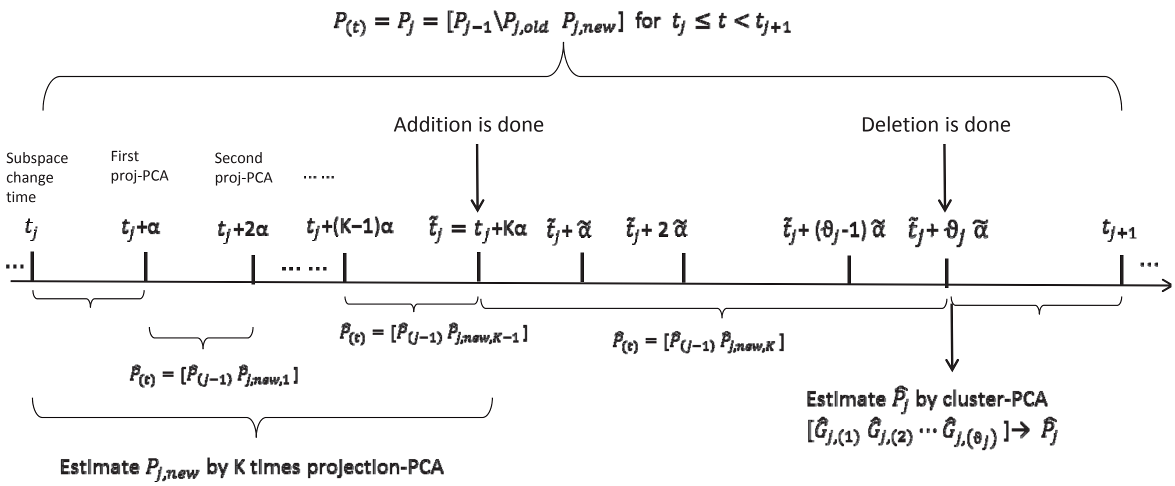


Figure 5.2 A diagram illustrating subspace estimation by ReProCS-cPCA

and $\tilde{\alpha}$ need to be at least five to ten times $c_{\max}$ and $\max_j \max_{i=1,2...\vartheta_j} \tilde{c}_{j,i}$ respectively. From simulation experiments, the algorithm is not very sensitive to the specific choice.

As explained in Sec. 4.2, the reason we use proj-PCA instead of standard PCA is because $e_t = \hat{L}_t - L_t$ is correlated with L_t.

5.3 Performance Guarantees

We state the main result first and then discuss it. We give its corollary for the case where f is small in Corollary 5.3.2. The proof is given in Sec 5.4.

Theorem 5.3.1 *Consider Algorithm 4. Let $c := c_{\max}$ and $r := r_0 + c$. Assume that L_t obeys the model given in Assumption 3.1.1. Also, assume that the initial subspace estimate is accurate enough, i.e. $\|(I - \hat{P}_0 \hat{P}_0')P_0\| \leq r_0 \zeta$, for a ζ that satisfies*

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

minimized. Keep doing this for d_1 steps. Notice that, with every splitting, $\tilde{g}_{\max}$ will either remain the same or reduce, however $\tilde{h}_{\max}$ will either remain same or increase. Then, do a set of merge steps: in each step find the pair of consecutive clusters to merge that will minimize $\tilde{h}_{\max}$.

Let $\xi_0(\zeta), \rho, K(\zeta), \alpha_{add}(\zeta), \alpha_{del}(\zeta),\ g_{j,k}$ be as defined in Definition 5.4.2. If the following conditions hold:

1. (algorithm parameters) $\xi = \xi_0(\zeta),\ 7\rho\xi \leq \omega \leq S_{\min} - 7\rho\xi,\ K = K(\zeta),\ \alpha \geq \alpha_{add}(\zeta),\ \tilde{\alpha} \geq \alpha_{del}(\zeta)$,

2. (denseness)

$$\max_j \kappa_{2s}(P_{j-1}) \leq \kappa^+_{2s,*} = 0.3,\ \max_j \kappa_{2s}(P_{j,new}) \leq \kappa^+_{2s,new} = 0.15,$$
$$\max_j \max_{0\leq k\leq K} \kappa_{2s}(D_{j,new,k}) \leq \kappa^+_s = 0.15,\ \max_j \max_{0\leq k\leq K} \kappa_{2s}(Q_{j,new,k}) \leq \tilde{\kappa}^+_{2s} = 0.15,$$
$$\max_j \kappa_s((I - \hat{P}_{j-1}\hat{P}'_{j-1} - \hat{P}_{j,new,K}\hat{P}'_{j,new,K})P_j) \leq \kappa^+_{s,e}$$

where $D_{j,new,k} := (I - \hat{P}_{j-1}\hat{P}'_{j-1} - \hat{P}_{j,new,k}\hat{P}'_{j,new,k})P_{j,new}$, *and* $Q_{j,new,k} := (I - P_{j,new}P_{j,new}{}')\hat{P}_{j,new,k}$ *and* $\hat{P}_{j,new,0} = [.]$,

3. (slow subspace change)

$$\max_j (t_{j+1} - t_j) > K\alpha + \vartheta_{\max}\tilde{\alpha},$$
$$\max_j \max_{t\in\mathcal{I}_{j,k}} \|a_{t,new}\|_\infty \leq \gamma_{new,k} := \min(1.2^{k-1}\gamma_{new}, \gamma_*),\ \textit{for all } k = 1, 2, \ldots K,$$
$$14\rho\xi_0(\zeta) \leq S_{\min},$$

4. (small average condition number of $Cov(a_{t,\text{new}})$) $g_{j,k} \leq g^+ := \sqrt{2}$,

5. (clustered eigenvalues) *Assumption 5.1.1 holds with* $\tilde{g}_{\max}, \tilde{h}_{\max}, \tilde{c}_{\min}$ *satisfying* $f_{dec}(\tilde{g}_{\max}, \tilde{h}_{\max}) - \frac{f_{inc}(\tilde{g}_{\max}, \tilde{h}_{\max})}{\tilde{c}_{\min}\zeta} > 0$ *where* $f_{dec}(\tilde{g}_{\max}, \tilde{h}_{\max})$ *and* $f_{inc}(\tilde{g}_{\max}, \tilde{h}_{\max})$ *are defined in Definition 5.4.3 (also see Remark D.2.5 which weakens this requirement),*

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

1. $\hat{T}_t = T_t$ *and* $\|e_t\|_2 = \|L_t - \hat{L}_t\|_2 = \|\hat{S}_t - S_t\|_2 \leq 0.18\sqrt{c}\gamma_{new} + 1.24\sqrt{\zeta}$.

2. the subspace error, $SE_{(t)}$ satisfies

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

$$\leq \begin{cases} 0.6^{k-1} + 10^{-2}\sqrt{\zeta} & \text{if } t \in \mathcal{I}_{j,k},\ k = 1, 2, \cdots, K \\ 10^{-2}\sqrt{\zeta} & \text{if } t \in (\cup_{k=1}^{\vartheta_j} \tilde{\mathcal{I}}_{j,k}) \cup \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases}$$

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

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

$$\leq \begin{cases} 0.18 \cdot 0.72^{k-1}\sqrt{c}\gamma_{new} + 1.17 \cdot 1.06\sqrt{\zeta} & \text{if } t \in \mathcal{I}_{j,k},\ k = 1, 2, \cdots, K \\ 1.17\sqrt{\zeta} & \text{if } t \in (\cup_{k=1}^{\vartheta_j} \tilde{\mathcal{I}}_{j,k}) \cup \tilde{\mathcal{I}}_{j,\vartheta_j+1} \end{cases}$$

The above result says the following. Assume that the initial subspace error is small enough. If the assumptions given in the theorem hold, then, w.h.p., we will get exact support recovery at all times. Moreover, the sparse recovery error (and the error in recovering L_t) will always be bounded by $0.18\sqrt{c}\gamma_{\text{new}}$ plus a constant times $\sqrt{\zeta}$. Since ζ is very small, $\gamma_{\text{new}} \ll S_{\min}$, and c is also small, the normalized reconstruction error for S_t will be small at all times, thus making this a meaningful result. In the second conclusion, we bound the subspace estimation error, $\text{SE}_{(t)}$. When a subspace change occurs, this error is initially bounded by one. The above result shows that, w.h.p., with each adddition proj-PCA step, this error decays roughly exponentially and falls below $(r+c)\zeta$ within K steps. After the cluster-PCA step, this error falls below $r\zeta$. By assumption, this occurs before the next subspace change time. Because of the choice of ζ, both $(r+c)\zeta$ and $r\zeta$ are below $0.01\sqrt{\zeta}$. The third conclusion shows that the sparse recovery error as well as the error in recovering L_t decay in a similar fashion.

Notice from Definition 5.4.2 that $K = K(\zeta)$ is larger if ζ is smaller. Also, both $\alpha_{\text{add}}(\zeta)$ and $\alpha_{\text{del}}(\zeta)$ are inversely 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. Because of slow subspace change, this means that we also require a larger delay between subspace change times, i.e. larger $t_{j+1} - t_j$.

The ReProCS algorithm is Algorithm 4 with step 3c removed and replaced by $\hat{P}_j \leftarrow [\hat{P}_{j-1}, \hat{P}_{j,\text{new},K}]$. Let us compare the above result with that for ReProCS for the subspace change model of Assumption 3.1.1. First, 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, ReProCS requires ζ to satisfy $\zeta \leq \min(\frac{10^{-4}}{(r_0+(J-1)c)^2}, \frac{1.5\times 10^{-4}}{(r_0+(J-1)c)^2 f}, \frac{1}{(r_0+(J-1)c)^3 \gamma_*^2})$ whereas in case of ReProCS-cPCA the denominators in the bound on ζ only contain $r + c = r_0 + 2c$ (instead of $r_0 + (J-1)c$).

Because of the above, in Theorem 5.3.1 for ReProCS-cPCA, the only place where J (the number of subspace change times) appears is in the definitions of α_{add} and α_{del}. Notice that α_{add} and α_{del} govern the delay between subspace change times, $t_{j+1} - t_j$. Thus, with ReProCS-cPCA, J can keep increasing, as long as $t_{j+1} - t_j$ also increases accordingly. Moreover, notice that the dependence of α_{add} and α_{del} on J is only logarithmic and thus $t_{j+1} - t_j$ needs to only increase in proportion to $\log J$. On the other hand, for ReProCS, J appears in the denseness assumption, in the bound on ζ and in the definition of α_{add}. Thus, ReProCS needs a bound on J that is indirectly imposed by the denseness assumption.

The main extra assumptions that ReProCS-cPCA needs are (i) the clustering assumption (Assumption 5.1.1 with $\tilde{h}_{\max}, \tilde{g}_{\max}$ being small enough to satisfying $f_{dec}(\tilde{g}_{\max}, \tilde{h}_{\max}) - \frac{f_{inc}(\tilde{g}_{\max}, \tilde{h}_{\max})}{\tilde{c}_{\min}\zeta} > 0$; and (ii) $\max_j \kappa_s((I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,\text{new},K}\hat{P}_{j,\text{new},K}')P_j) < \kappa_{s,e}^+$. The second assumption is similar to the denseness assumption on $D_{j,\text{new},k}$ which is required by both ReProCS and ReProCS-cPCA. The clustering assumption is a practically valid one. We verified it for a video of moving lake waters (see Sec. 3.4) as follows. We first "low-rankified" it to 90% energy as explained in Sec. 3.4. 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 eigenvalues are sufficiently clustered. When this was done, we observed that the clustering assumption holds with $\tilde{g}_{\max} = 7.2$ and $\tilde{h}_{\max} = 0.34$.

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

Consider the denseness assumptions. Let $\mathcal{L}_t = U\Sigma V'$ be its SVD. Then, for $t \in [t_j, t_{j+1}-1]$, $U = [P_0, P_{1,\text{new}}, P_{2,\text{new}}, \dots P_{j,\text{new}}]$ and $V = [a_1, a_2 \dots a_t]'\Sigma^{-1}$. The result for PCP [2] assumes denseness of U and of V: it requires $\kappa_1(U) \le \sqrt{\mu r/n}$ and $\kappa_1(V) \le \sqrt{\mu r/n}$ for a constant $\mu \ge 1$. Moreover, it also requires $\|UV'\|_{\max} \le \sqrt{\mu r}/n$. On the other hand, ReProCS-cPCA only requires $\kappa_{2s}(P_j) \le 0.3$ and $\kappa_{2s}(P_{j,\text{new}}) \le 0.15$. It does not need denseness of the entire U; it does not assume anything about denseness of V; and it does not need a bound on $\|UV'\|_{\max}$.

Another difference is that the result for PCP assumes that any element of the $n \times t$ matrix $\mathcal{S}_t$ is nonzero w.p. ϱ, and zero w.p. $1-\varrho$, independent of all others (in particular, this means that the support sets of the different S_t's are independent over time). Our result for ReProCS-cPCA does not put any such assumption. However it does require denseness of the matrix $D_{j,\text{new},k}$ whose columns span the unestimated part of span$(P_{j,\text{new}})$ for $t \in \mathcal{I}_{j,k+1}$. As demonstrated in Sec. 5.5, this reduces ($\kappa_s(D_{j,\text{new},k})$ increases) if the support sets of S_t's change very little over time. However, as long as, for most k, $\kappa_s(D_{j,\text{new},k})$ is anything smaller than one, which happens as long as there is at least one support change during $\mathcal{I}_{j,k}$, the subspace error does decay down to a small enough value within a finite number of steps. The number of steps required for this increases as $\kappa_s(D_{j,\text{new},k})$ increases. Since $\kappa_s(D_{j,\text{new},k})$ cannot be computed in polynomial time, for the above discussion, we computed $\|{I_{T_t}}'D_{j,\text{new},k}\|_2/\|D_{j,\text{new},k}\|_2$ at $t = t_j + k\alpha - 1$ for $k = 0, 1, \dots K$. In fact, our proof also only needs a bound on this latter quantity.

Also, some additional assumptions that ReProCS-cPCA needs are (a) accurate knowledge of the initial subspace and slow subspace change; (b) denseness of $Q_{j,\text{new},k}$; (c) the independence of a_t's over time; (d) condition number of the average covariance matrix of $a_{t,\text{new}}$ is not too large; and (e) the clustering assumption. Assumptions (a), (b), (c) are discussed in detail in Sec. 4.2 and (a) is also verified for real data. As explained in Sec. 4.3, (c) can possibly be replaced by a weaker random walk model assumption on a_t's if we use the matrix Azuma inequality [25] instead of matrix Hoeffding. Assumption (e) is discussed above. (d) is also an assumption made for simplicity. It can be removed if a clustering assumption similar to Assumption 5.1.1 holds for $(\Lambda_t)_{\text{new}} = \text{Cov}(a_{t,\text{new}})$ during $t \in [t_j, \tilde{t}_j - 1]$ and we use an approach similar to cluster-PCA. If there are $\vartheta_{\text{new},j}$ clusters, we will need $\vartheta_{\text{new},j}$ proj-PCA steps to estimate $\hat{P}_{\text{new},k}$ (instead of the current one step). At the l^{th} step, we use proj-PCA with P being $\hat{P}_{j-1}$ concatenated with the basis matrix estimates for the last $l-1$ clusters to recover the l^{th} cluster.

If in a problem, L_t has small magnitude for all times t, then f, which is the maximum condition number of $\text{Cov}(L_t)$ for any t, can be small. If this is the case, then the clustering assumption 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, $\vartheta_{\max} = 1$. In this case, the following corollary holds.

Corollary 5.3.2 *Assume that the initial subspace estimate is accurate enough as given in Theorem 5.3.1 with ζ as chosen there. Also assume that the first four conditions of Theorem 5.3.1 hold. Then, if f is small enough so that $f_{inc}(f, 0) \leq f_{dec}(f, 0)\tilde{c}_{\min}\zeta$, then, all conclusions of Theorem 5.3.1 hold.*

Notice that the above corollary does not need Assumption 5.1.1 to hold.

5.4 Proof of Theorem 5.3.1

We first define all the quantities that are needed for the proof. The proof outline is given in Sec 5.4.1.

Certain quantities are defined earlier in Assumptions 3.1.1 and 5.1.1, in Definitions 3.1.2 and 5.2.1, in Algorithm 4 and in Theorem 5.3.1.

Definition 5.4.1 *In the sequel, we let*

1. $c := c_{\max}$ and $r := r_{\max} = r_0 + c$ and so $r_j = r_0 + \sum_{i=1}^{j}(c_{i,new} - c_{i,old}) \leq r$,

2. $\phi^+ := 1.1735$

Definition 5.4.2 *We define here the parameters used in Theorem 5.3.1.*

1. *Define* $K(\zeta) := \left\lceil \frac{\log(0.6c\zeta)}{\log 0.6} \right\rceil$

2. *Define* $\xi_0(\zeta) := \sqrt{c}\gamma_{new} + 1.06\sqrt{\zeta}$

3. *Define* $\rho := \max_t\{\kappa_1(\hat{S}_{t,cs} - S_t)\}$. *Notice that* $\rho \leq 1$.

4. *Define the condition number of the average of* $Cov(a_{t,new})$ *over* $t \in \mathcal{I}_{j,k}$ *as*

$$g_{j,k} := \frac{{\lambda_{j,new,k}}^+}{{\lambda_{j,new,k}}^-} \text{ where}$$
$${\lambda_{j,new,k}}^+ := \lambda_{\max}\Big(\frac{1}{\alpha}\sum_{t \in \mathcal{I}_{j,k}} (\Lambda_t)_{new}\Big), \quad {\lambda_{j,new,k}}^- := \lambda_{\min}\Big(\frac{1}{\alpha}\sum_{t \in \mathcal{I}_{j,k}} (\Lambda_t)_{new}\Big),$$

5. *Let* $K = K(\zeta)$. *We define* $\alpha_{add}(\zeta)$ *as in Definition 5.4.2 the smallest value of* α *so that* $(p_K(\alpha,\zeta))^{KJ} \geq 1 - n^{-10}$, *where* $p_K(\alpha,\zeta)$ *is defined in Lemma D.1.3. An explicit value for it is*

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

In words, α_{add} *is the smallest value of the number of data points,* α, *needed for an addition proj-PCA step to ensure that Theorem 5.3.1 holds w.p. at least* $(1 - 2n^{-10})$.

6. *We define* $\alpha_{del}(\zeta)$ *as the smallest value of* α *so that* $\tilde{p}(\tilde{\alpha},\zeta)^{\vartheta_{\max}J} \geq 1 - n^{-10}$ *where* $\tilde{p}(\tilde{\alpha},\zeta)$ *is defined in Lemma D.2.8. We can compute an explicit value for it by using the fact that for any* $x \leq 1$ *and* $r \geq 1$, $(1-x)^r \geq 1 - rx$ *and that* $\sum_{i=1}^{6} e^{-\frac{\alpha}{d_i^2}} \leq 6e^{-\frac{\alpha}{\max_{i=1,2...6} d_i^2}}$. *We get*

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

where $b_7 := (\sqrt{r}\gamma_* + \phi^+\sqrt{\zeta})^2$ *and* $\phi^+ = 1.1732$. *In words,* α_{del} *is the smallest value of the number of data points,* $\tilde{\alpha}$, *needed for a deletion proj-PCA step to ensure that Theorem 5.3.1 holds w.p. at least* $(1 - 2n^{-10})$.

Definition 5.4.3 *Define the following.*

1. $\zeta_*^+ := r\zeta$

2. *define the series* $\{{\zeta_k}^+\}_{k=0,1,2,\cdots K}$ *as follows*

$$\zeta_0^+ := 1, \ \zeta_k^+ := \frac{b + 0.125c\zeta}{1 - (\zeta_*^+)^2 - (\zeta_*^+)^2 f - 0.25c\zeta - b}, \text{ for } k \geq 1, \tag{5.1}$$

where $b := C\kappa_s^+ g^+ \zeta_{k-1}^+ + \tilde{C}(\kappa_s^+)^2 g^+ (\zeta_{k-1}^+)^2 + C' f (\zeta_*^+)^2$, $\kappa_s^+ := 0.15$, $C := (\frac{2\kappa_s^+ \phi^+}{\sqrt{1-(\zeta_*^+)^2}} + \phi^+)$, $C' := ((\phi^+)^2 + \frac{2\phi^+}{\sqrt{1-(\zeta_*^+)^2}} + 1 + \phi^+ + \frac{\kappa_s^+ \phi^+}{\sqrt{1-(\zeta_*^+)^2}} + \frac{\kappa_s^+ (\phi^+)^2}{\sqrt{1-(\zeta_*^+)^2}})$, $\tilde{C} := ((\phi^+)^2 + \frac{\kappa_s^+ (\phi^+)^2}{\sqrt{1-(\zeta_*^+)^2}})$.

3. *define the series* $\{\tilde{\zeta}_k^+\}_{k=1,2,\cdots,\vartheta_j}$ *as follows*

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

where $f_{inc}(\tilde{g}, \tilde{h}) := (r + c)\zeta[3\kappa_{s,e}^+ \phi^+ \tilde{g} + [\kappa_{s,e}^+ \phi^+ + \kappa_{s,e}^+ (1 + 2\phi^+) \frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}]\tilde{h} + [\frac{r^2}{r+c}\zeta + 4r\zeta\kappa_{s,e}^+ \phi^+ + 2(r+c)\zeta(1 + {\kappa_{s,e}^+}^2){\phi^+}^2]f + 0.2\frac{1}{r+c}]$, *and* $f_{dec}(\tilde{g}, \tilde{h}) := 1 - \tilde{h} - 0.2\zeta - r^2\zeta^2 f - r^2\zeta^2 - f_{inc}(\tilde{g}, \tilde{h})$. *Notice that* $f_{inc}(\tilde{g}, \tilde{h})$ *is an increasing function of* $\tilde{g}, \tilde{h}$ *and* $f_{dec}(\tilde{g}, \tilde{h})$ *is a decreasing function of* $\tilde{g}, \tilde{h}$.

As we will see, ζ_*^+, ζ_k^+, $\tilde{\zeta}_k^+$ *are the high probability upper bounds on* $\zeta_{j,*}$, $\zeta_{j,k}$, $\tilde{\zeta}_{j,k}$ *(defined in Definition 5.4.8) under the assumptions of Theorem 5.3.1.*

Definition 5.4.4 *For the addition step, define*

1. $\Phi_{j,k} := I - \hat{P}_{j-1}\hat{P}_{j-1}' - \hat{P}_{j,new,k}\hat{P}_{j,new,k}'$ *and* $\Phi_{j,0} := I - \hat{P}_{j-1}\hat{P}_{j-1}'$.

2. $\phi_k := \max_j \max_{T:|T|\leq s} \|((\Phi_{j,k})_T{}'(\Phi_{j,k})_T)^{-1}\|_2$. *It is easy to see that* $\phi_k \leq \frac{1}{1-\max_j \delta_s(\Phi_{j,k})}$.

3. $D_{j,new,k} := \Phi_{j,k} P_{j,new}$ *and* $D_{j,new} := D_{j,new,0} = \Phi_{j,0} P_{j,new}$.

For the cluster-PCA step (for deletion), 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* $D_{j,undet,k} := \Psi_{j,k-1} G_{j,undet,k}$.

Here, $G_{j,det,k}$ *contains the directions that are already detected before the* k^{th} *step of cluster-PCA;* $G_{j,k}$ *contains the directions that are being detected in the current step;* $G_{j,undet,k}$ *contains the as yet undetected directions.*

Definition 5.4.5 *Let* $\kappa_{s,*} := \max_j \kappa_s(P_{j-1})$, $\kappa_{s,new} := \max_j \kappa_s(P_{j,new})$, $\kappa_{s,k} := \max_j \kappa_s(D_{j,new,k})$, $\tilde{\kappa}_{s,k} := \max_j \kappa_s((I - P_{j,new} P_{j,new}{}')\hat{P}_{j,new,k})$, $\kappa_{s,e} := \max_j \kappa_s(\Phi_K P_j)$.

Definition 5.4.6

1. *Let* $D_{j,k} \overset{QR}{=} E_{j,k} R_{j,k}$ *denote its QR decomposition. Here,* $E_{j,k}$ *is a basis matrix while* $R_{j,k}$ *is upper triangular.* [2]

2. *Let* $E_{j,k,\perp}$ *be a basis matrix for the orthogonal complement of* $span(E_{j,k}) = span(D_{j,k})$. *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*

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

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

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

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

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

[2] *Notice that* $0 < \sqrt{1 - r^2 \zeta^2} \le \sigma_i(R_{j,k})$ *by Lemma D.2.3, therefore,* $R_{j,k}$ *is invertible.*

4. *Define*

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

$$\tilde{\mathcal{H}}_{j,k} := \begin{bmatrix} E_{j,k} \, E_{j,k,\perp} \end{bmatrix} \begin{bmatrix} \tilde{H}_{j,k} & \tilde{B}'_{j,k} \\ \tilde{B}_{j,k} & \tilde{H}_{j,k,\perp} \end{bmatrix} \begin{bmatrix} {E_{j,k}}' \\ {E_{j,k,\perp}}' \end{bmatrix} \tag{5.2}$$

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 4 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 5.4.7 *Let* $\hat{P}_{j,*} := \hat{P}_{j-1} = \hat{P}_{(t_j-1)}$. *Recall that* $P_{j,*} := P_{(t_j-1)} = P_{j-1}$. *In the sequel, we use the subscript* $*$ *to denote the quantity at* $t = t_j - 1$.

Definition 5.4.8 (Subspace estimation errors)

1. *Recall that the subspace error at time* t *is* $SE_{(t)} := \|(I - \hat{P}_{(t)} \hat{P}'_{(t)}) P_{(t)}\|_2$.

2. *Define*

$$\zeta_{j,*} := \|(I - \hat{P}_{j,*} \hat{P}'_{j,*}) P_{j,*}\|_2.$$

This is the subspace error at $t = t_j - 1$, *i.e.* $\zeta_{j,*} = SE_{(t_j-1)}$.

3. *For* $k = 0, 1, 2, \cdots, K$, *define*

$$\zeta_{j,k} := \|(I - \hat{P}_{j-1} \hat{P}'_{j-1} - \hat{P}_{j,new,k} \hat{P}'_{j,new,k}) P_{j,new}\|_2.$$

This is the error in estimating span$(P_{j,new})$ *after the* k^{th} *iteration of the addition step.*

4. *For* $k = 1, 2, \cdots, \vartheta_j$, *define*

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

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

Remark 5.4.9 (Notational issue) *Notice that* ζ *is a given scalar satisfying the bound given in Theorem 5.3.1, while* $\zeta_{j,k}, \zeta_{j,*}$ *and* $\tilde{\zeta}_{j,k}$ *are as defined above. Since the basis matrix estimates are functions of the* $\hat{L}_t$*'s, which in turn are depend on the* L_t*'s and* $L_t = P_{(t)}a_t$*, thus,* $\zeta_{j,k}, \zeta_{j,*}$ *and* $\tilde{\zeta}_{j,k}$ *are functions of the* a_t*'s. Thus,* $\zeta_{j,k}, \zeta_{j,*}$ *and* $\tilde{\zeta}_{j,k}$ *are, in fact, random variables.*

Remark 5.4.10

1. *Notice that* $\zeta_{j,0} = \|D_{j,new}\|_2$, $\zeta_{j,k} = \|D_{j,new,k}\|_2$ *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)* $\hat{P}_{j,new,k}$ *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* $t \in \mathcal{I}_{j,k}$, $P_{(t)} = P_j = [(P_{j-1} \setminus P_{j,old}),\ P_{j,new}]$, $\hat{P}_{(t)} = [\hat{P}_{j-1}\ \hat{P}_{j,new,k}]$ *and*

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

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

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

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

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

Remark 5.4.11 *Recall that* $e_t := \hat{S}_t - S_t$*. Notice from Algorithm 4 that*

1. $e_t = L_t - \hat{L}_t$.

2. *If* $\hat{T}_t = T_t$*, then* $e_t = I_{T_t}[(\Phi_{(t)})_{T_t}{}'(\Phi_{(t)})_{T_t}]^{-1} I_{T_t}{}'\Phi_{(t)} P_{(t)} a_t$*. This follows using the definition of* $\hat{S}_t$ *given in step 1d of the algorithm and the fact that* $(\Phi_{(t)})_T'\Phi_{(t)} = (\Phi_{(t)}I_T)'\Phi_{(t)} = I_T'\Phi_{(t)}$ *for any set* T*. Thus, for* $t \in [t_j, t_{j+1}-1]$*,*

$$\begin{aligned} e_t &= I_{T_t}[(\Phi_{(t)})_{T_t}{}'(\Phi_{(t)})_{T_t}]^{-1} I_{T_t}{}'\Phi_{(t)} P_j a_t \\ &= I_{T_t}[(\Phi_{(t)})_{T_t}{}'(\Phi_{(t)})_{T_t}]^{-1} I_{T_t}{}'\Phi_{(t)} [P_{j,*}a_{t,*} + P_{j,new}a_{t,new}] \end{aligned} \tag{5.3}$$

with

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

Definition 5.4.12 *Define the random variable*

$$X_{j,k_1,k_2} := \{a_1, a_2, \cdots, a_{t_j+k_1\alpha+k_2\tilde{\alpha}-1}\}.$$

Recall that a_t*'s are mutually independent over* t*.*

Definition 5.4.13 *Define the set* $\check{\Gamma}_{j,k_1,k_2}$ *as follows.*

$$\begin{aligned} \check{\Gamma}_{j,k,0} &:= \{X_{j,k,0} : \zeta_{j,k} \le \zeta_k^+,\ \text{and } \hat{T}_t = T_t \text{ and } e_t \text{ satisfies } (5.3) \\ &\qquad \text{for all } t \in \mathcal{I}_{j,k}\},\ k = 1,2,\ldots K,\ j = 1,2,3,\ldots J \\ \check{\Gamma}_{j,K,k} &:= \{X_{j,K,k} : \tilde{\zeta}_{j,k} \le \tilde{c}_{j,k}\zeta,\ \text{and } \hat{T}_t = T_t \text{ and } e_t \text{ satisfies } (5.3) \\ &\qquad \text{for all } t \in \tilde{\mathcal{I}}_{j,k}\},\ k = 1,2,\ldots \vartheta_j,\ j = 1,2,3,\ldots J \\ \check{\Gamma}_{j,K,\vartheta_j+1} &:= \{X_{j+1,0,0} : \hat{T}_t = T_t \text{ and } e_t \text{ satisfies } (5.3) \\ &\qquad \text{for all } t \in \tilde{\mathcal{I}}_{j,\vartheta_j+1}\},\ j = 1,2,3,\ldots J \end{aligned}$$

Define the set Γ_{j,k_1,k_2} *as follows.*

$$\Gamma_{1,0,0} := \{X_{1,0,0} : \zeta_{1,*} \le r\zeta, \text{ and } \hat{T}_t = T_t \text{ and } e_t \text{ satisfies } (5.3) \text{ for all } t \in [t_{train}, t_1 - 1]\},$$

$$\Gamma_{j,k,0} := \Gamma_{j,k-1,0} \cap \check{\Gamma}_{j,k,0}, \ k = 1, 2, \dots K, \ j = 1, 2, 3, \dots J$$

$$\Gamma_{j,K,k} := \Gamma_{j,K,k-1} \cap \check{\Gamma}_{j,K,k}, \ k = 1, 2, \dots \vartheta_j, \ j = 1, 2, 3, \dots J$$

$$\Gamma_{j+1,0,0} := \Gamma_{j,K,\vartheta_j} \cap \check{\Gamma}_{j,K,\vartheta_j+1}, \ j = 1, 2, 3, \dots J$$

Recall from the notation section that the event $\Gamma^e_{j,k_1,k_2} := \{X_{j,k_1,k_2} \in \Gamma_{j,k_1,k_2}\}$.

Remark 5.4.14 *Notice that the subscript* j *always appears as the first subscript, while* k *is the last one. At many places in this paper, we remove the subscript* j *for simplicity. Whenever there is only one subscript, it refers to the value of* k*, e.g.,* Φ_0 *refers to* $\Phi_{j,0}$*,* $\hat{P}_{new,k}$ *refers to* $\hat{P}_{j,new,k}$ *and so on.*

5.4.1 Proof Outline of Theorem 5.3.1

The first part of the proof that analyzes the projected CS step and the addition step is essentially the same as that for ReProCS. The only difference is that, now, $\zeta_*^+ = r\zeta$ instead of $\zeta_*^+ = (r_0 + (j-1)c)\zeta$. In Lemma 5.4.15, the final conclusions for this part are summarized: it shows that, for all $k = 1, 2, \dots K$, ζ_k^+ decays roughly exponentially with k and it bounds the probability of $\Gamma^e_{j,k,0}$ given $\Gamma^e_{j,k-1,0}$. The second part of the proof analyzes the projected CS step and the cluster-PCA step. The final conclusion for this part is summarized in Lemma 5.4.16: it bounds the probability of $\Gamma^e_{j,K,k}$ given $\Gamma^e_{j,K,k-1}$. Theorem 5.3.1 follows essentially by applying Lemmas 5.4.16 and 5.4.15 for each j and k and using Lemma 2.3.2.

Lemma 5.4.16, in turn, follows by combining the results of Lemma D.2.2 (which shows exact support recovery and bounds the sparse recovery error for $t \in \tilde{I}_{j,k}$ conditioned on $\Gamma^e_{j,K,k-1}$), and Lemma D.2.8 (which bounds the subspace recovery error at the k^{th} step of cluster-PCA conditioned on $\Gamma^e_{j,K,k-1}$).

Lemma D.2.2 uses the result of Lemma D.2.1 which bounds the RIC of Φ_k in terms of ζ_*, ζ_k and the denseness coefficients of P_* and P_{new}. Lemma D.2.8 is obtained as follows. In Lemma D.2.4, we show that, under the theorem's assumptions, $\tilde{\zeta}_k^+ \le \tilde{c}_{j,k}\zeta$. In Lemma D.2.6, we bound

Table 5.1 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
done at $t = t_j + k\alpha - 1$
goal: keep improving estimates of span$(P_{j,\text{new}})$
compute $\hat{P}_{j,\text{new},k}$ by proj-PCA on $[\hat{L}_t : t \in \mathcal{I}_{j,k}]$ with $P = \hat{P}_{j-1}$
start with $\\|(I - \hat{P}_{j-1}\hat{P}_{j-1}')P_{j-1}\\|_2 \leq r\zeta$ and $\zeta_{j,k-1} \leq \zeta_{k-1}^{+} \leq 0.6^{k-1} + 0.4c\zeta$
need small $g_{j,k}$ which is the average of the condition number of $\text{Cov}(P_{j,\text{new}}' L_t)$ over $t \in \mathcal{I}_{j,k}$
no undetected subspace
$\zeta_{j,k}$ is the subspace error in estimating span$(P_{j,\text{new}})$ after the k^{th} step
end with $\zeta_{j,k} \leq \zeta_k^{+} \leq 0.6^k + 0.4c\zeta$ w.h.p.
stop when $k = K$ with K chosen so that $\zeta_{j,K} \leq c\zeta$
after K^{th} iteration: $\hat{P}_{(t)} \leftarrow [\hat{P}_{j-1} \ \hat{P}_{j,\text{new},K}]$ and $SE_{(t)} \leq (r+c)\zeta$
k^{th} iteration of cluster-PCA in the deletion step
done at $t = t_j + K\alpha + \vartheta_j\tilde{\alpha} - 1$
goal: re-estimate span(P_j) and thus "delete" span$(P_{j,\text{old}})$
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 $\\|(I - \hat{G}_{j,\text{det},k}\hat{G}_{j,\text{det},k}')G_{j,\text{det},k}\\|_2 \leq r\zeta$ and $\zeta_{j,K} \leq c\zeta$
need small $\tilde{g}_{j,k}$ which is the maximum of the condition number of $\text{Cov}(G_{j,k}' L_t)$ over $t \in \tilde{\mathcal{I}}_{j,k}$
extra issue: ensure perturbation due to span$(G_{j,\text{undet},k})$ is small; need small $\tilde{h}_{j,k}$ to ensure it
$\tilde{\zeta}_{j,k}$ is the subspace error in estimating span$(G_{j,k})$ after the k^{th} step
end with $\tilde{\zeta}_{j,k} \leq \tilde{c}_{j,k}\zeta$ w.h.p.
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 ϑ_j^{th} iteration: $\hat{P}_{(t)} \leftarrow [\hat{G}_{j,1}, \cdots, \hat{G}_{j,\vartheta_j}]$ and $SE_{(t)} \leq r\zeta$

$\tilde{\zeta}_k$ in terms of $\lambda_{\min}(A_k)$, $\lambda_{\max}(A_{k,\perp})$ and $\|\mathcal{H}_k\|_2$ using Lemma 2.2.1. Next, in Lemma D.2.7, (i) we use Lemma D.2.2 and the Hoeffding corollaries (Corollaries 2.3.4 and 2.3.5) to bound each of these terms and (ii) then we use Lemma D.2.6 and these bounds to bound $\tilde{\zeta}_k$ by $\tilde{\zeta}_k^{+}$ with a certain probability conditioned on $\Gamma_{j,K,k-1}^{e}$. Finally, Lemma D.2.8 follows by combining Lemma D.2.4 and Lemma D.2.7.

Our strategy for analyzing cluster-PCA and hence for proving Theorem 5.3.1 is a generalization of that used to analyze the k^{th} addition proj-PCA step for ReProCS. We discuss this in Table 5.1.

5.4.2 Key Lemmas

The theorem is a direct consequence of Lemmas 5.4.15 and 5.4.16 given below.

Lemma 5.4.15 is a slight modification of Lemma 4.4.21. It summarizes the final conclusions of the addition step.

Lemma 5.4.15 (Final lemma for addition step) *Assume that all the conditions in Theorem 5.3.1 holds. Also assume that* $\mathbf{P}(\Gamma^e_{j,k-1,0}) > 0$. *Then*

1. $\zeta_0^+ = 1$, $\zeta_k^+ \leq 0.6^k + 0.4c\zeta$ *for all* $k = 1, 2, \dots K$;

2. $\mathbf{P}(\Gamma^e_{j,k,0} \, | \Gamma^e_{j,k-1,0}) \geq p_k(\alpha, \zeta) \geq p_K(\alpha, \zeta)$ *for all* $k = 1, 2, \dots K$.

where ζ_k^+ *is defined in Definition 5.4.3 and* $p_k(\alpha, \zeta)$ *is defined in Lemma 4.4.16.*

The proof of the above lemma follows using the exact same approach as in the proof of Lemma 4.4.21 but with $\zeta_*^+ = r\zeta$ instead of $(r_0 + (j-1)c_{\max})\zeta$ everywhere. We give the proof outline in Appendix D.

Lemma 5.4.16 summarizes the final conclusions for the cluster-PCA step. It is proved using lemmas given in Sec D.2.

Lemma 5.4.16 (Final lemma for cluster-PCA) *Assume that all the conditions in Theorem 5.3.1 hold. Also assume that* $\mathbf{P}(\Gamma^e_{j,K,k-1}) > 0$. *Then,*

1. *for all* $k = 1, 2, \dots \vartheta_j$, $\mathbf{P}(\Gamma^e_{j,K,k} \mid \Gamma^e_{j,K,k-1}) \geq \tilde{p}(\tilde{\alpha}, \zeta)$ *where* $\tilde{p}(\tilde{\alpha}, \zeta)$ *is defined in Lemma D.2.8;*

2. $\mathbf{P}(\Gamma^e_{j+1,0,0} \mid \Gamma^e_{j,K,\vartheta_j}) = 1$.

proof Notice that $\mathbf{P}(\Gamma^e_{j,K,k} \mid \Gamma^e_{j,K,k-1}) = \mathbf{P}(\tilde{\zeta}_k \leq \tilde{c}_k\zeta \text{ and } \hat{T}_t = T_t, \text{ and } e_t \text{ satisfies (5.3) for all } t \in \tilde{I}_{j,k} \mid \Gamma^e_{j,K,k-1})$ and $\mathbf{P}(\Gamma^e_{j+1,0,0} \mid \Gamma^e_{j,K,\vartheta_j}) = \mathbf{P}(\hat{T}_t = T_t \text{ and } e_t \text{ satisfies (5.3) for all } t \in \mathcal{I}_{j,\vartheta_j+1})$. The first claim of the lemma follows by combining Lemma D.2.8 and the last claim of Lemma D.2.2, both given below in Sec D.2. The second claim follows using the last claim of Lemma D.2.2. ∎

Remark 5.4.17 *Under the assumptions of Theorem 5.3.1, it is easy to see that the following holds.*

1. *For any* $k = 1, 2 \ldots K$, $\Gamma^e_{j,k,0}$ *implies that (i)* $\zeta_{j,*} \leq \zeta_*^+ := r\zeta$ *and (ii)* $\zeta_{j,k'} \leq 0.6^{k'} + 0.4c\zeta$ *for all* $k' = 1, 2, \ldots k$

   - *(i) follows from the definition of* $\Gamma^e_{j,k,0}$ *and* $\zeta_{j,*} \leq \sum_{k=1}^{\vartheta_{j-1}} \tilde{\zeta}_{j-1,k'} \leq \sum_{k=1}^{\vartheta_{j-1}} \tilde{c}_{j-1,k'}\zeta = r_{j-1}\zeta \leq r\zeta = \zeta_*^+$*; and (ii) follows from the definition of* $\Gamma^e_{j,k,0}$ *and the first claim of Lemma 5.4.15.*

2. *For any* $k = 1, 2 \ldots \vartheta_j + 1$, $\Gamma^e_{j,K,k}$ *implies (i)* $\zeta_{j,*} \leq \zeta_*^+$*, (ii)* $\zeta_{j,k'} \leq 0.6^{k'} + 0.4c\zeta$ *for all* $k' = 1, 2, \ldots K$*, (iii)* $\zeta_{j,K} \leq c\zeta$*, (iv)* $\|\Phi_{j,K} P_j\|_2 \leq (r + c)\zeta$*, (v)* $\tilde{\zeta}_{j,k'} \leq \tilde{c}_{j,k'}\zeta$ *for* $k' = 1, 2, \ldots k$ *and (vi)* $\sum_{k'=1}^{k} \tilde{\zeta}_{j,k'} \leq r_j\zeta \leq r\zeta$.

   - *(i) and (ii) follow because* $\Gamma^e_{j,K,0} \subseteq \Gamma^e_{j,K,k}$*, (iii) follows from (ii) using the definition of* K*, (iv) follows from (i) and (iii) using* $\|\Phi_{j,K} P_j\|_2 \leq \|\Phi_{j,K}[P_{j,*}, P_{j,new}]\|_2 \leq \zeta_{j,*} + \zeta_{j,K}$*, and (v) follows from the definition of* $\Gamma^e_{j,K,k}$.

3. $\Gamma^e_{J+1,0,0}$ *implies (i)* $\zeta_{j,*} \leq \zeta_*^+$ *for all* j*, (ii)* $\zeta_{j,k} \leq 0.6^k + 0.4c\zeta$ *for all* $k = 1, \cdots, K$ *and all* j*, (iii)* $\zeta_{j,K} \leq c\zeta$ *for all* j.

5.4.3 Proof of Theorem 5.3.1

The theorem is a direct consequence of Lemmas 5.4.15 and 5.4.16 and Lemma 2.3.2.

Notice that $\Gamma^e_{j,0,0} \supseteq \Gamma^e_{j,1,0} \cdots \supseteq \Gamma^e_{j,K,0} \supseteq \Gamma^e_{j,K,1} \supseteq \Gamma^e_{j,K,2} \cdots \supseteq \Gamma^e_{j,K,\vartheta} \supseteq \Gamma^e_{j+1,0,0}$. Thus, by Lemma 2.3.2,

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

and $\mathbf{P}(\Gamma_{J+1,0,0}|\Gamma_{1,0,0}) = \prod_{j=1}^{J} \mathbf{P}(\Gamma^e_{j+1,0,0}|\Gamma^e_{j,0,0})$. Using Lemmas 5.4.15 and 5.4.16, and the fact that $p_k(\alpha,\zeta) \geq p_K(\alpha,\zeta)$, we get $\mathbf{P}(\Gamma^e_{J+1,0,0}|\Gamma_{1,0,0}) \geq p_K(\alpha,\zeta)^{KJ} \tilde{p}(\tilde{\alpha},\zeta)^{\vartheta_{\max} J}$. Also, $\mathbf{P}(\Gamma^e_{1,0,0}) = 1$. This follows by the assumption on $\hat{P}_0$ and Lemma D.2.2. Thus, $\mathbf{P}(\Gamma^e_{J+1,0,0}) \geq p_K(\alpha,\zeta)^{KJ} \tilde{p}(\tilde{\alpha},\zeta)^{\vartheta_{\max} J}$.

Using the definitions of $\alpha_{\text{add}}(\zeta)$ and $\alpha_{\text{del}}(\zeta)$ and $\alpha \geq \alpha_{\text{add}}$ and $\tilde{\alpha} \geq \alpha_{\text{del}}$, $\mathbf{P}(\Gamma^e_{J+1,0,0}) \geq p_K(\alpha,\zeta)^{KJ}\tilde{p}(\tilde{\alpha},\zeta)^{\vartheta_{\max}J} \geq (1-n^{-10})^2 \geq 1-2n^{-10}$.

The event $\Gamma^e_{J+1,0,0}$ implies that $\hat{T}_t = T_t$ and e_t satisfies (5.3) for all $t < t_{J+1}$. Using Remark 5.4.10 and the third claim of Remark 5.4.17, $\Gamma^e_{J+1,0,0}$ implies that all the bounds on the subspace error hold. Using these, Remark 5.4.11, $\|a_{t,\text{new}}\|_2 \leq \sqrt{c}\gamma_{\text{new},k}$ and $\|a_t\|_2 \leq \sqrt{r}\gamma_*$, $\Gamma^e_{J+1,0,0}$ implies that all the bounds on $\|e_t\|_2$ hold (the bounds are obtained in in Lemmas D.2.2 and D.1.2).

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

5.5 Experimental Results

The simulated data is generated as follows.

The measurement matrix $\mathcal{M}_t := [M_1, M_2, \cdots, M_t]$ is of size 2048×5200. 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 $\mathcal{S}_t := [S_1, S_2, \cdots, S_t]$ is generated as follows. For $1 \leq t \leq t_{\text{train}} = 200$, $S_t = 0$. For $t_{\text{train}} < t \leq 5200$, 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 $t > t_{\text{train}} + \Delta$). The signs of the nonzero elements of S_t are $P'_{1\to 2}1$ with equal probability and the magnitudes are uniformly distributed between 2 and 3. Thus, $S_{\min} = 2$.

The low rank matrix $\mathcal{L}_t := [L_1, L_2, \cdots, L_t]$ where $L_t := P_{(t)}a_t$ is generated as follows: There are a total of $J = 2$ subspace change times, $t_1 = 301$ and $t_2 = 2501$. $r_0 = 36$, $c_{1,\text{new}} = c_{2,\text{new}} = 1$ and $c_{1,\text{old}} = c_{2,\text{old}} = 3$. Let U be an $2048 \times (r_0 + c_{1,\text{new}} + c_{2,\text{new}})$ orthonormalized random Gaussian matrix. For $1 \leq t \leq t_1 - 1$, $P_{(t)} = P_0$ has rank r_0 with $P_0 = U_{[1,2,\cdots,36]}$. For $t_1 \leq t \leq t_2 - 1$, $P_{(t)} = P_1 = [P_0 \setminus P_{1,\text{old}} \ P_{1,\text{new}}]$ 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 $t \geq t_2$, $P_{(t)} = P_2 = [P_1 \setminus P_{2,\text{old}} \ P_{2,\text{new}}]$ has rank $r_2 = r_1 + c_{2,\text{new}} - c_{2,\text{old}} = 32$ with $P_{2,\text{new}} = U_{[38]}$ and $P_{1\text{old}} = U_{[8,17,35]}$. a_t is independent over

t. The various $(a_t)_i$'s are also mutually independent for different i. For $1 \leq t < t_1$, we let $(a_t)_i$ be uniformly distributed between $-\gamma_{i,t}$ and $\gamma_{i,t}$, where

$$\gamma_{i,t} = \begin{cases} 400 & \text{if } i = 1, 2, \cdots, 9, \forall t, \\ 30 & \text{if } i = 10, 11, \cdots, 18, \forall t. \\ 2 & \text{if } i = 19, 20, \cdots, 27, \forall t. \\ 1 & \text{if } i = 28, 29 \cdots, 36, \forall t. \end{cases} \tag{5.4}$$

For $t_1 \leq t < t_2$, $a_{t,*}$ is an $r_0 - c_{1,\text{old}}$ length vector, $a_{t,\text{new}}$ is a $c_{1,\text{new}}$ length vector and $L_t := P_{(t)} a_t = P_1 a_t = (P_0 \setminus P_{1,\text{old}}) a_{t,*,nz} + P_{1,\text{new}} a_{t,\text{new}}$. Now, $(a_{t,*,nz})_i$ is uniformly distributed between $-\gamma_{i,t}$ and $\gamma_{i,t}$ for $i = 1, 2, \cdots, 35$ and $a_{t,\text{new}}$ is uniformly distributed between $-\gamma_{\text{new},t}$ and $\gamma_{\text{new},t}$, where

$$\gamma_{i,t} = \begin{cases} 400 & \text{if } i = 1, 2, \cdots, 8, \forall t, \\ 30 & \text{if } i = 9, 10, \cdots, 16 \forall t. \\ 2 & \text{if } i = 17, 18, \cdots, 24, \forall t. \\ 1 & \text{if } i = 25, 26, \cdots, 33, \forall t. \end{cases}$$

$$\gamma_{\text{new},t} = \begin{cases} 1.1^{k-1} & \text{if } t_1 + (k-1)\alpha \leq t \leq t_1 + k\alpha - 1, k = 1, 2, 3, 4, \\ 1.1^{4-1} = 1.331 & \text{if } t \geq t_1 + 4\alpha. \end{cases} \tag{5.5}$$

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

$$\gamma_{i,t} = \begin{cases} 400 & \text{if } i = 1, 2, \cdots, 7, \forall t, \\ 30 & \text{if } i = 8, 9, \cdots, 14, \forall t. \\ 2 & \text{if } i = 15, 16, \cdots, 21, \forall t. \\ 1.331 & \text{if } i = 22, \forall t. \\ 1 & \text{if } i = 23, 24, \cdots, 31, \forall t. \end{cases} \tag{5.6}$$

$$\gamma_{\text{new},t} = \begin{cases} 1.1^{k-1} & \text{if } t_2 + (k-1)\alpha \le t \le t_2 + k\alpha - 1, k = 1, 2, \cdots, 7, \\ 1.1^{7-1} = 1.7716 & \text{if } t \ge t_2 + 7\alpha. \end{cases} \tag{5.7}$$

Thus for the above model, $S_{\min} = 2$, $\gamma_* = 400$, $\gamma_{\text{new}} = 1$, $\lambda^+ = 53333$, $\lambda^- = 0.3333$ and $f := \frac{\lambda^+}{\lambda^-} = 1.6 \times 10^5$. One way to get the clusters of $\{1, 2, \cdots, r_j\}$ is as follows.

1. For $t_1 \le t < t_2$ with $j = 1$, let $\mathcal{G}_{1,(1)} = \{1, 2, \cdots, 8\}$, $\mathcal{G}_{1,(2)} = \{9, 10, \cdots, 16\}$ and $\mathcal{G}_{1,(3)} = \{17, 18, \cdots, 34\}$. Thus, $\tilde{c}_{1,1} = \tilde{c}_{1,2} = 8$, $\tilde{c}_{1,3} = 18$, $\tilde{g}_{j,1} = \tilde{g}_{j,2} = 1$, $\tilde{g}_{j,3} = 4$, $\tilde{h}_{j,1} = 0.0056$, $\tilde{h}_{j,2} = 0.0044$.

2. For $t \ge t_2$ with $j = 2$, let $\mathcal{G}_{1,(1)} = \{1, 2, \cdots, 7\}$, $\mathcal{G}_{1,(2)} = \{8, 10, \cdots, 14\}$ and $\mathcal{G}_{1,(3)} = \{17, 18, \cdots, 32\}$. Thus, $\tilde{c}_{1,1} = \tilde{c}_{1,2} = 7$, $\tilde{c}_{1,3} = 16$, $\tilde{g}_{j,1} = \tilde{g}_{j,2} = 1$, $\tilde{g}_{j,3} = 4$, $\tilde{h}_{j,1} = 0.0056$, $\tilde{h}_{j,2} = 0.0044$.

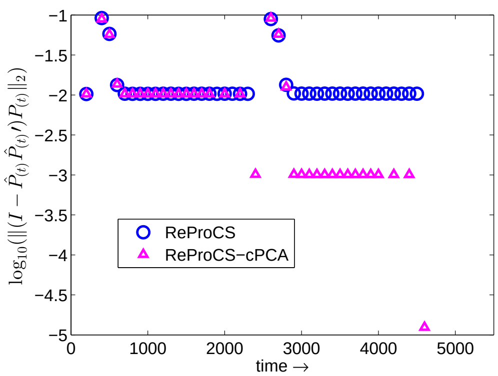
3. Therefore, $\tilde{g}_{\max} = 4$, $\tilde{h}_{\max} = 0.0056$ and $\tilde{c}_{\min} = 7$.

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.

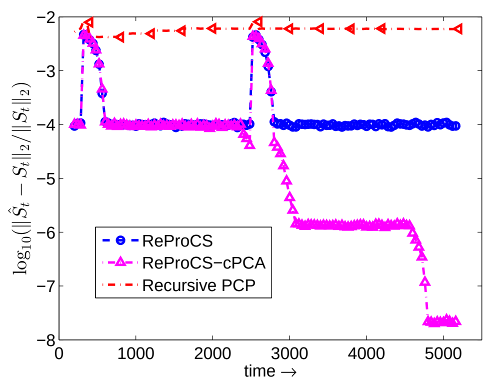
For Fig. 5.3 and Fig. 5.4, we used $s = 20$. We used $\Delta = 10$ for Fig. 5.3 and $\Delta = 50$ for Fig. 5.4. Because of the correlated support change, the $2048 \times t$ sparse matrix $\mathcal{S}_t = [S_1, S_2, \cdots, S_t]$ is rank deficient in either case, e.g. for Fig. 5.3, $\mathcal{S}_t$ has rank $29, 39, 49, 259$ at $t = 300, 400, 500, 2600$; for Fig. 5.4, $\mathcal{S}_t$ has rank $21, 23, 25, 67$ 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.

As can be seen from Fig. 5.3 and Fig. 5.4, the subspace error $\text{SE}_{(t)}$ of ReProCS and ReProCS-cPCA decreased exponentially and stabilized. Furthermore, ReProCS-cPCA outperforms over ReProCS greatly when deletion steps are done (i.e., at $t = 2400$ and 4600). The averaged normalized error for S_t followed a similar trend.

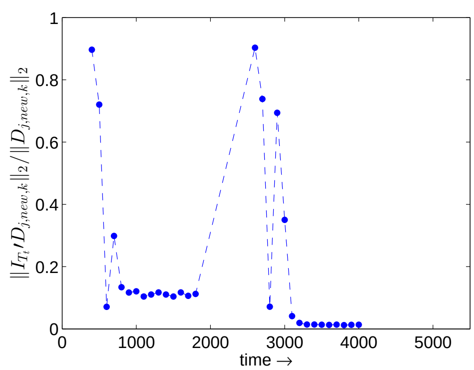
We also compared against PCP [2]. At every $t = t_j + 4k\alpha$, we solved (1.1) with $\lambda = 1/\sqrt{\max(n, t)}$ as suggested in [2] to recover $\mathcal{S}_t$ and $\mathcal{L}_t$. We used the estimates of S_t for the last



(a) subspace error, $\mathrm{SE}_{(t)}$

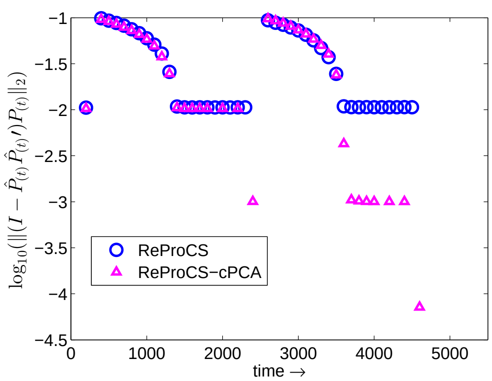


(b) recon error of S_t

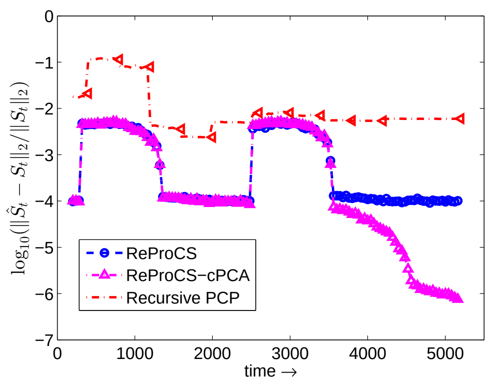


(c) plot of $\frac{\|I_{T_t}{}'D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$

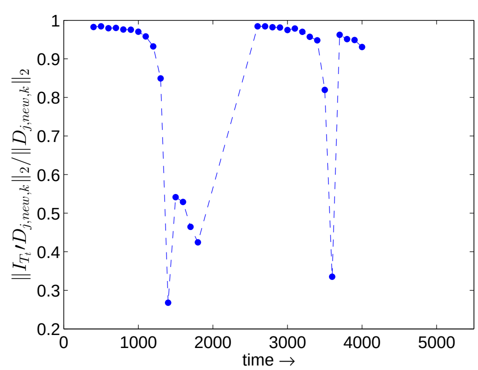
Figure 5.3 ReProCS-cPCA with $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 10$.



(a) subspace error, $\mathrm{SE}_{(t)}$



(b) recon error of S_t



(c) plot of $\frac{\|I_{T_t}{}' D_{j,\mathrm{new},k}\|_2}{\|D_{j,\mathrm{new},k}\|_2}$

Figure 5.4 ReProCS-cPCA with $r_0 = 36$, $s = \max_t |T_t| = 20$ and $\Delta = 50$

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.

We also plot the ratio $\frac{\|I_{T_t}{}' D_{j,\text{new},k}\|_2}{\|D_{j,\text{new},k}\|_2}$ at the projection PCA times. This serves as a proxy for $\kappa_s(D_{j,\text{new},k})$ (which has exponential computational complexity). As can be seen from Fig. 5.3 and Fig. 5.4, this ratio is less than 1 and it becomes larger when Δ increases (T_t becomes more correlated over t).

We implemented ReProCS-cPCA using Algorithm 4 with $\alpha = 100$, $\tilde{\alpha} = 200$ and $K = 15$. The algorithm is not very sensitive to these choices. Also, we let $\xi = \xi_t$ and $\omega = \omega_t$ vary with time. Recall that ξ_t is the upper bound on $\|\beta_t\|_2$. We do not know β_t. All we have is an estimate of β_t from $t-1$, $\hat{\beta}_{t-1} = (I - \hat{P}_{t-1}\hat{P}_{t-1}')\hat{L}_{t-1}$. We used a value a little larger than $\|\hat{\beta}_{t-1}\|_2$; we let $\xi_t = 2\|\hat{\beta}_{t-1}\|_2$. The parameter ω_t is the support estimation threshold. One reasonable way to pick this is to use a percentage energy threshold of $\hat{S}_{t,\text{cs}}$ [40]. For a vector v, define the 99%-energy set of v as $T_{0.99}(v) := \{i : |v_i| \geq v^{0.99}\}$ where the 99% energy threshold, $v^{0.99}$, is the largest value of $|v_i|$ so that $\|v_{T_{0.99}}\|_2^2 \geq 0.99\|v\|_2^2$. It is computed by sorting $|v_i|$ in non-increasing order of magnitude. One keeps adding elements to $T_{0.99}$ until $\|v_{T_{0.99}}\|_2^2 \geq 0.99\|v\|_2^2$. We used $\omega_t = 0.5(\hat{S}_{t,\text{cs}})^{0.99}$.

Algorithm 4 Recursive Projected CS with cluster-PCA (ReProCS-cPCA)

Parameters: algorithm parameters: ξ, ω, α, $\tilde{\alpha}$, K, model parameters: t_j, r_0, $c_{j,\text{new}}$, 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 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($[\hat{L}_t; t \in \mathcal{I}_{j,k}], \hat{P}_{j-1}, c_{j,\text{new}}$)
      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 span(P_j) by cluster-PCA:** If $t = t_j + K\alpha + \vartheta_j \tilde{\alpha} - 1$,
      i. For $i = 1, 2, \cdots, \vartheta_j$,
         - $\hat{G}_{j,i} \leftarrow$ proj-PCA($[\hat{L}_t; t \in \tilde{\mathcal{I}}_{j,k}], [\hat{G}_{j,1}, \hat{G}_{j,2}, \ldots \hat{G}_{j,i-1}], \tilde{c}_{j,i}$)

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

CHAPTER 6. Conclusions and Future Work

We studied the problem of recursive sparse recovery in the presence of large but structured noise (noise lying in a "slowly changing" low dimensional subspace). We introduced ReProCS and ReProCS with cluster-PCA (ReProCS-cPCA) algorithm that addresses some of the limitations of PCP [2]. ReProCS assumes that the subspace in which the most recent several L_t's lie can only grow over time and hence it needs to assume a bound on the total number of subspace changes, J. Unlike ReProCS, ReProCS-cPCA does not bound the number of allowed subspace changes, J, as long as the delay between change times is increased in proportion to $\log J$. Under mild assumptions, we showed that, w.h.p., ReProCS and ReProCS-cPCA can exactly recover the support set of S_t at all times; and the reconstruction errors of both S_t and L_t are upper bounded by a time-invariant and small value at all times.

In ongoing work, we are studying the undersampled measurements case. On the other hand, open questions also include (i) how to analyze a practical version of ReProCS-cPCA (which does not assume knowledge of signal model parameters), and (ii) how to study the correlated a_t's case (e.g. the case where a_t's satisfy a linear random walk model). The starting point for (ii) would be to try to use the matrix Azuma inequality [25] instead of Hoeffdding.

APPENDIX A. Proof of the Lemmas and Corollaries in Chapter 2

A.1 Proof of Lemma 2.2.4

proof: 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. $\|PP' - \hat{P}\hat{P}'\|_2 = \|PP - \hat{P}\hat{P}'PP' + \hat{P}\hat{P}'PP' - \hat{P}\hat{P}'\|_2 \leq \|(I - \hat{P}\hat{P}')PP'\|_2 + \|(I - PP')\hat{P}\hat{P}'\|_2 = 2\zeta_*$.

3. Since $Q'P = 0$, then $\|Q'\hat{P}\|_2 = \|Q'(I - PP')\hat{P}\|_2 \leq \|(I - PP')\hat{P}\|_2 = \zeta_*$.

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, $\lambda_{\max}(Q'(I - \hat{P}\hat{P}')Q) \leq 1$. By Weyl's Theorem, $\lambda_{\min}(Q'(I - \hat{P}\hat{P}')Q) \geq 1 - \lambda_{\max}(Q'\hat{P}\hat{P}'Q) = 1 - \|Q'\hat{P}\|_2^2 \geq 1 - \zeta_*^2$. Therefore, $\sqrt{1 - \zeta_*^2} \leq \sigma_i((I - \hat{P}\hat{P}')Q) \leq 1$.

∎

A.2 Proof of Lemma 2.3.1

proof: 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 $\mathbf{E}[\mathbb{I}_{\mathcal{B}}(X,Y)|X] \geq p$ for all $X \in \mathcal{C}$, this means that $\mathbf{E}[\mathbb{I}_{\mathcal{B}}(X,Y)|X]\mathbb{I}_{\mathcal{C}}(X) \geq p\mathbb{I}_{\mathcal{C}}(X)$. This, in turn, implies that

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

Recall from Definition 1.1.3 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 $\mathbf{P}(\mathcal{B}^e|X) \geq p$ for all $X \in \mathcal{C}$, then $\mathbf{P}(\mathcal{B}^e, \mathcal{C}^e) \geq p\mathbf{P}(\mathcal{C}^e)$. Using the definition of $\mathbf{P}(\mathcal{B}^e|\mathcal{C}^e)$, the claim follows. ■

A.3 Proof of Corollary 2.3.4

proof:

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 2.3.3, $\sigma^2 = \|\sum_t (b_2 - b_1)^2 I\|_2 = \alpha(b_2 - b_1)^2$. For any $X \in \mathcal{C}$, applying Theorem 2.3.3 for $\{Y_t\}$'s conditioned on X, we get that, for any $\epsilon > 0$,

$$\mathbf{P}(\lambda_{\max}(\frac{1}{\alpha}\sum_t Y_t) \leq \epsilon|X) > 1 - n\exp(-\frac{\alpha\epsilon^2}{8(b_2 - b_1)^2}) \text{ for all } X \in \mathcal{C}$$

By Weyl's theorem, $\lambda_{\max}(\frac{1}{\alpha}\sum_t Y_t) = \lambda_{\max}(\frac{1}{\alpha}\sum_t (Z_t - \mathbf{E}(Z_t|X)) \geq \lambda_{\max}(\frac{1}{\alpha}\sum_t Z_t) + \lambda_{\min}(\frac{1}{\alpha}\sum_t -\mathbf{E}(Z_t|X))$. Since $\lambda_{\min}(\frac{1}{\alpha}\sum_t -\mathbf{E}(Z_t|X)) = -\lambda_{\max}(\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X)) \geq -b_4$, thus $\lambda_{\max}(\frac{1}{\alpha}\sum_t Y_t) \geq \lambda_{\max}(\frac{1}{\alpha}\sum_t Z_t) - b_4$. Therefore,

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

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 2.3.3, we

get that for any $\epsilon > 0$,

$$\mathbf{P}(\lambda_{\max}(\frac{1}{\alpha}\sum_t Y_t) \leq \epsilon | X) > 1 - n\exp(-\frac{\alpha\epsilon^2}{8(b_2 - b_1)^2}) \text{ for all } X \in \mathcal{C}$$

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

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

■

A.4 Proof of Corollary 2.3.5

proof: Define the dilation of an $n_1 \times n_2$ matrix M as $\text{dilation}(M) := \begin{bmatrix} 0 & M' \\ M & 0 \end{bmatrix}$. Notice that this is an $(n_1 + n_2) \times (n_1 + n_2)$ Hermitian matrix [25]. As shown in [25, equation 2.12],

$$\lambda_{\max}(\text{dilation}(M)) = \|\text{dilation}(M)\|_2 = \|M\|_2 \tag{A.1}$$

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 (A.1), 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 $\text{dilation}(Z_t)$'s. Thus, applying Corollary 2.3.4 for the sequence $\{\text{dilation}(Z_t)\}$, we get that,

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

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

APPENDIX B. Proof of Lemma 3.3.2

proof 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,

$$\delta_s(I - PP') = \max_{|T|\leq s}(1 - \lambda_{\min}(I - I_T' PP' I_T)) \tag{B.1}$$

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 $\|I_T' P\|_2^2 = \lambda_{\max}(I_T' PP' I_T) = 1 - \lambda_{\min}(I - I_T' PP' I_T)$ [1], and so

$$\kappa_s^2(P) = \max_{|T|\leq s}(1 - \lambda_{\min}(I - I_T' PP' I_T)) \tag{B.2}$$

From (B.1) and (B.2), we get $\delta_s(I - PP') = \kappa_s^2(P)$. ■

[1] 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 C. Proof of the Lemmas in Chapter 4

C.1 Proof of Lemma 4.4.10

proof:

1. Since P is a basis matrix, $\kappa_s^2(P) = \max_{|T|\le s} \|I_T{}'P\|_2^2$. Also, $\|I_T{}'P\|_2^2 = \|I_T{}'[P_1, P_2] [P_1, P_2]'I_T\|_2 = \|I_T{}'(P_1P_1' + P_2P_2')I_T\|_2 \le \|I_T{}'P_1P_1'I_T\|_2 + \|I_T{}'P_2P_2'I_T\|_2$. Thus, the inequality follows.

2. For any set T with $|T| \le s$, $\|I_T{}'\hat{P}_*\|_2^2 = \|I_T{}'\hat{P}_*\hat{P}_*'I_T\|_2 = \|I_T{}'(\hat{P}_*\hat{P}_*' - P_*P_*{}' + P_*P_*{}')I_T\|_2 \le \|I_T{}'(\hat{P}_*\hat{P}_*' - P_*P_*{}')I_T\|_2 + \|I_T{}'P_*P_*{}'I_T\|_2 \le 2\zeta_* + \kappa_{s,*}^2$. The last inequality follows using Lemma 2.2.4 with $P = P_*$ and $\hat{P} = \hat{P}_*$.

3. By Lemma 2.2.4 with $P = P_*$, $\hat{P} = \hat{P}_*$ and $Q = P_{\text{new}}$, $\|P_{\text{new}}{}'\hat{P}_*\|_2 \le \zeta_*$. By Lemma 2.2.4 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 $|T| \le s$, $\|I_T{}'\hat{P}_{\text{new},k}\|_2 \le \|I_T{}'(I - P_{\text{new}}P_{\text{new}}')\hat{P}_{\text{new},k}\|_2 + \|I_T{}'P_{\text{new}}P_{\text{new}}'\hat{P}_{\text{new},k}\|_2 \le \tilde{\kappa}_{s,k}\|(I - P_{\text{new}}P_{\text{new}}{}')\hat{P}_{\text{new},k}\|_2 + \|I_T{}'P_{\text{new}}\|_2 = \tilde{\kappa}_{s,k}\|(I - \hat{P}_{\text{new},k}\hat{P}_{\text{new},k}')P_{\text{new}}\|_2 + \|I_T{}'P_{\text{new}}\|_2 \le \tilde{\kappa}_{s,k}\|D_{\text{new},k}\|_2 + \tilde{\kappa}_{s,k}\|\hat{P}_*\hat{P}_*'P_{\text{new}}\|_2 + \|I_T{}'P_{\text{new}}\|_2 \le \tilde{\kappa}_{s,k}\zeta_k + \tilde{\kappa}_{s,k}\zeta_* + \kappa_{s,\text{new}} \le \tilde{\kappa}_{s,k}\zeta_k + \zeta_* + \kappa_{s,\text{new}}$. Taking max over $|T| \le s$ the claim follows.

4. This follows using Lemma 3.3.2 and the second claim of this lemma.

5. This follows using Lemma 3.3.2 and the first three claims of this lemma.

■

C.2 Simple facts

Let ζ_k^+ denote the bound on $\zeta_{j,k}$ for any j. We obtain an expression for ζ_k^+ later.

Fact C.2.1 *Suppose* $\kappa_{2s,*} \leq \kappa_{2s,*}^+ = 0.3$*,* $\kappa_{2s,new} \leq \kappa_{2s,new}^+ = 0.15$*,* $\tilde{\kappa}_{2s,k} \leq \tilde{\kappa}_{2s}^+ = 0.15$*, and* $\kappa_{s,k} \leq \kappa_s^+ = 0.15$*. Pick* ζ *as in Theorem 4.3.1 and set* $\zeta_*^+ = (r_0 + (j-1)c)\zeta$*. Then,*

1. $\zeta\gamma_* \leq \frac{\sqrt{\zeta}}{(r_0+(J-1)c)^{3/2}} \leq \sqrt{\zeta}$
2. $\zeta_*^+ \leq \frac{10^{-4}}{(r_0+(J-1)c)} \leq 10^{-4}$
3. $\zeta_*^+ \gamma_{new,k}^2 \leq \zeta_*^+ \gamma_*^2 \leq \frac{1}{(r_0+(J-1)c)^2} \leq 1$
4. $\zeta_*^+ \gamma_{new,k} \leq \zeta_*^+ \gamma_* \leq \frac{\sqrt{\zeta}}{\sqrt{r_0+(J-1)c}} \leq \sqrt{\zeta}$
5. $\zeta_*^+ f \leq \frac{1.5\times 10^{-4}}{r_0+(J-1)c} \leq 1.5 \times 10^{-4}$
6. *If* $\zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$*, then* $\zeta_{k-1}^+ \gamma_{new,k} \leq (0.6 \cdot 1.2)^{k-1}\gamma_{new} + 0.4c\zeta\gamma_* \leq 0.72^{k-1}\gamma_{new} + \frac{0.4\sqrt{\zeta}}{\sqrt{r_0+(J-1)c}} \leq 0.72^{k-1}\gamma_{new} + 0.4\sqrt{\zeta}$
7. *If* $\zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$*, then* $\zeta_{k-1}^+ \gamma_{new,k}^2 \leq (0.6 \cdot 1.2^2)^{k-1}\gamma_{new}^2 + 0.4c\zeta\gamma_*^2 \leq 0.864^{k-1}\gamma_{new}^2 + \frac{0.4}{(r_0+(J-1)c)^2} \leq 0.864^{k-1}\gamma_{new}^2 + 0.4$
8. *If* $\zeta_* \leq \zeta_*^+$*,* $\zeta_k \leq \zeta_k^+$ *and* $\zeta_k^+ \leq 0.6^k + 0.4c\zeta$*, then*

   (a) $\delta_s(\Phi_0) \leq \delta_{2s}(\Phi_0) \leq {\kappa_{2s,*}^+}^2 + 2\zeta_*^+ < 0.1 < 0.1479$

   (b) $\delta_s(\Phi_k) \leq \delta_{2s}(\Phi_k) \leq {\kappa_{2s,*}^+}^2 + 2\zeta_*^+ + (\kappa_{2s,new}^+ + \tilde{\kappa}_{2s,k}^+ \zeta_k^+ + \zeta_*^+)^2 < 0.1479$

   (c) $\phi_k \leq \frac{1}{1-\delta_s(\Phi_k)} < 1.1735$

proof: The first seven items follow directly. The eighth item follows using Lemma 4.4.10. ■

C.3 Proof of Lemma 4.4.11

proof:

1. 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, $\|\beta_t\|_2 \leq \zeta_*\sqrt{r}\gamma_* + \zeta_{k-1}\sqrt{c}\gamma_{\text{new},k} \leq \sqrt{c}0.72^{k-1}\gamma_{\text{new}} + \sqrt{\zeta}(\sqrt{r} + 0.4\sqrt{c}) \leq \xi_0$. The second last inequality follows using Fact C.2.1.

2. By Fact C.2.1 and condition 2 of the theorem, $\delta_{2s}(\Phi_{k-1}) < 0.15 < \sqrt{2}-1$. Given $|T_t| \leq s$, $\|\beta_t\|_2 \leq \xi_0 = \xi$ and $\delta_s(\Phi_{k-1}) < \sqrt{2}-1$, by Theorem 2.1.1, the CS error satisfies $\|\hat{S}_{t,\text{cs}} - S_t\|_2 \leq \frac{4\sqrt{1+\delta_{2s}(\Phi_{k-1})}}{1-(\sqrt{2}+1)\delta_{2s}(\Phi_{k-1})}\xi_0 < 7\xi_0$.

3. Using the above and the definition of ρ, $\|\hat{S}_{t,\text{cs}} - S_t\|_\infty \leq 7\rho\xi_0$. Since $\min_t |(S_t)_{T_t}| \geq S_{\min}$ and $(S_t)_{T_t^c} = 0$, $\min_t |(\hat{S}_{t,cs})_{T_t}| \geq S_{\min} - 7\rho\xi_0$ and $\min_t |(\hat{S}_{t,\text{cs}})_{T_t^c}| \leq 7\rho\xi_0$. If $\omega < S_{\min} - 7\rho\xi_0$, then $\hat{T}_t \supseteq T_t$. On the other hand, if $\omega > 7\rho\xi_0$, then $\hat{T}_t \subseteq T_t$. Since $S_{\min} > 14\rho\xi_0$(condition 3 of the theorem) and ω satisfies $7\rho\xi_0 \leq \omega \leq S_{\min} - 7\rho\xi_0$(condition 1 of the theorem), then the support of S_t is exactly recovered, i.e. $\hat{T}_t = T_t$.

4. 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 (4.2). Thus, using Fact C.2.1 and condition 2 of the theorem, $\|e_t\|_2 \leq \phi^+(\zeta_*^+\sqrt{r}\gamma_* + \kappa_{s,k-1}\zeta_{k-1}^+\sqrt{c}\gamma_{\text{new},k} \leq 1.2(\sqrt{r}\sqrt{\zeta} + \sqrt{c}0.15(0.72)^{k-1} + \sqrt{c}0.06\sqrt{\zeta}) = 0.18\sqrt{c}0.72^{k-1}\gamma_{\text{new}} + 1.2\sqrt{\zeta}(\sqrt{r} + 0.06\sqrt{c})$.

■

C.4 Proof of Lemma 4.4.12

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 2.2.2), $\lambda_{c_{\text{new}}+1}(\mathcal{A}_k + \mathcal{H}_k) \leq \lambda_{c_{\text{new}}+1}(\mathcal{A}_k) + \|\mathcal{H}_k\|_2 = \|A_{k,\perp}\|_2 + \|\mathcal{H}_k\|_2$. Therefore, $\lambda_{\max}(\Lambda_{k,\perp}) \leq \|A_{k,\perp}\|_2 + \|\mathcal{H}_k\|_2$ and hence $\lambda_{\min}(A_k) - \lambda_{\max}(\Lambda_{k,\perp}) \geq \lambda_{\min}(A_k) - \|A_{k,\perp}\|_2 - \|\mathcal{H}_k\|_2 > 0$. Apply the $\sin\theta$ theorem (Theorem 2.2.1) with $\lambda_{\min}(A_k) - \lambda_{\max}(\Lambda_{k,\perp}) > 0$, we get

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

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

■

C.5 Key facts for proving Lemmas 4.4.14 and 4.4.15

In this and the next two subsections, we use $\frac{1}{\alpha}\sum_t$ to denote $\frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}}$.

Lemmas 4.4.14 and 4.4.15 can be proved using the following facts and Corollaries 2.3.4 and 2.3.5. Under the assumptions of these lemmas, the following are true.

1. Recall from the model (Sec 3.1) and from condition 3 of Theorem 4.3.1 that (i) $a_{t,\text{new}}$ and $a_{t,*}$ are mutually uncorrelated, (ii) $\|a_{t,*}\|_2 \le \sqrt{r}\gamma_*$, (iii) for $t \in \mathcal{I}_{j,k}$, $\|a_{t,\text{new}}\|_2 \le \sqrt{c}\gamma_{\text{new},k}$ and $\|a_{t,*}a_{t,\text{new}}\|_2 \le \sqrt{cr}\gamma_{\text{new},k}\gamma_*$.

2. Recall that

   (a) $f := \lambda^+/\lambda^-$ where $\lambda^+ := \max_t \lambda_{\max}(\Lambda_t)$ and $\lambda^- := \min_t \lambda_{\min}(\Lambda_t)$ and so $\lambda^+_{\text{new},k} \le \lambda^+$, $\lambda^-_{\text{new},k} \ge \lambda^-$

   (b) $\Phi_0 = I - \hat{P}_*\hat{P}_*'$, $\Phi_{k-1} = I - \hat{P}_*\hat{P}_*' - \hat{P}_{\text{new},k-1}\hat{P}'_{\text{new},k-1}$, $D_{\text{new},k-1} = \Phi_{k-1}P_{\text{new}}$, $D_{\text{new}} = D_{\text{new},0} = \Phi_0 P_{\text{new}} \overset{QR}{=} E_{\text{new}}R_{\text{new}}$, $D_* = \Phi_0 P_*$, $\zeta_* = \|D_*\|$, $\zeta_{k-1} = \|D_{\text{new},k-1}\|$ with $\zeta_0 = \|D_{\text{new}}\|$.

   (c) Conditions 2 and 4 of Theorem 4.3.1 imply that $\kappa_{2s,*} \le \kappa^+_{2s,*} = 0.3$ and $\kappa_{2s,\text{new}} \le \kappa^+_{2s,\text{new}} = 0.15$, $\tilde{\kappa}_{2s,k} \le \tilde{\kappa}^+_{2s} = 0.15$, $\kappa_{s,k} \le \kappa^+_s = 0.15$ and $g_{j,k} \le g^+ = \sqrt{2}$.

   (d) The r.v. $X_{j,k-1}$ and the set $\Gamma_{j,k-1}$ are defined in Lemma 4.4.14.

3. 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 $\|E_{\text{new}}{}'\Phi_0 e_t\| = \|(R'_{\text{new}})^{-1}D'_{\text{new}}\Phi_0 e_t\| = \|(R_{\text{new}})^{-1}D'_{\text{new}}e_t\| \le \|(R'_{\text{new}})^{-1}D'_{\text{new}}I_{T_t}\|\|e_t\| \le \frac{\kappa_s^+}{\sqrt{1-\zeta_*^2}}\|e_t\|$. The bounds on $\|R_{\text{new}}\|$ and $\|(R_{\text{new}})^{-1}\|$ follows using Lemma 2.2.4 and the fact that $\sigma_i(R_{\text{new}}) = \sigma_i(D_{\text{new}})$.

4. $X_{j,k-1} \in \Gamma_{j,k-1}$ implies that $\zeta_{k-1} \le \zeta^+_{k-1}$ and $\zeta_* \le \zeta^+_*$. We prove this below. This, in turn, implies that

   (a) $\lambda_{\min}(R_{\text{new}}R_{\text{new}}{}') \ge 1 - (\zeta^+_*)^2$. This follows from Lemma 2.2.4 and the fact that $\sigma_{\min}(R_{\text{new}}) = \sigma_{\min}(D_{\text{new}})$.

(b) $\|I_{T_t}{}'\Phi_{k-1}P_*\|_2 \leq \|\Phi_{k-1}P_*\|_2 \leq \zeta_* \leq \zeta_*^+$, $\|I_{T_t}{}'D_{\text{new},k-1}\|_2 \leq \kappa_{s,k-1}\zeta_{k-1} \leq \kappa_s^+\zeta_{k-1}^+$.

(c) $\phi_{k-1} := \|[(\Phi_{k-1})_{T_t}{}'(\Phi_{k-1})_{T_t}]^{-1}\|_2 \leq \phi^+ = 1.2$. This follows from Fact C.2.1.

5. $\mathbf{P}(\{\hat{T}_t = T_t$ and e_t satisfies (4.2) for all $t \in \mathcal{I}_{j,k}\}|X_{j,k-1}) = 1$ for all $X_{j,k-1} \in \Gamma_{j,k-1}$. We prove this below. In other words, conditioned on $X_{j,k-1}$, $\hat{T}_t = T_t$ and e_t satisfies

$$e_t = I_{T_t}[(\Phi_{k-1})_{T_t}{}'(\Phi_{k-1})_{T_t}]^{-1}I_{T_t}{}'[(\Phi_{k-1}P_*)a_{t,*} + D_{\text{new},k-1}a_{t,\text{new}}]$$

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

6. The matrices D_{new}, R_{new}, E_{new}, D_*, $D_{\text{new},k-1}$, Φ_{k-1} are functions of the r.v. $X_{j,k-1}$ (defined in Lemma 4.4.14).

   (a) Thus, all terms that we bound in the proof of Lemma 4.4.14 are of the form $\frac{1}{\alpha}\sum_{t\in\mathcal{I}_{j,k}} Z_t$ where Z_t can be rewritten as either $Z_t = f_1(X_{j,k-1})a_{t,*}a_{t,*}' f_2(X_{j,k-1})$ or $Z_t = f_1(X_{j,k-1})a_{t,\text{new}}a_{t,\text{new}}' f_2(X_{j,k-1})$ or $Z_t = f_1(X_{j,k-1})a_{t,*}a_{t,\text{new}}' f_2(X_{j,k-1})$ for some functions $f_1(.)$ and $f_2(.)$.

   (b) Conditioned on $X_{j,k-1}$, all terms that we bound in the proof of Lemma 4.4.15 are also of the above form, whenever $X_{j,k-1} \in \Gamma_{j,k-1}$. This follows using item 5 (all terms that we bound in the proof of this lemma contain e_t).

7. $X_{j,k-1}$ is independent of any $a_{t,*}$ or $a_{t,\text{new}}$ for $t \in \mathcal{I}_{j,k}$, and hence the same is true for the matrices D_{new}, R_{new}, E_{new}, D_*, $D_{\text{new},k-1}$, Φ_{k-1} (which are functions of $X_{j,k-1}$). Also, $a_{t,*}$'s for different $t \in \mathcal{I}_{j,k}$ are mutually independent and the same is true for $a_{t,\text{new}}$'s for $t \in \mathcal{I}_{j,k}$.

8. Combining the previous two facts, for Lemma 4.4.14, conditioned on $X_{j,k-1}$, the Z_t's given in item 6 are mutually independent. For Lemma 4.4.15, conditioned on $X_{j,k-1}$, the Z_t's given in item 6 are mutually independent, whenever $X_{j,k-1} \in \Gamma_{j,k-1}$.

9. The assumption that $\zeta_{k-1} \leq 0.6^{k-1} + 0.4c\zeta$ is combined with Fact C.2.1 to get simple expressions for the probabilities with which the bounds hold.

10. The statement *"conditioned on r.v. X, the event $\mathcal{E}^e$ holds w.p. one for all $X \in \Gamma$" is equivalent to "$\mathbf{P}(\mathcal{E}^e|X) = 1$, for all $X \in \Gamma$"*. We often use the former statement in our proofs since it is often easier to interpret.

Proof of item 4: $\zeta_{k-1} \le \zeta_{k-1}^+$ follows from the definition of $\Gamma_{j,k-1}$. Also, the definition implies that $\zeta_{1,*} \le r_0\zeta$ and $\zeta_{j',K} \le \zeta_K^+$ for all $j' \le j-1$. Using the definition of K from Theorem 4.3.1 and using the assumption on ζ_k^+, this implies that $\zeta_{j',K} \le 0.6^K + 0.4c\zeta \le c\zeta$ for all $j' \le (j-1)$. Using Remark 4.4.4, this implies that $\zeta_* \le r_0\zeta + (j-1)c\zeta = \zeta_*^+$.

Proof of item 5: $X_{j,k-1} \in \Gamma_{j,k-1}$ implies that $\zeta_{k-1} \le \zeta_{k-1}^+$ and $\zeta_* \le \zeta_*^+ = r_0 + (j-1)\zeta$. This follows using item 4. By assumption, $\zeta_{k-1}^+ \le 0.6^{k-1} + 0.4c\zeta$ and the four conditions of Theorem 4.3.1 hold. Thus, conditioned on $X_{j,k-1}$, all conditions of Lemma D.1.2 hold as long as $X_{j,k-1} \in \Gamma_{j,k-1}$. Applying Lemma D.1.2, (i) $\hat{T}_t = T_t$ for all $t \in \mathcal{I}_{j,k}$; and (ii) for this duration, e_t satisfies (4.2), i.e. the claim follows.

C.6 Proof of Lemma 4.4.14

proof: In this proof, we frequently refer to items from the previous subsection, i.e. Sec. C.5.

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

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

Consider $\sum_t Z_t = \sum_t R_{\text{new}} a_{t,\text{new}} a_{t,\text{new}}{}' R'_{\text{new}}$. (a) Using item 8 of Sec. C.5, the Z_t's are conditionally independent given $X_{j,k-1}$. (b) Using item 3, Ostrowoski's theorem (Theorem 2.2.3), and item 4, for all $X_{j,k-1} \in \Gamma_{j,k-1}$,

$$\lambda_{\min}(\mathbf{E}(\frac{1}{\alpha}\sum_t Z_t|X_{j,k-1})) = \lambda_{\min}(R_{\text{new}}\frac{1}{\alpha}\sum_t \mathbf{E}(a_{t,\text{new}} a_{t,\text{new}}{}')R_{\text{new}}{}')$$
$$\ge \lambda_{\min}(R_{\text{new}} R_{\text{new}}{}')\lambda_{\min}(\frac{1}{\alpha}\sum_t \mathbf{E}(a_{t,\text{new}} a_{t,\text{new}}{}')) \ge (1-(\zeta_*^+)^2)\lambda_{\text{new},k}^-$$

(c) Finally, using items 3 and 1, conditioned on $X_{j,k-1}$, $0 \preceq Z_t \preceq c\gamma_{\text{new},k}^2 I \le c\max((1.2)^{2k}\gamma_{\text{new}}^2, \gamma_*^2)I$ holds w.p. one for all $X_{j,k-1} \in \Gamma_{j,k-1}$.

Thus, applying Corollary 2.3.4 with $\epsilon = \frac{c\zeta\lambda^-}{24}$, we get

$$\mathbf{P}(\lambda_{\min}(\frac{1}{\alpha}\sum_t Z_t) \geq (1-(\zeta_*^+)^2)\lambda_{\text{new},k}^- - \frac{c\zeta\lambda^-}{24}|X_{j,k-1}) \geq 1 - c\exp(-\frac{\alpha\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot\min(1.2^{4k}\gamma_{\text{new}}^4,\gamma_*^4)}) \text{ for all } X_{j,k-1}\in\Gamma_{j,k-1} \quad (C.2)$$

Consider $Y_t = R_{\text{new}}a_{t,\text{new}}a_{t,*}{}'D_*{}'E_{\text{new}} + E_{\text{new}}{}'D_*a_{t,*}a_{t,\text{new}}{}'R_{\text{new}}{}'$. (a) Using item 8, the Y_t's are conditionally independent given $X_{j,k-1}$. (b) Using items 3 and 1, $\mathbf{E}(\frac{1}{\alpha}\sum_t Y_t|X_{j,k-1}) = 0$ for all $X_{j,k-1}\in\Gamma_{j,k-1}$. (c) Using items 1, 3, 4 and Fact C.2.1, conditioned on $X_{j,k-1}$, $\|Y_t\| \leq 2\sqrt{cr}\zeta_*^+\gamma_*\gamma_{\text{new},k} \leq 2\sqrt{cr}\zeta_*^+\gamma_*^2 \leq 2$ 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 2.3.4 with $\epsilon = \frac{c\zeta\lambda^-}{24}$, we get

$$\mathbf{P}(\lambda_{\min}(\frac{1}{\alpha}\sum_t Y_t) \geq -\frac{c\zeta\lambda^-}{24}|X_{j,k-1}) \geq 1 - c\exp(-\frac{\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot(2b)^2}) \text{ for all } X_{j,k-1}\in\Gamma_{j,k-1} \quad (C.3)$$

Combining (C.1), (C.2) and (C.3) and using the union bound, $\mathbf{P}(\lambda_{\min}(A_k) \geq \lambda_{\text{new},k}^-(1-(\zeta_*^+)^2) - \frac{c\zeta\lambda^-}{12}|X_{j,k-1}) \geq 1 - p_a(\alpha,\zeta)$ for all $X_{j,k-1}\in\Gamma_{j,k-1}$. The first claim of the lemma follows by using $\lambda_{\text{new},k}^- \geq \lambda^-$ and then applying Lemma 2.3.1 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 3, $E_{\text{new},\perp}{}'\Phi_0 L_t = E_{\text{new},\perp}{}'D_* a_{t,*}$. 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 $(n-c)\times(n-c)$. (a) As before, given $X_{j,k-1}$, the Z_t's are independent. (b) Using items 4, 1 and Fact C.2.1, conditioned on $X_{j,k-1}$, $0 \preceq Z_t \preceq r(\zeta_*^+)^2\gamma_*^2 I \preceq \zeta I$ w.p. one for all $X_{j,k-1}\in\Gamma_{j,k-1}$. (c) Using items 3, 2, $\mathbf{E}(\frac{1}{\alpha}\sum_t Z_t|X_{j,k-1}) \preceq (\zeta_*^+)^2\lambda^+ I$.

Thus applying Corollary 2.3.4 with $\epsilon = \frac{c\zeta\lambda^-}{24}$, we get

$$\mathbf{P}(\lambda_{\max}(A_{k,\perp}) \leq (\zeta_*^+)^2\lambda^+ + \frac{c\zeta\lambda^-}{24}|X_{j,k-1}) \geq 1-(n-c)\exp(-\frac{\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2\zeta}) \text{ for all } X_{j,k-1}\in\Gamma_{j,k-1}$$

The second claim follows using $\lambda_{\text{new},k}^- \geq \lambda^-$ and $f = \lambda^+/\lambda^-$ in the above expression and then applying Lemma 2.3.1. ■

C.7 Proof of Lemma 4.4.15

proof: In this proof, we frequently refer to items from Sec. C.5.

The first claim of the lemma follows using item 5 of Sec. C.5 and Lemma 2.3.1.

For the second claim, using the expression for $\mathcal{H}_k$ given in Definition 4.4.6, it is easy to see that

$$\|\mathcal{H}_k\|_2 \leq \max\{\|H_k\|_2, \|H_{k,\perp}\|_2\} + \|B_k\|_2 \leq \|\frac{1}{\alpha}\sum_t e_t {e_t}'\|_2 + \max(\|T2\|_2, \|T4\|_2) + \|B_k\|_2 \tag{C.4}$$

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) $\max(\|T1\|_2, \|T3\|_2) \leq \|\frac{1}{\alpha}\sum_t e_t {e_t}'\|_2$. Next, we obtain high probability bounds on each of the terms on the RHS of (C.4) using the Hoeffding corollaries.

Consider $\|\frac{1}{\alpha}\sum_t e_t {e_t}'\|_2$. Let $Z_t = e_t {e_t}'$. (a) Using item 8, conditioned on $X_{j,k-1}$, the various Z_t's in the summation are independent, for all $X_{j,k-1} \in \Gamma_{j,k-1}$. (b) Using items 1, 2, 4, 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 $b_1 := (\kappa_s^+ \zeta_{k-1}^+ \phi^+ \sqrt{c}\gamma_{\text{new},k} + \zeta_*^+ \phi^+ \sqrt{r}\gamma_*)^2$. (c) Using items 1, 2, 4, $0 \preceq \frac{1}{\alpha}\sum_t \mathbf{E}(Z_t | X_{j,k-1}) \preceq b_2 I$, $b_2 := (\kappa_s^+)^2 (\zeta_{k-1}^+)^2 (\phi^+)^2 \lambda_{\text{new},k}^+ + (\zeta_*^+)^2 (\phi^+)^2 \lambda^+$ for all $X_{j,k-1} \in \Gamma_{j,k-1}$.

Thus, applying Corollary 2.3.4 with $\epsilon = \frac{c\zeta\lambda^-}{24}$,

$$\mathbf{P}(\|\frac{1}{\alpha}\sum_t e_t {e_t}'\|_2 \leq b_2 + \frac{c\zeta\lambda^-}{24} | X_{j,k-1}) \geq 1 - n\exp(-\frac{\alpha c^2 \zeta^2 (\lambda^-)^2}{8 \cdot 24^2 b_1^2}) \text{ for all } X_{j,k-1} \in \Gamma_{j,k-1} \tag{C.5}$$

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 $T2 = \frac{1}{\alpha}\sum_t Z_t$. (a) Using item 8, 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 items 2 and 3, ${E_{\text{new}}}'\Phi_0 L_t = R_{\text{new}} a_{t,\text{new}} + {E_{\text{new}}}' D_* a_{t,*}$ and ${E_{\text{new}}}'\Phi_0 e_t = ({R_{\text{new}}}')^{-1} {D_{\text{new}}}' e_t$. (b) Thus, using items 2, 3, 4, 1, it follows that conditioned on $X_{j,k-1}$, $\|Z_t\|_2 \leq 2\tilde{b}_3 \leq 2b_3$ 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 $b_3 := \frac{1}{\sqrt{1-(\zeta_*^+)^2}}(\phi^+ c {\kappa_s^+}^2 \zeta_{k-1}^+ \gamma_{\text{new},k}^2 + \phi^+ \sqrt{rc}{\kappa_s^+}^2 \zeta_{k-1}^+ \zeta_*^+ \gamma_{\text{new},k}\gamma_* + \phi^+\sqrt{rc}\kappa_s^+ \zeta_*^+ \gamma_* \gamma_{\text{new},k} + \phi^+ r {\zeta_*^+}^2 \gamma_*^2)$.
(c) Also, $\|\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t | X_{j,k-1})\|_2 \leq 2\tilde{b}_4 \leq 2b_4$ 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 2.3.5 with $\epsilon = \frac{c\zeta\lambda^-}{24}$,

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

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 $T4 = \frac{1}{\alpha}\sum_t Z_t$. (a) Using item 8, 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 items 2, 3, $E_{\text{new},\perp}{}'\Phi_0 L_t = E_{\text{new},\perp}{}'D_* a_{t,*}$. (b) 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 2, 4, 1. (c) 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 2.3.5 with $\epsilon = \frac{c\zeta\lambda^-}{24}$,

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

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 $1-(n-c)\exp(-\frac{\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot 4b_5^2}) > 1-(n-c)\exp(-\frac{\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2\cdot 4b_3^2})$. Therefore, for all $X_{j,k-1}\in\Gamma_{j,k-1}$,

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

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

$$\mathbf{P}(\max(\|T2\|_2, \|T4\|_2) \le 2b_4 + \frac{c\zeta\lambda^-}{24}|X_{j,k-1}) \ge 1 - n\exp(-\frac{\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2\cdot 4b_3^2}) \qquad \text{(C.6)}$$

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 $B_k = \frac{1}{\alpha}\sum_t Z_t$. (a) Using item 8, 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 items 2, 3, $E_{\text{new},\perp}{}'\Phi_0(L_t - e_t) = E_{\text{new},\perp}{}'(D_* a_{t,*} - \Phi_0 e_t)$, $E_{\text{new}}{}'\Phi_0(L_t - e_t) = R_{\text{new}} a_{t,\text{new}} + E_{\text{new}}{}'D_* a_{t,*} + (R_{\text{new}}')^{-1}D_{\text{new}}' e_t$. (b) Thus, conditioned on $X_{j,k-1}$, $\|Z_t\|_2 \le b_7$ w.p. one for all $X_{j,k-1}\in\Gamma_{j,k-1}$. Here $b_7 := (\sqrt{r}\zeta_*^+(1+\phi^+)\gamma_* + (\kappa_s^+)\zeta_{k-1}^+\phi^+\sqrt{c}\gamma_{\text{new},k})(\sqrt{c}\gamma_{\text{new},k} + \sqrt{r}\zeta_*^+(1+\frac{1}{\sqrt{1-(\zeta_*^+)^2}}\kappa_s^+\phi^+)\gamma_* + \frac{1}{\sqrt{1-(\zeta_*^+)^2}}{\kappa_s^+}^2\zeta_{k-1}^+\phi^+\sqrt{c}\gamma_{\text{new},k})$. This follows using items 2, 3, 4, 1.
(c) Also, $\|\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X_{j,k-1})\|_2 \le b_8$ where $b_8 := (\kappa_s^+\zeta_{k-1}^+\phi^+ + \frac{1}{\sqrt{1-(\zeta_*^+)^2}}(\kappa_s^+)^3(\zeta_{k-1}^+)^2(\phi^+)^2)$

$\lambda^+_{\text{new},k} + (\zeta^+_*)^2(1+\phi^+ + \frac{1}{\sqrt{1-(\zeta^+_*)^2}}\kappa^+_s\phi^+ + \frac{1}{\sqrt{1-(\zeta^+_*)^2}}\kappa^+_s(\phi^+)^2)\lambda^+$ for all $X_{j,k-1} \in \Gamma_{j,k-1}$. Thus, applying Corollary 2.3.5 with $\epsilon = \frac{c\zeta\lambda^-}{24}$,

$$\mathbf{P}(\|B_k\|_2 \leq b_8 + \frac{c\zeta\lambda^-}{24}|X_{j,k-1}) \geq 1 - n\exp(-\frac{\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2 b_7^2}) \text{ for all } X_{j,k-1}\in\Gamma_{j,k-1} \tag{C.7}$$

Using (C.4), (D.9), (D.10) and (D.11) and the union bound, for any $X_{j,k-1}\in\Gamma_{j,k-1}$,

$$\mathbf{P}(\|\mathcal{H}_k\|_2 \leq b_9 + \frac{c\zeta\lambda^-}{8}|X_{j,k-1}) \geq 1 - n\exp(-\frac{\alpha c^2\zeta^2(\lambda^-)^2}{8\cdot 24^2 b_1^2}) - n\exp(-\frac{\alpha c^2\zeta^2(\lambda^-)^2}{32\cdot 24^2\cdot 4b_3^2}) \\ -n\exp(-\frac{\alpha c^2\zeta^2(\lambda^-)^2\epsilon^2}{32\cdot 24^2 b_7^2}) \tag{C.8}$$

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

$$\begin{aligned} b_9 =& ((\frac{2(\kappa^+_s)^2\phi^+}{\sqrt{1-(\zeta^+_*)^2}} + \kappa^+_s\phi^+)\zeta^+_{k-1} + ((\kappa^+_s)^2(\phi^+)^2 + \frac{(\kappa^+_s)^3(\phi^+)^2}{\sqrt{1-(\zeta^+_*)^2}})(\zeta^+_{k-1})^2)\lambda^+_{\text{new},k} \\ &+((\phi^+)^2 + \frac{2\phi^+}{\sqrt{1-(\zeta^+_*)^2}} + 1 + \phi^+ + \frac{\kappa^+_s\phi^+}{\sqrt{1-(\zeta^+_*)^2}} + \frac{\kappa^+_s(\phi^+)^2}{\sqrt{1-(\zeta^+_*)^2}})(\zeta^+_*)^2\lambda^+ \\ =& C(\zeta^+_{k-1};\zeta^+_*)\lambda^+_{\text{new},k} + O(\zeta^+_*, \zeta^+_* f)\lambda^+ \end{aligned} \tag{C.9}$$

where $C(x;u,v)$ and $O(u,v)$ are defined in Definition 4.4.13. Using $\lambda^-_{\text{new},k} \geq \lambda^-$ and $f := \lambda^+/\lambda^-$, $b_9 + \frac{c\zeta\lambda^-}{8} \leq \lambda^-_{\text{new},k} g_{inc}(\zeta^+_{k-1};\zeta^+_*, \zeta^+_* f, c\zeta)$. Using Fact C.2.1 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 2.3.1, the result follows. ∎

C.8 Proof of Lemma 4.4.18

proof: Conditions 2, 4 of Theorem 4.3.1 imply that $\kappa_{2s,*} \leq \kappa^+_{2s,*} = 0.3$, $\kappa_{2s,\text{new}} \leq \kappa^+_{2s,\text{new}} = 0.15$, $\tilde{\kappa}_{2s,k} \leq \tilde{\kappa}^+_{2s} = 0.15$, $\kappa_{s,k} \leq \kappa^+_s = 0.15$ and $g_{j,k} \leq g^+ = \sqrt{2}$. Using Lemma 4.4.10, this implies that $\phi_k \leq \phi^+ = 1.1735$. Using Fact C.2.1, $\zeta^+_* \leq 10^{-4}$; $\zeta^+_* f \leq 1.5\times 10^{-4}$; and $c\zeta \leq 10^{-4}$.

1. By definition, $\zeta^+_0 = 1$. We prove the first claim by induction.
   - Base case: For $k=1$, $\zeta^+_1 = f_{inc}(1;\zeta^+_*, \zeta^+_* f, c\zeta) \leq f_{inc}(1; 10^{-4}, 1.5\times 10^{-4}, 10^{-4})$ $< 0.5985 < 1 = \zeta^+_0$.

- Induction step: Assume that $\zeta_{k-1}^+ \leq \zeta_{k-2}^+$ for $k > 1$. Since f_{inc} is an increasing function of its arguments, $\zeta_k^+ = f_{inc}(\zeta_{k-1}^+; \zeta_*^+, \zeta_*^+ f, c\zeta) \leq f_{inc}(\zeta_{k-2}^+; \zeta_*^+, \zeta_*^+ f, c\zeta) = \zeta_{k-1}^+$.

2. For the second claim, let $\theta_a(x; u, v, w) := \frac{1}{x} \frac{C(x;u) g^+}{g_{dec}(x;u,v,w)}$ and $\theta_b(x; u, v, w) := \frac{1}{c\zeta} \frac{O(u,v) f + 0.125 w}{g_{dec}(x;u,v,w)}$. Then, $f_{inc}(x; u, v, w) = \theta_a(x; u, v, w) x + \theta_b(x, u, v, w) c\zeta$.

   - Notice that θ_a, θ_b are also increasing functions of all their arguments. Thus, $\theta_a(\zeta_{k-1}^+; \zeta_*^+, \zeta_*^+ f, c\zeta) \leq \theta_a(0.5985; 10^{-4}, 1.5 \times 10^{-4}, 10^{-4}) \approx 0.4471 < 0.6$ and $\theta_b(\zeta_{k-1}^+; \zeta_*^+, \zeta_*^+ f, c\zeta) \leq \theta_b(0.5985; 10^{-4}, 1.5 \times 10^{-4}, 10^{-4}) = 0.1598 < 0.16$. Thus,

$$\begin{aligned} \zeta_k^+ &= \theta_a(\zeta_{k-1}^+; \zeta_*^+, \zeta_*^+ f, c\zeta)\zeta_{k-1}^+ + \theta_b(\zeta_{k-1}^+; \zeta_*^+, \zeta_*^+ f, c\zeta) c\zeta \\ &\leq 0.6\zeta_{k-1}^+ + 0.16c\zeta \\ &\leq 0.6^{k-1}\zeta_1^+ + (0.6^{k-2} + 0.6^{k-3} + \cdots + 1)0.16c\zeta \\ &\leq 0.6^k + \frac{0.16c\zeta}{1-0.6} = 0.6^k + 0.4c\zeta \end{aligned} \tag{C.10}$$

3. Since $\zeta_k^+ \leq 0.5985$ and g_{dec} is a decreasing function of its arguments, $g_{dec}(\zeta_k^+; \zeta_*^+, \zeta_*^+ f, c\zeta) \geq g_{dec}(0.5985; 10^{-4}, 1.5 \times 10^{-4}, 10^{-4}) > 0$.

■

C.9 Proof of Lemma 4.4.21

proof: By Lemma 4.4.18, ζ_k^+ defined in Definition 4.4.17 satisfies $\zeta_k^+ \leq 0.6^k + 0.4c\zeta$ for all $k \leq K$ and $g_{dec}(\zeta_k^+; \zeta_*^+, \zeta_*^+ f, c\zeta) > 0$. Thus, we can apply Lemma 4.4.16 and Lemma 4.4.15. By Lemma 4.4.16, $\mathbf{P}(\zeta_k \leq \zeta_k^+ | \Gamma_{j,k-1}^e) \geq p_k(\alpha, \zeta)$. By Lemma 4.4.15, $\mathbf{P}(\{\hat{T}_t = T_t \text{ and } e_t \text{ satisfies (4.2) for all } t \in \mathcal{I}_{j,k}\} | \Gamma_{j,k-1}^e) = 1$. Combining these two facts, $\mathbf{P}(\tilde{\Gamma}_{j,k}^e | \Gamma_{j,k-1}^e) \geq p_k(\alpha, \zeta)$ for all $1 \leq k \leq K$.

Since $\Gamma_{j,K}^e$ holds and since $\zeta_k^+ \leq 0.6^k + 0.4c\zeta$ for all $k \leq K$, thus $\zeta_* \leq \zeta_*^+$ and $\zeta_K \leq \zeta_K^+ \leq 0.6^K + 0.4c\zeta$. This is proved in Sec. C.5 (item 4). Using this and applying Lemma 4.4.11, the last claim follows. ■

APPENDIX D. Proof of the Lemmas in Chapter 5

D.1 Proof of Lemma 5.4.15

The proof follows by using the following three lemmas.

Lemma D.1.1 (Exponential decay of ζ_k^+) *Assume that all the conditions of Theorem 5.3.1 hold. Let $\zeta_*^+ = r\zeta$. Define the series ${\zeta_k}^+$ as in Definition 5.4.3. Then,*

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

proof This lemma is the same as Lemma 4.4.18 but with ζ_*^+ defined differently. ∎

Lemma D.1.2 (Sparse recovery, support recovery and expression for e_t) *Assume that all conditions of Theorem 5.3.1 hold.*

1. *If $\zeta_* \leq \zeta_*^+ := r\zeta$ and $\zeta_{k-1} \leq \zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$, then for all $t \in \mathcal{I}_{j,k}$, for any $k = 1, 2, \ldots K$,*
   (a) *the projection noise β_t satisfies $\|\beta_t\|_2 \leq \zeta_{k-1}^+ \sqrt{c}\gamma_{new,k} + \zeta_*^+ \sqrt{r}\gamma_* \leq \sqrt{c}0.72^{k-1}\gamma_{new} + 1.06\sqrt{\zeta} \leq \xi$.*
   (b) *the CS error satisfies $\|\hat{S}_{t,cs} - S_t\|_2 \leq 7\xi$.*
   (c) *$\hat{T}_t = T_t$*
   (d) *e_t satisfies (5.3) and $\|e_t\|_2 \leq \phi^+[\kappa_s^+ \zeta_{k-1}^+ \sqrt{c}\gamma_{new,k} + \zeta_*^+ \sqrt{r}\gamma_*] \leq 0.18 \cdot 0.72^{k-1}\sqrt{c}\gamma_{new} + 1.17 \cdot 1.06\sqrt{\zeta}$*
2. *For all $k = 1, 2, \ldots K$, $\mathbf{P}(\hat{T}_t = T_t$ and e_t satisfies (5.3) for all $t \in \mathcal{I}_{j,k} | X_{j,k-1,0}) = 1$ for all $X_{j,k-1,0} \in \Gamma_{j,k-1,0}$.*

3. *For all* $k = 1, 2, \dots K$, $\mathbf{P}(\hat{T}_t = T_t$ *and* e_t *satisfies (5.3) for all* $t \in \mathcal{I}_{j,k} | \Gamma^e_{j,k-1,0}) = 1$.

proof The first claim is the same as Lemma 4.4.11 but with ζ_*^+ defined differently. The proof follows in an analogous fashion. The second claim follows from the first using Remark 5.4.17. The third claim follows using Lemma 2.3.1. ■

Lemma D.1.3 (High probability bound on ζ_k) *Assume that all the conditions of Theorem 5.3.1 hold. Let* $\zeta_*^+ = r\zeta$. *Then, for all* $k = 1, 2, \dots K$,

$$\mathbf{P}(\zeta_k \leq \zeta_k^+ | \Gamma^e_{j,k-1,0}) \geq p_k(\alpha, \zeta)$$

where ζ_k^+ *is defined in Definition 5.4.3 and* $p_k(\alpha, \zeta)$ *is defined in Lemma 4.4.16.*

proof Using Lemma D.1.1, (i) $\zeta_0^+ = 1$ and $\zeta_{k-1}^+ \leq 0.6^{k-1} + 0.4c\zeta$ and (ii) the denominator of ζ_k^+ is positive. Using this and the theorem's conditions, the above lemma follows exactly as in Lemma D.1.1. The only difference is that ζ_*^+ is defined differently. Also, $\Gamma_{j,k} := \Gamma_{j,k,0}$. The proof proceeds by first bounding ζ_k (in a fashion similar to the bound in Lemma D.2.6); using Lemma D.1.2 to get an expression for e_t; and finally using Corollaries 2.3.4 and 2.3.5 to get high probability bounds on each of the terms in the bound on ζ_k. ■

Lemma 5.4.15 follows by combining Lemma D.1.3 and the third claim of Lemma D.1.2 and using the fact that

$$\mathbf{P}(\Gamma^e_{j,k,0} | \Gamma^e_{j,k-1,0}) = \mathbf{P}(\zeta_k \leq \zeta_k^+, \ \hat{T}_t = T_t \text{ and } e_t \text{ satisfies (5.3) for all } t \in \mathcal{I}_{j,k} | \Gamma^e_{j,k-1,0})$$

D.2 Lemmas used to prove Lemma 5.4.16

In this section, we remove the subscript j at most places. The convention of Remark 5.4.14 applies.

D.2.0.1 Showing exact support recovery and getting an expression for e_t

Lemma D.2.1 (Bounding the RIC of Φ_k) *The following hold.*

1. $\delta_s(\Phi_0) = \kappa_s^2(\hat{P}_*) \leq \kappa_{s,*}^2 + 2\zeta_*$

2. $\delta_s(\Phi_k) = \kappa_s^2([\hat{P}_* \ \hat{P}_{new,k}]) \leq \kappa_s^2(\hat{P}_*) + \kappa_s^2(\hat{P}_{new,k}) \leq \kappa_{s,*}^2 + 2\zeta_* + (\kappa_{s,new} + \tilde{\kappa}_{s,k}\zeta_k + \zeta_*)^2$ *for* $k = 1, 2 \dots K$

proof The above lemma is the same as the last two claims of Lemma D.2.1. It follows using Lemma 3.3.2 and some linear algebraic manipulations. ■

Lemma D.2.2 (Sparse recovery, support recovery and expression for e_t) *Assume that the conditions of Theorem 5.3.1 hold.*

1. *For all* $k = 1, 2, \dots \vartheta + 1$, $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ *implies that*
   (a) $\zeta_* \leq \zeta_*^+ := r\zeta$, $\zeta_K \leq c\zeta$, $\|\Phi_K P_j\|_2 \leq (r + c)\zeta$,
   (b) $\delta_s(\Phi_K) \leq 0.1479$ *and* $\phi_K \leq \phi^+ := 1.1735$
   (c) *for any* $t \in \tilde{\mathcal{I}}_{j,k}$,
      i. *the projection noise* $\beta_t := (I - \hat{P}_{(t-1)}\hat{P}'_{(t-1)})L_t$ *satisfies* $\|\beta_t\|_2 \leq \sqrt{\zeta}$,
      ii. *the CS error satisfies* $\|\hat{S}_{t,cs} - S_t\|_2 \leq 7\sqrt{\zeta}$,
      iii. $\hat{T}_t = T_t$,
      iv. e_t *satisfies (5.3) and* $\|e_t\|_2 \leq \phi^+\sqrt{\zeta}$.
2. *For all* $k = 1, 2, \dots \vartheta + 1$, $\mathbf{P}(T_t = \hat{T}_t$ *and* e_t *satisfies (5.3) for all* $t \in \tilde{\mathcal{I}}_{j,k} \,|X_{j,K,k-1}) = 1$ *for all* $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.
3. *For all* $k = 1, 2, \dots \vartheta + 1$, $\mathbf{P}(T_t = \hat{T}_t$ *and* e_t *satisfies (5.3) for all* $t \in \tilde{\mathcal{I}}_{j,k} \,|\Gamma^e_{j,K,k-1}) = 1$.

proof

Claim 1-a follows using Remark 5.4.17. Claim 1-b) follows using claim 1-a) and Lemma D.2.1. Claim 1-c) follows in a fashion similar to the proof of Lemma 4.4.11. The main difference is that everywhere we use $\Phi_K L_t = \Phi_K P_j a_t$ and $\|\Phi_K P_j\|_2 \leq (r + c)\zeta$. Claim 1-c-i) uses this and the fact that for $t \in \tilde{\mathcal{I}}_{j,k}$, $\Phi_{(t)} = \Phi_K$, and $\sqrt{\zeta} \leq \sqrt{\gamma_*^2/(r + c)^3}$. Claim 1-c-ii) uses c-i), $\sqrt{\zeta} \leq \xi$ (defined in the theorem), $\delta_{2s}(\Phi_K) \leq 0.1479$, and Theorem 2.1.1. Claim 1-c-iii) uses c-ii), the definition of ρ, the choice of ω and the lower bound on $S_{\min}$ given in the theorem.

Claim 1-c-iv) uses claim c-iii) and Remark 5.4.11. To get the bound on $\|e_t\|_2$ we use the first expression of (5.3), $\phi_K \leq \phi^+ := 1.1735$, and $\sqrt{\zeta} \leq \sqrt{\gamma_*^2/(r+c)^3}$.

Claim 2) is just a rewrite of claim 1). Claim 3) follows from claim 2) by Lemma 2.3.1.

■

D.2.1 A lemma needed for bounding the subspace error, $\tilde{\zeta}_k$

Lemma D.2.3 *Assume that $\tilde{\zeta}_{k'} \leq \tilde{c}_{k'}\zeta$ for $k' = 1, \cdots, k-1$. Then*

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

proof The first claim essentially follows by using the fact that $\hat{G}_1, \cdots, \hat{G}_{k-1}$ are mutually orthonormal and triangle inequality. Recall that $\Psi_{k-1} = (I - \hat{G}_{\text{det},k} \hat{G}'_{\text{det},k})$. The last three claims use this and the first claim and apply Lemma 2.2.4. The last claim also uses the definition of D_k and its QR decomposition. ■

D.2.2 Bounding on the subspace error, $\tilde{\zeta}_k$

Lemma D.2.4 (Bounding $\tilde{\zeta}_k{}^+$) *If*

$$f_{dec}(\tilde{g}_{\max}, \tilde{h}_{\max}) - \frac{f_{inc}(\tilde{g}_{\max}, \tilde{h}_{\max})}{\tilde{c}_{\min}\zeta} > 0 \tag{D.1}$$

then $f_{dec}(\tilde{g}_k, \tilde{h}_k) > 0$ *and* $\tilde{\zeta}_k^+ \leq \tilde{c}_k\zeta$.

proof Recall that $f_{inc}(.)$, $f_{dec}(.)$ are defined in Definition 5.4.3 and $\tilde{\zeta}_k{}^+ := \frac{f_{inc}(\tilde{g},\tilde{h})}{f_{dec}(\tilde{g},\tilde{h})}$. Notice that $f_{inc}(.)$ is a non-decreasing function of $\tilde{g}, \tilde{h}$, and $f_{dec}(.)$ is a non-increasing function. Using the definition of $\tilde{g}_{\max}, \tilde{h}_{\max}, \tilde{c}_{\min}$ given in Assumption 5.1.1, the result follows. ■

Remark D.2.5 *If we ignore the small terms of $f_{inc}(.)$ and $f_{dec}(.)$, the above condition simplifies to requiring that $\frac{3\kappa_{s,e}^+ \phi^+ \tilde{g}_{\max} + \kappa_{s,e}^+ \phi^+ \tilde{h}_{\max}}{1-\tilde{h}_{\max}} \leq \frac{\tilde{c}_{\min}}{r+c}$. Since $\tilde{g}_{\max} \geq 1$, the first term of the numerator is the largest one. To ensure that this condition holds we need $\kappa_{s,e}^+$ to be very small. However, as explained in Sec D.2.3, if we also assume denseness of D_k, i.e. if we assume $\kappa_s(D_k) \leq \kappa_{s,D}^+$ for a small enough $\kappa_{s,D}^+$, then the first term of the numerator can be replaced by $\max(3\kappa_{s,e}^+ \kappa_{s,D}^+ \phi^+ \tilde{g}_{\max}, \kappa_{s,e}^+ \phi^+ \tilde{h}_{\max})$. This will relax the requirement on $\kappa_{s,e}^+$, e.g. now $\kappa_{s,e}^+ = \kappa_{s,D}^+ = 0.3$ will work.*

Lemma D.2.6 (Bounding $\tilde{\zeta}_k$) *If $\lambda_{\min}(\tilde{A}_k) - \lambda_{\max}(\tilde{A}_{k,\perp}) - \|\tilde{\mathcal{H}}_k\|_2 > 0$, then*

$$\tilde{\zeta}_k \leq \frac{\|\tilde{\mathcal{H}}_k\|_2}{\lambda_{\min}(\tilde{A}_k) - \lambda_{\max}(\tilde{A}_{k,\perp}) - \|\tilde{\mathcal{H}}_k\|_2} \tag{D.2}$$

proof Recall that $\tilde{A}_k$, $\tilde{A}_{k,\perp}$, $\tilde{\mathcal{H}}_k$ are defined in Definition 5.4.6. The result follows by using the fact that $\tilde{\zeta}_k = \|(I - \hat{G}_k \hat{G}_k')D_{j,k}\|_2 = \|(I - \hat{G}_k \hat{G}_k')E_k R_k\|_2 \leq \|(I - \hat{G}_k \hat{G}_k')E_k\|_2$ and applying Lemma 2.2.1 with $E \equiv E_k$ and $F \equiv \hat{G}_k$. ∎

Lemma D.2.7 (High probability bounds for each terms in the $\tilde{\zeta}_k$ bound and for $\tilde{\zeta}_k$) *Assume that the conditions of Theorem 5.3.1 hold. Also, assume that $\mathbf{P}(\Gamma_{j,K,k-1}^e) > 0$. Then, for all $1 \leq k \leq \vartheta_j$,*

1. *$\mathbf{P}(\lambda_{\min}(\tilde{A}_k) \geq \lambda_k^-(1 - r^2\zeta^2 - 0.1\zeta)|\Gamma_{j,K,k-1}^e) > 1 - \tilde{p}_1(\tilde{\alpha}, \zeta)$ with $\tilde{p}_1(\tilde{\alpha}, \zeta)$ given in (D.6).*

2. *$\mathbf{P}(\lambda_{\max}(\tilde{A}_{k,\perp}) \leq \lambda_k^-(\tilde{h}_k + r^2\zeta^2 f + 0.1\zeta)|\Gamma_{j,K,k-1}^e) > 1 - \tilde{p}_2(\tilde{\alpha}, \zeta)$ with $\tilde{p}_2(\tilde{\alpha}, \zeta)$ given in (D.7).*

3. *$\mathbf{P}(\|\tilde{\mathcal{H}}_k\|_2 \leq \lambda_k^- f_{inc}(\tilde{g}_k, \tilde{h}_k)\ |\Gamma_{j,K,k-1}^e) \geq 1 - \tilde{p}_3(\tilde{\alpha}, \zeta)$ with $\tilde{p}_3(\tilde{\alpha}, \zeta)$ given in (D.12).*

4. *$\mathbf{P}(\lambda_{\min}(\tilde{A}_k) - \lambda_{\max}(\tilde{A}_{k,\perp}) - \|\tilde{\mathcal{H}}_k\|_2 \geq \lambda_k^- f_{dec}(\tilde{g}_k, \tilde{h}_k)\ |\Gamma_{j,K,k-1}^e) \geq \tilde{p}(\tilde{\alpha}, \zeta) := 1 - \tilde{p}_1(\tilde{\alpha}, \zeta) - \tilde{p}_2(\tilde{\alpha}, \zeta) - \tilde{p}_3(\tilde{\alpha}, \zeta)$.*

5. *If $f_{dec}(\tilde{g}_k, \tilde{h}_k) > 0$, then $\mathbf{P}(\tilde{\zeta}_k \leq \tilde{\zeta}_k^+\ |\Gamma_{j,K,k-1}^e) \geq \tilde{p}(\tilde{\alpha}, \zeta)$*

proof Recall that $f_{inc}(.)$, $f_{dec}(.)$ and $\tilde{\zeta}_k^+$ are defined in Definition 5.4.3. The proof of the first three claims is given in Sec D.2.3. The fourth claim follows directly from the first three using the union bound on probabilities. The fifth claim follows from the fourth using Lemma D.2.6. ■

Lemma D.2.8 (High probability bound on $\tilde{\zeta}_k$) *Assume that the conditions of Theorem 5.3.1 hold. Then,*

$$\mathbf{P}(\tilde{\zeta}_k \leq \tilde{c}_k \zeta \; | \Gamma^e_{j,K,k-1}) \geq \tilde{p}(\tilde{\alpha}, \zeta)$$

proof This follows by combining Lemma D.2.4 and the last claim of Lemma D.2.7. ■

D.2.3 Proof of Lemma D.2.7

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 2.3.4 and 2.3.5.

1. The statement *"conditioned on r.v. X, the event $\mathcal{E}^e$ holds w.p. one for all $X \in \Gamma$" is equivalent to "$\mathbf{P}(\mathcal{E}^e|X) = 1$, for all $X \in \Gamma$"*. We often use the former statement in our proofs since it is often easier to interpret.

2. 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. $X_{j,K,k-1}$. 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 $Z_t = f_1(X_{j,K,k-1}) Y_t f_2(X_{j,K,k-1})$, Y_t is a sub-matrix of $a_t a_t'$ and $f_1(.)$ and $f_2(.)$ are functions of $X_{j,K,k-1}$. For instance, one of the terms while bounding $\lambda_{\min}(\mathcal{A}_k)$ is $\frac{1}{\tilde{\alpha}} \sum_t R_k a_{t,k} {a_{t,k}}' {R_k}'$.

3. $X_{j,K,k-1}$ 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 $X_{j,K,k-1}$, the Z_t's defined above are mutually independent.

4. All the terms that we bound for the third claim contain e_t. Using the second claim of Lemma D.2.2, conditioned on $X_{j,K,k-1}$, e_t satisfies (5.3) w.p. one whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Conditioned on $X_{j,K,k-1}$, 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 $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Thus, conditioned on $X_{j,K,k-1}$, the Z_t's for these terms are mutually independent, whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.

5. By Remark 5.4.17, $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ implies that $\zeta_* \leq r\zeta$, $\tilde{\zeta}_{k'} \leq c_{k'}\zeta$, for all $k' = 1, 2, \ldots k-1$, $\zeta_K \leq \zeta_K^+ \leq c\zeta$, (iv) $\phi_K \leq \phi^+$ (by Lemma D.2.2); (v) $\|\Phi_K P_j\|_2 \leq (r+c)\zeta$; and (vi) all conclusions of Lemma D.2.3 hold.

6. By the clustering assumption, $\lambda_k^- \leq \lambda_{\min}(\mathbf{E}(a_{t,k}{a_{t,k}}')) \leq \lambda_{\max}(\mathbf{E}(a_{t,k}{a_{t,k}}')) \leq \lambda_k^+$; $\lambda_{\max}(\mathbf{E}(a_{t,\text{det}}{a_{t,\text{det}}}')) \leq \lambda_1^+ = \lambda^+$; and $\lambda_{\max}(\mathbf{E}(a_{t,\text{undet}}{a_{t,\text{undet}}}')) \leq \lambda_{k+1}^+$.
   Also, $\lambda_{\max}(\mathbf{E}(a_t a_t')) \leq \lambda^+$.

7. By Weyl's theorem, for a sequence of matrices B_t, $\lambda_{\min}(\sum_t B_t) \geq \sum_t \lambda_{\min}(B_t)$ and $\lambda_{\max}(\sum_t B_t) \leq \sum_t \lambda_{\max}(B_t)$.

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

$$\tilde{A}_k \succeq \frac{1}{\tilde{\alpha}}\sum_t Z_t + \frac{1}{\tilde{\alpha}}\sum_t Y_t \tag{D.3}$$

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 $X_{j,K,k-1}$. (b) Using Ostrowoski's theorem and Lemma D.2.3, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\begin{aligned}
\lambda_{\min}(\mathbf{E}(\frac{1}{\tilde{\alpha}}\sum_t Z_t | X_{j,K,k-1})) &= \lambda_{\min}(R_k \frac{1}{\tilde{\alpha}}\sum_t \mathbf{E}(a_{t,k}{a_{t,k}}'){R_k}') \\
&\geq \lambda_{\min}(R_k {R_k}')\lambda_{\min}(\frac{1}{\tilde{\alpha}}\sum_t \mathbf{E}(a_{t,k}{a_{t,k}}')) \\
&\geq (1 - r^2\zeta^2)\lambda_k^-
\end{aligned}$$

(c) Finally, using $\|R_k\|_2 \leq 1$ and $\|a_{t,k}\|_2 \leq \sqrt{\tilde{c}_k}\gamma_*$, conditioned on $X_{j,K,k-1}$, $0 \preceq Z_t \preceq \tilde{c}_k\gamma_*^2 I$ holds w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.

Thus, applying Corollary 2.3.4 with $\epsilon = 0.1\zeta\lambda^-$, and using $\tilde{c}_k \le r$, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\mathbf{P}(\lambda_{\min}(\frac{1}{\tilde{\alpha}}\sum_t Z_t) \ge (1-r^2\zeta^2)\lambda_k^- - 0.1\zeta\lambda^- | X_{j,K,k-1}) \ge 1 - \tilde{c}_k \exp(-\frac{\tilde{\alpha}\epsilon^2}{8(\tilde{c}_k\gamma_*^2)^2}) \\ \ge 1 - r\exp(-\frac{\tilde{\alpha}\cdot(0.1\zeta\lambda^-)^2}{8r^2\gamma_*^4}) \quad \text{(D.4)}$$

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 $X_{j,K,k-1}$. (b) Since $\mathbf{E}[a_t] = 0$ and $\text{Cov}[a_t] = \Lambda_t$ is diagonal, $\mathbf{E}(\frac{1}{\alpha}\sum_t Y_t | X_{j,K,k-1}) = 0$ whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (c) Conditioned on $X_{j,K,k-1}$, $\|Y_t\|_2 \le 2\sqrt{\tilde{c}_k r}\gamma_*^2 r\zeta(1+\frac{r\zeta}{\sqrt{1-r^2\zeta^2}}) \le 2r^2\zeta\gamma_*^2(1+\frac{10^{-4}}{\sqrt{1-10^{-4}}}) \le \frac{2}{r}(1+\frac{10^{-4}}{\sqrt{1-10^{-4}}}) < 2.1$ holds w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. This follows because $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ 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 2.3.4 with $\epsilon = 0.1\zeta\lambda^-$, we get

$$\mathbf{P}(\lambda_{\min}(\frac{1}{\tilde{\alpha}}\sum_t Y_t) \ge -0.1\zeta\lambda^- | X_{j,K,k-1}) \ge 1 - r\exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8(4.2)^2}) \text{ for all } X_{j,K,k-1} \in \Gamma_{j,K,k-1} \quad \text{(D.5)}$$

Combining (D.3), (D.4) and (D.5) and using the union bound, $\mathbf{P}(\lambda_{\min}(\tilde{A}_k) \ge \lambda_k^-(1-r^2\zeta^2) - 0.2\zeta\lambda^- | X_{j,K,k-1}) \ge 1 - \tilde{p}_1(\tilde{\alpha},\zeta)$ for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ where

$$\tilde{p}_1(\tilde{\alpha},\zeta) := r\exp(-\frac{\tilde{\alpha}\cdot(0.1\zeta\lambda^-)^2}{8r^2\gamma_*^4}) + r\exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8(4.2)^2}) \quad \text{(D.6)}$$

The first claim of the lemma follows by using $\lambda_k^- \ge \lambda^-$ and applying Lemma 2.3.1 with $X \equiv X_{j,K,k-1}$ and $\mathcal{C} \equiv \Gamma_{j,K,k-1}$.

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 $X_{j,K,k-1}$, the Z_t's are independent. (b) Conditioned on $X_{j,K,k-1}$, $0 \preceq Z_t \preceq r\gamma_*^2 I$ w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (c) $\mathbf{E}(\frac{1}{\alpha}\sum_t Z_t | X_{j,K,k-1}) \preceq (\lambda_{k+1}^+ + r^2\zeta^2\lambda^+)I$ for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.

Thus applying Corollary 2.3.4 with $\epsilon = 0.1\zeta\lambda^-$ and using $\tilde{c}_k \geq \tilde{c}_{\min}$, we get

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

where

$$\tilde{p}_2(\tilde{\alpha},\zeta) := (n - \tilde{c}_{\min}) \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8r^2\gamma_*^4}) \tag{D.7}$$

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 2.3.1.

Consider the third claim. Using the expression for $\tilde{\mathcal{H}}_k$ given in Definition 5.4.6, it is easy to see that

$$\|\tilde{\mathcal{H}}_k\|_2 \leq \max\{\|\tilde{H}_k\|_2, \|\tilde{H}_{k,\perp}\|_2\} + \|\tilde{B}_k\|_2 \leq \|\frac{1}{\tilde{\alpha}}\sum_t e_t {e_t}'\|_2 + \max(\|T2\|_2, \|T4\|_2) + \|\tilde{B}_k\|_2 \tag{D.8}$$

with $T2 := \frac{1}{\tilde{\alpha}}\sum_t {E_k}'\Psi_{k-1}(L_t{e_t}' + e_t{L_t}')\Psi_{k-1}E_k$ 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 (D.8) 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 $X_{j,K,k-1}$, the various Z_t's in the summation are independent whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (b) Conditioned on $X_{j,K,k-1}$, $0 \preceq Z_t \preceq b_1 I$ w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Here $b_1 := {\phi^+}^2\zeta$. (c) Using $\|\Phi_K P_j\|_2 \leq (r+c)\zeta$, $0 \preceq \frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X_{j,K,k-1}) \preceq b_2 I$, $b_2 := (r+c)^2\zeta^2{\phi^+}^2\lambda^+$ for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$.

Thus, applying Corollary 2.3.4 with $\epsilon = 0.1\zeta\lambda^-$,

$$\mathbf{P}(\|\frac{1}{\tilde{\alpha}}\sum_t e_t{e_t}'\|_2 \leq b_2 + 0.1\zeta\lambda^- | X_{j,K,k-1}) \geq 1 - n\exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{8 \cdot b_1^2}) \text{ for all } X_{j,K,k-1} \in \Gamma_{j,K,k-1} \tag{D.9}$$

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 $X_{j,K,k-1}$, the various Z_t's used in the summation are mutually independent whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (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 $E_k{}'\Psi_{k-1}e_t = (R_k^{-1})'D_k' I_{T_t}[(\Phi_K)'_{T_t}(\Phi_K)_{T_t}]^{-1}I_{T_t}{}'\Phi_K P_j a_t$. Thus conditioned on $X_{j,K,k-1}$, $\|Z_t\|_2 \leq 2b_3$ w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. 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, $\|\frac{1}{\alpha}\sum_t \mathbf{E}(Z_t|X_{j,K,k-1})\|_2 \leq 2b_4$ where $b_4 := \kappa_{s,e}(r+c)\zeta\phi^+(\lambda_k^+ + r\zeta\lambda^+ + \frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}\lambda_{k+1}^+)$.

Thus, applying Corollary 2.3.5 with $\epsilon = 0.1\zeta\lambda^-$, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

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

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 $X_{j,K,k-1}$, the various Z_t's used in the summation are mutually independent whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (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 $X_{j,K,k-1}$, $\|Z_t\|_2 \leq 2b_5$ w.p. one for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. Here $b_5 := \sqrt{r\zeta}\phi^+\gamma_*$. (c) Also, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$, $\|\frac{1}{\tilde{\alpha}}\sum_t \mathbf{E}(Z_t|X_{j,K,k-1})\|_2 \leq 2b_6$, $b_6 := \kappa_{s,e}(r+c)\zeta\phi^+(\lambda_{k+1}^+ + r\zeta\lambda^+)$. Applying Corollary 2.3.5 with $\epsilon = 0.1\zeta\lambda^-$, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\begin{aligned}\mathbf{P}(\|T4\|_2 \leq 2b_6 + 0.1\zeta\lambda^-|X_{j,K,k-1}) &\geq 1 - (n-\tilde{c}_k)\exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32\cdot 4b_5^2}) \\ &\geq 1 - (n-\tilde{c}_{\min})\exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32\cdot 4b_5^2})\end{aligned}$$

Consider $\max(\|T2\|_2, \|T4\|_2)$. Since $b_3 = b_5$ and $b_4 > b_6$, so $2b_6 + \epsilon < 2b_4 + \epsilon$. Therefore, for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

$$\mathbf{P}(\|T4\|_2 \leq 2b_4 + 0.1\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - (n-\tilde{c}_k)\exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32\cdot 4b_3^2})$$

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

$$\mathbf{P}(\max(\|T2\|_2, \|T4\|_2) \leq 2b_4 + 0.1\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - n\exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32\cdot 4b_3^2}) \tag{D.10}$$

Notice that if we also introduce an extra denseness coefficient $\kappa_{s,D} := \max_j \max_k \kappa_s(D_k)$, then $\mathbf{P}(\|T2\|_2 \leq 2\kappa_{s,D}b_4 + 0.1\zeta\lambda^-|X_{j,K,k-1}) \geq 1 - \tilde{c}_k \exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32\cdot 4b_3^2})$. Thus,

$\mathbf{P}(\max(\|T2\|_2, \|T4\|_2) \leq 2\max(\kappa_{s,D} b_4, b_6) + 0.1\zeta\lambda^- | X_{j,K,k-1}) \geq 1 - n\exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot 4b_3^2})$. This would help to get a looser bounds on $\tilde{g}_{\max}$ and $\tilde{h}_{\max}$ in Theorem 5.3.1.

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 $X_{j,K,k-1}$, the various Z_t's used in the summation are mutually independent whenever $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$. (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 $X_{j,K,k-1}$, $\|Z_t\|_2 \leq b_7$ 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) $\|\frac{1}{\tilde{\alpha}}\sum_t \mathbf{E}(Z_t | X_{j,K,k-1})\|_2 \leq b_8$ for all $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$ where

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

Thus, applying Corollary 2.3.5 with $\epsilon = 0.1\zeta\lambda^-$,

$$\mathbf{P}(\|\tilde{B}_k\|_2 \leq b_8 + 0.1\zeta\lambda^- | X_{j,K,k-1}) \geq 1 - n\exp(-\frac{\tilde{\alpha}(0.1\zeta\lambda^-)^2}{32 \cdot b_7^2}) \text{ for all } X_{j,K,k-1} \in \Gamma_{j,K,k-1} \tag{D.11}$$

Using (D.8), (D.9), (D.10) and (D.11) and the union bound, for any $X_{j,K,k-1} \in \Gamma_{j,K,k-1}$,

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

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

$$\tilde{p}_3(\tilde{\alpha}, \zeta) := n\exp(-\frac{\tilde{\alpha}\epsilon^2}{8 \cdot b_1^2}) + n\exp(-\frac{\tilde{\alpha}\epsilon^2}{32 \cdot 4b_3^2}) + n\exp(-\frac{\tilde{\alpha}\epsilon^2}{32 \cdot b_7^2}) \tag{D.12}$$

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 2.3.1, the third claim of the lemma follows. ■

D.3 Proof of Lemma D.2.3

proof

1. The first claim follows because $\|D_{\text{det},k}\|_2 = \|\Psi_{k-1}G_{\text{det},k}\|_2 = \|\Psi_{k-1}[G_1G_2\cdots G_{k-1}]\|_2 \leq \sum_{k_1=1}^{k-1}\|\Psi_{k-1}G_{k_1}\|_2 \leq \sum_{k_1=1}^{k-1}\|\Psi_{k_1}G_{k_1}\|_2 = \sum_{k_1=1}^{k-1}\tilde{\zeta}_{k_1} \leq \sum_{k_1=1}^{k-1}\tilde{c}_{k_1}\zeta \leq r\zeta$. 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, $\|(I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_{\text{det},k}\|_2 = \|\Psi_{k-1}G_{\text{det},k}\|_2 \leq r\zeta$. By item 2) of Lemma 2.2.4 with $P = G_{\text{det},k}$ and $\hat{P} = \hat{G}_{\text{det},k}$, the result $\|G_{\text{det},k}G_{\text{det},k}{}' - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k}\|_2 \leq 2r\zeta$ follows.

3. Recall that $D_k \overset{QR}{=} E_kR_k$ 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 $\|(I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_{\text{det},k}\|_2 = \|\Psi_{k-1}G_{\text{det},k}\|_2 \leq r\zeta$ and $G'_kG_{\text{det},k} = 0$, by item 4) of Lemma 2.2.4 with $P = G_{\text{det},k}$, $\hat{P} = \hat{G}_{\text{det},k}$ and $Q = G_k$, we have $\sqrt{1 - r^2\zeta^2} \leq \sigma_i((I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_k) = \sigma_i(D_k) \leq 1$.

4. Since $D_k \overset{QR}{=} E_kR_k$, so $\|D_{\text{undet},k}{}'E_k\|_2 = \|D_{\text{undet},k}{}'D_kR_k^{-1}\|_2 = \|G_{\text{undet},k}{}'\Psi'_{k-1}\Psi_{k-1}G_kR_k^{-1}\|_2$ $= \|G_{\text{undet},k}{}'\Psi_{k-1}G_kR_k^{-1}\|_2 = \|G_{\text{undet},k}{}'D_kR_k^{-1}\|_2 = \|G_{\text{undet},k}{}'E_k\|_2$. Since $E_k = D_kR_k^{-1} = (I - \hat{G}_{\text{det},k}\hat{G}'_{\text{det},k})G_kR_k^{-1}$,

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

By item 3) of Lemma 2.2.4 with $P = G_{\text{det},k}$, $\hat{P} = \hat{G}_{\text{det},k}$ and $Q = G_{\text{undet},k}$, we get $\|G_{\text{undet},k}{}'\hat{G}_{\text{det},k}\|_2 \leq r\zeta$. By item 3) of Lemma 2.2.4 with $\hat{P} = \hat{G}_{\text{det},k}$ and $Q = G_k$, we get $\|\hat{G}'_{\text{det},k}G_k\|_2 \leq r\zeta$. Therefore, $\|G_{\text{undet},k}{}'E_k\|_2 = \|E_k{}'G_{\text{undet},k}\|_2 \leq \frac{r^2\zeta^2}{\sqrt{1-r^2\zeta^2}}$.

■

BIBLIOGRAPHY

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

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

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

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

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

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

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

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

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

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

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

[12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, no. 6, 2012.

[13] Y. Hu, S. Goud, and M. Jacob, "A fast majorize-minimize algorithm for the recovery of sparse and low-rank matrices," *IEEE Transactions on Image Processing*, vol. 21, no. 2, p. 742=753, Feb 2012.

[14] A. E. Waters, A. C. Sankaranarayanan, and R. G. Baraniuk, "Sparcs: Recovering low-rank and sparse matrices from compressive measurements," in *Proc. of Neural Information Processing Systems(NIPS)*, 2011.

[15] E. Richard, P.-A. Savalle, and N. Vayatis, "Estimation of simultaneously sparse and low rank matrices," *arXiv:1206.6474, appears in Proceedings of the 29th International Conference on Machine Learning (ICML 2012).*

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

[17] M. Mardani, G. Mateos, and G. B. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *arXiv:1204.6537*.

[18] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *arXiv:1202.4596*.

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

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

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

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

[23] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. iii," *SIAM Journal on Numerical Analysis*, Mar. 1970.

[24] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

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

[26] B. Nadler, "Finite sample approximation results for principal component analysis: A matrix perturbation approach," *The Annals of Statistics*, vol. 36, no. 6, 2008.

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

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

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

[30] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Oxford University Press, 2001.

[31] C. Qiu, N. Vaswani, and L. Hogben, "Recursive robust pca or recursive sparse recovery in large but structured noise," in *IEEE Intl. Conf. Acoustics, Speech, Sig. Proc. (ICASSP)*, 2013.

[32] ——, "Recursive robust pca or recursive sparse recovery in large but structured noise," *arXiv: 1211.3754[cs.IT], submitted to IEEE Tran. Info. Th., shortor version to appear in ICASSP 2013.*

[33] C. Qiu and N. Vaswani, "Recursive robust pca or recursive sparse recovery in large but structured noise," in *IEEE Intl. Symp. on Information Theory (ISIT)*, 2013.

[34] ——, "Recursive sparse recovery in large but structured noise - part 2," *arXiv:1303.1144[cs.IT], submitted to IEEE Tran. Info. Th., shortor version to appear in ISIT 2013.*

[35] G. Li and Z. Chen., "Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and monte carlo," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 759–766, 1985.

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

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

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

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

[40] N. Vaswani and W. Lu, "Modified-cs: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Signal Processing*, September 2010.