Correlated-PCA: Principal Components' Analysis when Data and Noise are Correlated

Namrata Vaswani and Han Guo
Iowa State University, Ames, IA, USA

Abstract

Given a matrix of observed data, Principal Components Analysis (PCA) computes a small number of orthogonal directions that contain most of its variability. Provably accurate solutions for PCA have been in use for decades. However, to the best of our knowledge, all existing theoretical guarantees for it assume that the data and the corrupting noise are mutually independent, or at least uncorrelated. This is valid in practice often, but not always. In this paper, we study the PCA problem in the setting where the data and noise can be correlated. Such noise is often referred to as "data-dependent noise". We obtain a correctness result for the standard eigenvalue decomposition (EVD) based solution to PCA under simple assumptions on the data-noise correlation. We also develop and analyze a generalization of EVD, called cluster-EVD, and argue that it reduces the sample complexity of EVD in certain regimes.

I. Introduction

Principal Components Analysis (PCA) is among the most frequently used tools for dimension reduction. Given a matrix of data, PCA computes a small number of orthogonal directions that contain all (or most) of the variability of the data. The subspace spanned by these directions is called the "principal subspace". In order to use PCA for dimension reduction, one projects the observed data onto this subspace.

The standard solution to PCA is to compute the reduced singular value decomposition (SVD) of the data matrix, or, equivalently, to compute the reduced eigenvalue decomposition (EVD) of the empirical covariance matrix of the data. If all eigenvalues are nonzero, a threshold is used and all eigenvectors with eigenvalues above the threshold are retained. This solution, which we henceforth refer to as *simple-EVD*, or sometimes just *EVD*, has been used for many decades and is well-studied in literature, e.g., see [1] and references therein. However, to the best of our knowledge, all existing results for it assume that the

A part of this paper is under submission to NIPS 2016.

true data and the corrupting noise in the observed data are independent, or, at least, uncorrelated. This is valid in practice often, but not always. In this paper we study the PCA problem in the setting where the data and noise vectors can be correlated (correlated-PCA). Such noise is sometimes referred to as "data-dependent" noise. Two example situations where this problem occurs are the PCA with missing data problem and a specific instance of the robust PCA problem described in Sec. I-B. A third example is the subspace update step of our recently proposed online dynamic robust PCA algorithm, ReProCS [2], [3], [4]. This is discussed in Sec. VII-A. These works inspired the current work.

Contributions. (1) We show that, under simple assumptions, for a fixed desired subspace error level, the sample complexity of simple-EVD for correlated-PCA scales as $f^2 r^2 \log n$ where n is the data vector length, f is the condition number of the true data covariance matrix and r is its rank. Here "sample complexity" refers to the number of samples needed to get a small enough subspace recovery error with high probability (whp). The dependence on f^2 is problematic for datasets with large condition numbers, and especially in the high dimensional setting when n itself is large. As we show in Sec. III-A, a large f is common. (2) To address this issue, we also develop a generalization of simple-EVD that we call *cluster-EVD*. Under an eigenvalues' "clustering" assumption, and under certain other mild assumptions, we argue that, cluster-EVD weakens the dependence on f. This assumption can be understood as a generalization of the eigen-gap condition needed by the block power method, which is a fast algorithm for obtaining the k top eigenvectors of a matrix [5], [6]. As we verify in Sec. III-A, the clustering assumption is valid for data that has variations across multiple scales. Common examples of such data include video textures such as moving waters or moving trees in a forest. (3) Finally, we also provide a guarantee for the problem of correlated-PCA with partial subspace knowledge, and we explain how this result can be used to significantly simplify the proof of correctness of ReProCS for online dynamic robust PCA given in [3].

Other somewhat related recent works include works such as [7], [8] that develop and study stochastic optimization based techniques for speeding up PCA; and works such [9], [10], [11], [12] that study incremental or online PCA solutions.

Notation. We use the interval notation $[a, b]$ to mean all of the integers between a and b, inclusive, and similarly for $[a, b)$ etc. We use $\mathcal{J}_u^\alpha$ to denote a time interval of length α beginning at $t = u\alpha$, i.e. $\mathcal{J}_u^\alpha := [u\alpha, (u+1)\alpha)$. For a set $\mathcal{T}$, $|\mathcal{T}|$ denotes its cardinality. We use $'$ to denote a vector or matrix transpose. The l_p-norm of a vector or the induced l_p-norm of a matrix are both denoted by $\|\cdot\|_p$. When the subscript p is missing, i.e., when we just use $\|.\|$, it denotes the l_2 norm of a vector or the induced l_2 norm of a matrix.

$\boldsymbol{I}$ denotes the identity matrix. The notation $\boldsymbol{I}_{\mathcal{T}}$ refers to an $n \times |\mathcal{T}|$ matrix of columns of the identity matrix indexed by entries in $\mathcal{T}$. For a matrix $\boldsymbol{A}$, $\boldsymbol{A}_{\mathcal{T}} := \boldsymbol{A}\boldsymbol{I}_{\mathcal{T}}$. For a Hermitian matrices $\boldsymbol{H}$ and $\boldsymbol{H}_2$, $\boldsymbol{H} \overset{\text{EVD}}{=} \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$ denotes its reduced eigenvalue decomposition with the eigenvalues in $\boldsymbol{\Lambda}$ arranged in non-increasing order; and the notation $\boldsymbol{H} \preceq \boldsymbol{H}_2$ means that $\boldsymbol{H}_2 - \boldsymbol{H}$ is positive semi-definite. $\mathrm{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{\Lambda}_3)$ defines a block diagonal matrix with blocks $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{\Lambda}_3$.

A tall matrix with orthonormal columns, i.e., a matrix $\boldsymbol{P}$ with $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$, is referred to a *basis matrix*. For a basis matrix $\boldsymbol{P}$, $\boldsymbol{P}_\perp$ denotes a basis matrix that is such that the square matrix $[\boldsymbol{P}\ \boldsymbol{P}_\perp]$ is unitary.

For basis matrices $\hat{\boldsymbol{P}}$ and $\boldsymbol{P}$, we quantify the subspace error (SE) between their range spaces using

$$\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) := \|(\boldsymbol{I} - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{P}\|. \tag{1}$$

Paper Organization. We describe the correlated-PCA problem next followed by example applications. In Sec. II, we give the performance guarantee for simple-EVD for correlated-PCA. We develop the cluster-EVD algorithm and its guarantee in Sec. III. The results are discussed in Sec. IV. We prove the two results in Sec. V and VI respectively. The main lemma used here is proved in Appendix A. In Sec. VII, we explain how the same ideas can be extended to correlated-PCA with partial subspace knowledge and to analyzing the subspace tracking step of ReProCS [2], [3], [4]. Numerical experiments are given in Sec. VIII. We conclude in Sec. IX.

A. Correlated-PCA: Problem Definition

We observe the data matrix $\boldsymbol{Y} := \boldsymbol{L} + \boldsymbol{W}$ of size $n \times m$, where $\boldsymbol{L}$ is the low-rank true data matrix and $\boldsymbol{W}$ is the noise matrix. We assume a column-wise linear model of correlation, i.e., each column $\boldsymbol{w}_t$ of $\boldsymbol{W}$ satisfies $\boldsymbol{w}_t = \boldsymbol{M}_t\boldsymbol{\ell}_t$. Thus,

$$\boldsymbol{y}_t = \boldsymbol{\ell}_t + \boldsymbol{w}_t, \text{ with } \boldsymbol{w}_t = \boldsymbol{M}_t\boldsymbol{\ell}_t \text{ for all } t = 1, 2, \dots, m \tag{2}$$

In Sec. III-D, we also give guarantees for a generalized version of this problem where there is another noise component, $\boldsymbol{\nu}_t$, that is independent of $\boldsymbol{\ell}_t$ and $\boldsymbol{w}_t$, i.e.,

$$\boldsymbol{y}_t = \boldsymbol{\ell}_t + \tilde{\boldsymbol{w}}_t, \text{ with } \tilde{\boldsymbol{w}}_t = \boldsymbol{w}_t + \boldsymbol{\nu}_t = \boldsymbol{M}_t\boldsymbol{\ell}_t + \boldsymbol{\nu}_t \tag{3}$$

The *goal* is to estimate the column space of $\boldsymbol{L}$ under the following assumptions on $\boldsymbol{\ell}_t$ and on the data-noise correlation matrix, $\boldsymbol{M}_t$.

Model 1.1 (Model on $\boldsymbol{\ell}_t$)**.** *The true data vectors, $\boldsymbol{\ell}_t$, are zero mean, mutually independent and bounded random vectors, with covariance matrix $\boldsymbol{\Sigma} \overset{\text{EVD}}{=} \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$ and with*

$$0 < \lambda^- \le \lambda_{\min}(\boldsymbol{\Lambda}) \le \lambda_{\max}(\boldsymbol{\Lambda}) \le \lambda^+ < \infty.$$

Let $f := \frac{\lambda^+}{\lambda^-}$ *and let* $r := \text{rank}(\boldsymbol{\Sigma})$.

Since the $\boldsymbol{\ell}_t$*'s are bounded, for all* $j = 1, 2, \ldots, r$*, there exists a constant* γ_j *such that* $\max_t |\boldsymbol{P}_j{}'\boldsymbol{\ell}_t| \leq \gamma_j$. *Define*

$$\eta := \max_{j=1,2,\ldots r} \frac{\gamma_j^2}{\lambda_j}.$$

Thus, $(\boldsymbol{P}_j{}'\boldsymbol{\ell}_t)^2 \leq \eta\lambda_j$. For most bounded distributions, η will be a small constant more than one. For example, if the distribution of all entries of $\boldsymbol{a}_t := \boldsymbol{P}'\boldsymbol{\ell}_t$ is iid zero mean uniform, then $\eta = 3$.

Model 1.2 (Model on $\boldsymbol{M}_t$, parameters: q, α, β)**.** *Decompose* $\boldsymbol{M}_t$ *as* $\boldsymbol{M}_t = \boldsymbol{M}_{2,t}\boldsymbol{M}_{1,t}$. *Assume that*

$$\|\boldsymbol{M}_{2,t}\| \leq 1, \ \|\boldsymbol{M}_{1,t}\boldsymbol{P}\| \leq q < 1, \tag{4}$$

Also, for any sequence of positive semi-definite Hermitian matrices, $\boldsymbol{A}_t$*, and for all time intervals* $\mathcal{J}_u^\alpha \subseteq [1, m]$*, the following holds with a* $\beta < \alpha$*:*

$$\left\| \frac{1}{\alpha} \sum_{t \in \mathcal{J}_u^\alpha} \boldsymbol{M}_{2,t}\boldsymbol{A}_t\boldsymbol{M}_{2,t}{}' \right\| \leq \frac{\beta}{\alpha} \max_{t \in \mathcal{J}_u^\alpha} \|\boldsymbol{A}_t\|. \tag{5}$$

We will need the above model to hold for all $\alpha \geq \alpha_0$ and for all $\beta \leq c_0\alpha$ with a $c_0 \ll 1$. We set α_0 and c_0 in Theorems 2.1 and 3.5; both will depend on q.

To understand the last assumption of this model, notice that, if we allow $\beta = \alpha$, then (5) will always hold and it is not an assumption. Let $\boldsymbol{B} = \frac{1}{\alpha}\sum_{t \in \mathcal{J}_u^\alpha} \boldsymbol{M}_{2,t}\boldsymbol{A}_t\boldsymbol{M}_{2,t}{}'$. One example situation when (5) will hold with a $\beta \ll \alpha$ is if $\boldsymbol{B}$ is block-diagonal with blocks $\boldsymbol{A}_t$. In this case, in fact, (5) will hold with $\beta = 1$. The matrix $\boldsymbol{B}$ will be of this form if $\boldsymbol{M}_{2,t} = \boldsymbol{I}_{\mathcal{T}_t}$ with all the sets $\mathcal{T}_t$ being mutually disjoint. This means that the matrices $\boldsymbol{A}_t$ will be of size $|\mathcal{T}_t| \times |\mathcal{T}_t|$, and that $n \geq \sum_{t \in \mathcal{J}_u^\alpha} |\mathcal{T}_t|$ (necessary condition for all the α sets, $\mathcal{T}_t$, to be mutually disjoint).

More generally, even if $\boldsymbol{B}$ is block-diagonal with blocks given by the summation of $\boldsymbol{A}_t$'s over at most $\beta_0 < \alpha$ time instants, the assumption holds with $\beta = \beta_0$. This will happen if $\boldsymbol{M}_{2,t} = \boldsymbol{I}_{\mathcal{T}_t}$ with $\mathcal{T}_t = \mathcal{T}^{[k]}$ for at most β time instants and if the distinct sets $\mathcal{T}^{[k]}$ are mutually disjoint. Finally, the $\mathcal{T}^{[k]}$'s need not even be mutually disjoint. As long as they are such that $\boldsymbol{B}$ is a matrix with nonzero blocks on only the main diagonal and on a few diagonals near it, e.g., if it is block tri-diagonal, it can be shown that the above assumption holds. This example is generalized in Model 1.3 given below. Lemma 1.4 relies on the above intuition to prove that it is indeed a special case of (5).

B. Examples of correlated-PCA problems

One key example of correlated-PCA is the *PCA with missing data (PCA-missing)* problem. Let $\mathcal{T}_t$ denote the set of missing entries at time t. Suppose, for simplicity, we set the missing entries of $\boldsymbol{y}_t$ to

zero. Then $\boldsymbol{y}_t$ can be expressed as

$$\boldsymbol{y}_t = \boldsymbol{\ell}_t - \boldsymbol{I}_{\mathcal{T}_t}\boldsymbol{I}_{\mathcal{T}_t}{}'\boldsymbol{\ell}_t. \tag{6}$$

In this case $\boldsymbol{M}_{2,t} = \boldsymbol{I}_{\mathcal{T}_t}$ and $\boldsymbol{M}_{1,t} = -\boldsymbol{I}_{\mathcal{T}_t}{}'$ and q is an upper bound on $\|\boldsymbol{I}_{\mathcal{T}_t}{}'\boldsymbol{P}\|$. Thus, to ensure that q is small, we need the columns of $\boldsymbol{P}$ to be dense vectors. For the reader familiar with low-rank matrix completion (MC), e.g., [13], [14], [15], [16], [17], *PCA-missing* can also be solved by first solving the low-rank matrix completion problem to recover $\boldsymbol{L}$, followed by PCA on the completed matrix. This would, of course, be much more expensive than directly solving *PCA-missing* and may need more assumptions. For example, recovering $\boldsymbol{L}$ correctly requires both the left singular vectors of $\boldsymbol{L}$ (columns of $\boldsymbol{P}$) and the right singular vectors of $\boldsymbol{L}$ be dense [14], [17]. We discuss this further in Sec. IV.

Another example is that of robust PCA (low-rank + sparse formulation) [18], [19], [20] when the sparse component's magnitude is correlated with $\boldsymbol{\ell}_t$, but its support is independent of $\boldsymbol{\ell}_t$. Let $\mathcal{T}_t$ denote the support set of $\boldsymbol{w}_t$ and let $\boldsymbol{x}_t$ be the $|\mathcal{T}_t|$-length vector of its nonzero entries. If we assume linear dependency, we can rewrite $\boldsymbol{y}_t$ as

$$\boldsymbol{y}_t = \boldsymbol{\ell}_t + \boldsymbol{I}_{\mathcal{T}_t}\boldsymbol{x}_t,\ \ \boldsymbol{x}_t = \boldsymbol{M}_{s,t}\boldsymbol{\ell}_t. \tag{7}$$

Thus $\boldsymbol{M}_{2,t} = \boldsymbol{I}_{\mathcal{T}_t}$ and $\boldsymbol{M}_{1,t} = \boldsymbol{M}_{s,t}$. In this case, a solution for the PCA problem *will work only when the corrupting sparse component* $\boldsymbol{w}_t = \boldsymbol{I}_{\mathcal{T}_t}\boldsymbol{M}_{s,t}\boldsymbol{\ell}_t$ *has magnitude that is small compared to that of* $\boldsymbol{\ell}_t$. In the rest of the paper, we refer to this problem is *"PCA with sparse data-dependent corruptions (PCA-SDDC)"*.

One key application where this problem occurs is in video analytics of videos consisting of a slow changing background sequence (modeled as being approximately low-rank) and a sparse foreground image sequence consisting typically of one or more moving objects [18]. This is a PCA-SDDC problem if the goal is to estimate the background sequence's subspace. For this problem, $\boldsymbol{\ell}_t$ is the background image at time t, $\mathcal{T}_t$ is the support set of the foreground image at t, and $\boldsymbol{x}_t$ is the difference between foreground and background intensities on $\mathcal{T}_t$ [1]. For PCA-SDDC, again, an alternative solution approach is to use an RPCA solution such as principal components' pursuit (PCP) [18], [19] or Alternating-Minimization (Alt-Min-RPCA) [21] to first recover the matrix $\boldsymbol{L}$ followed by PCA on $\boldsymbol{L}$. As demonstrated in Sec. VIII, this approach will be much slower. Moreover, it will work only if the required incoherence assumptions hold, e.g., as shown in Sec. VIII, Tables I, III, if the columns of $\boldsymbol{P}$ are sparse, this will fail.

[1]If all the entries in $\boldsymbol{x}_t$ are large, so that the foreground support is easily detectable, $\mathcal{T}_t$ can be assumed to be known. This application then becomes an instance of the PCA-missing problem.

A third example is the subspace update step of ReProCS for online robust PCA [2], [4]. We discuss this in Sec. VII-A.

In all three of the above applications, the assumptions on the data-noise correlation matrix given in Model 1.2 hold if there are "enough" changes in the set of missing or corrupted entries, $\mathcal{T}_t$. One example situation, inspired by the video application, is that of a 1D object of length s or less that remains static for at most β frames at a time. When it moves, it moves by at least a certain fraction of s pixels. The following model is inspired by the object's support.

Model 1.3 (model on $\mathcal{T}_t$, parameters: α, β)**.** *In any interval $\mathcal{J}_u^\alpha := [u\alpha, (u+1)\alpha) \subseteq [1, m]$, the following holds.*

Let l denote the number of times the set $\mathcal{T}_t$ changes in this interval (so $0 \le l \le \alpha - 1$). Let $t^0 := u\alpha$; let t^k, with $t^k < t^{k+1}$, denote the time instants in this interval at which $\mathcal{T}_t$ changes; and let $\mathcal{T}^{[k]}$ denote the distinct sets. In other words, $\mathcal{T}_t = \mathcal{T}^{[k]}$ for $t \in [t^k, t^{k+1}) \subseteq \mathcal{J}_u^\alpha$, for each $k = 1, 2, \dots, l$. Assume that the following hold with a $\beta < \alpha$:

1) *$(t^{k+1} - t^k) \le \tilde{\beta}$ and $|\mathcal{T}^{[k]}| \le s$;*
2) *$\rho^2 \tilde{\beta} \le \beta$ where ρ is the smallest positive integer so that, for any $0 \le k \le l$, $\mathcal{T}^{[k]}$ and $\mathcal{T}^{[k+\rho]}$ are disjoint;*
3) *for any k_1, k_2 satisfying $0 \le k_1 < k_2 \le l$, the sets $(\mathcal{T}^{[k_1]} \setminus \mathcal{T}^{[k_1+1]})$ and $(\mathcal{T}^{[k_2]} \setminus \mathcal{T}^{[k_2+1]})$ are disjoint.*

An implicit assumption for condition 3 to hold is that $\sum_{k=0}^{l} |\mathcal{T}^{[k]} \setminus \mathcal{T}^{[k+1]}| \le n$. As will be evident from the example given next, conditions 2 and 3 enforce an upper bound on the maximum support size s.

To connect this model with the moving object example given above, condition 1 holds if the object's size is at most s and if it moves at least once every $\tilde{\beta}$ frames. Condition 2 holds, if, every time it moves, it moves in the same direction and by at least $\frac{s}{\rho}$ pixels. Condition 3 holds if, every time it moves, it moves in the same direction and by at most $d_0 \ge \frac{s}{\rho}$ pixels, with $d_0 \alpha \le n$ (or, more generally, the motion is such that, if the object were to move at each frame, and if it started at the top of the frame, it does not reach the bottom of the frames in a time interval of length α).

The following lemma taken from [3] shows that, with this model on $\mathcal{T}_t$, both the PCA-missing and PCA-SDDC problems satisfy Model 1.2 assumed earlier.

Lemma 1.4. *[[3], Lemmas 5.2 and 5.3] Assume that Model 1.3 holds. For any sequence of $|\mathcal{T}_t| \times |\mathcal{T}_t|$*

symmetric positive-semi-definite matrices $\boldsymbol{A}_t$, and for any interval $\mathcal{J}_u^\alpha \subseteq [1, m]$,

$$\| \sum_{t \in \mathcal{J}_u^\alpha} \boldsymbol{I}_{\mathcal{T}_t} \boldsymbol{A}_t \boldsymbol{I}_{\mathcal{T}_t}{}' \| \leq (\rho^2 \tilde{\beta}) \max_{t \in \mathcal{J}_u^\alpha} \|\boldsymbol{A}_t\| \leq \beta \max_{t \in \mathcal{J}_u^\alpha} \|\boldsymbol{A}_t\|$$

Thus, if $\|\boldsymbol{I}_{\mathcal{T}_t}{}' \boldsymbol{P}\| \leq q < 1$, then the PCA-missing problem satisfies Model 1.2. If $\|\boldsymbol{M}_{s,t} \boldsymbol{P}\| \leq q < 1$, then the PCA-SDDC problem satisfies Model 1.2.

1) Generalizations: The above is one simple example of a support change model that would work. If instead of one object, there are k objects, and each of their supports satisfies Model 1.3, then again, with some modifications, it is possible to show that both the PCA-missing and PCA-SDDC problems satisfy Model 1.2. Moreover, notice that Model 1.3 does not require the entries in $\mathcal{T}_t$ to be contiguous at all (they need not correspond to the support of one or a few objects). Similarly, we can replace the condition that $\mathcal{T}_t$ be constant for at most $\tilde{\beta}$ time instants in Model 1.3 by $|\{t : \mathcal{T}_t = \mathcal{T}^{[k]}\}| \leq \tilde{\beta}$.

Thirdly, the requirement of the object(s) always moving in one direction may seem too stringent. As explained in [3, Lemma 9.4], a Bernoulli-Gaussian "constant velocity with random acceleration" motion model will also work whp. It allows the object to move at each frame with probability p and not move with probability $1 - p$ independent of past or future frames; when the object moves, it moves with an iid Gaussian velocity that has mean $1.1s/\rho$ and variance σ^2; σ^2 needs to be upper bounded and p needs to be lower bounded.

Lastly, if $s < c_1 \alpha$ for $c_1 \ll 1$, another model that works is that of an object of length s or less moving by at least one pixel and at most b pixels at each time [3, Lemma 9.5].

II. Simple EVD

Simple EVD computes the top eigenvectors of the empirical covariance matrix, $\frac{1}{\alpha} \sum_{t=1}^{\alpha} \boldsymbol{y}_t \boldsymbol{y}_t{}'$, of the observed data. The following can be shown. This, and all our later results, use SE defined in (1) as the subspace error metric.

Theorem 2.1 (simple-EVD result)**.** *Let $\hat{\boldsymbol{P}}$ denote the matrix containing all the eigenvectors of $\frac{1}{\alpha} \sum_{t=1}^{\alpha} \boldsymbol{y}_t \boldsymbol{y}_t{}'$ with eigenvalues above a threshold, λ_{thresh}, as its columns. Pick a ζ so that $r\zeta \leq 0.01$. Suppose that $\boldsymbol{y}_t$'s satisfy* (2) *and the following hold.*

1) Model 1.1 on $\boldsymbol{\ell}_t$ holds. Define

$$\alpha_0 := C\eta^2 \frac{r^2 11 \log n}{(r\zeta)^2} \max(f, qf, q^2 f)^2, \; C := \frac{32}{0.01^2}.$$

2) Model 1.2 on $\boldsymbol{M}_t$ holds for any $\alpha \geq \alpha_0$ and for any β satisfying

$$\frac{\beta}{\alpha} \leq \left(\frac{1 - r\zeta}{2}\right)^2 \min\left(\frac{(r\zeta)^2}{4.1(qf)^2}, \frac{(r\zeta)}{q^2 f}\right)$$

3) Set algorithm parameters $\lambda_{\text{thresh}} = 0.95\lambda^-$ *and* $\alpha \geq \alpha_0$.

Then, with probability at least $1 - 6n^{-10}$, $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \leq r\zeta$.

Proof: The proof is given in Section V.

Consider the lower bound on α. We refer to this as the "sample complexity". Since $q < 1$, and η is a small constant (e.g., for the uniform distribution, $\eta = 3$), for a fixed error level, $r\zeta$, α_0 simplifies to $Cf^2r^2 \log n$. Notice that the dependence on n is logarithmic. It is possible to show that the sample complexity scales as $\log n$ because we assume that the $\boldsymbol{\ell}_t$'s are bounded random variables (r.v.s). As a result we can apply the matrix Hoeffding inequality [22] to bound the perturbation between the observed data's empirical covariance matrix and that of the true data. The bounded r.v. assumption is actually a more practical one than the usual Gaussian assumption since most sources of data have finite power. The dependence on f^2 can be problematic when it is large.

Consider the upper bound on β/α. Clearly, the smaller term is the first one. This depends on $1/(qf)^2$. Thus, when f is large and q is not small enough, the bound required may be impractically small. As will be evident from the proof (see Remark 5.3), we get this bound because $\boldsymbol{w}_t$ is correlated with $\boldsymbol{\ell}_t$ and so $\mathbb{E}[\boldsymbol{\ell}_t \boldsymbol{w}_t{}'] \neq 0$. If $\boldsymbol{w}_t$ and $\boldsymbol{\ell}_t$ were uncorrelated, f would get replaced by $\frac{\lambda_{\max}(\text{Cov}(\boldsymbol{w}_t))}{\lambda^-}$ in the upper bound on β/α. If $\boldsymbol{w}_t$ is small, this would be much smaller than f.

2) Corollaries for PCA-missing and PCA-SDDC: Using Lemma 1.4, we have the following corollaries.

Corollary 2.2 (PCA-missing)**.** *Consider the PCA-missing model,* (6)*, and assume that* $\max_t \|\boldsymbol{I}_{\mathcal{T}_t}{}'\boldsymbol{P}\| \leq q < 1$. *Assume that everything in Theorem 2.1 holds except that we replace Model 1.2 by Model 1.3 with* $\alpha \geq \alpha_0$ *and with* β *satisfying the upper bound given there. Then, with probability at least* $1 - 6n^{-10}$, $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \leq r\zeta$.

Corollary 2.3 (PCA-SDDC)**.** *Consider the PCA-SDDC model,* (7)*, and assume that* $\max_t \|\boldsymbol{M}_{s,t}\boldsymbol{P}\| \leq q < 1$. *Everything else stated in Corollary 2.2 holds.*

III. CLUSTER-EVD

To try to relax the strong dependence on f^2 of the result above, we develop a generalization of simple-EVD that we call *cluster-EVD*. This requires the clustering assumption.

A. Clustering assumption

To state the assumption, define the following partition of the index set $\{1, 2, \ldots r\}$ based on the eigenvalues of $\boldsymbol{\Sigma}$. Let λ_i denote its i-th largest eigenvalue.

Definition 3.1 (g-condition-number partition of $\{1, 2, \dots r\}$)**.** *Define $\mathcal{G}_1 = \{1, 2, \dots r_1\}$ where r_1 is the index for which $\frac{\lambda_1}{\lambda_{r_1}} \leq g$ and $\frac{\lambda_1}{\lambda_{r_1+1}} > g$. In words, to define $\mathcal{G}_1$, start with the index of the first (largest) eigenvalue and keep adding indices of the smaller eigenvalues to the set until the ratio of the maximum to the minimum eigenvalue first exceeds g.*

For each $k > 1$, let $r_* = (\sum_{i=1}^{k-1} r_i)$. Define $\mathcal{G}_k = \{r_* + 1, r_* + 2, \dots, r_* + r_k\}$ where r_k is the index such that $\frac{\lambda_{r_*+1}}{\lambda_{r_*+r_k}} \leq g$ and $\frac{\lambda_{r_*+1}}{\lambda_{r_*+r_k+1}} > g$. In words, to define $\mathcal{G}_k$, start with the index of the $(r_* + 1)$-th eigenvalue, and repeat the above procedure.

Keep incrementing k and doing the above until $\lambda_{r_*+r_k+1} = 0$, i.e., until there are no more nonzero eigenvalues. Define $\vartheta = k$ as the number of sets in the partition. Thus $\{\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_\vartheta\}$ is the desired partition.

Definition 3.2. *Define $\boldsymbol{G}_0 = [.]$ and $\boldsymbol{G}_k := (\boldsymbol{P})_{\mathcal{G}_k}$.*

Definition 3.3. *Define $\lambda_k^+ := \max_{i \in \mathcal{G}_k} \lambda_i(\boldsymbol{\Lambda})$ and $\lambda_k^- := \min_{i \in \mathcal{G}_k} \lambda_i(\boldsymbol{\Lambda})$.*

By definition,

$$\frac{\lambda_k^+}{\lambda_k^-} \leq g.$$

We quantify the "distance" between consecutive sets of the partition using

$$\chi := \max_{k=1,2,\dots,\vartheta} \frac{\lambda_{k+1}^+}{\lambda_k^-}.$$

Clearly, $g \geq 1$ and $\chi \leq 1$ always.

Model 3.4 (Clustered eigenvalues, parameters: g^+, χ^+, ϑ, r_k's)**.** *For a $1 \leq g^+ < f$ and a $\chi^+ < 1$, assume that there exists a g satisfying $1 \leq g \leq g^+$ for which we can define a g-condition-number partition of $\{1, 2, \dots r\}$ that satisfies $\chi \leq \chi^+$. The number of sets in the partition is ϑ.*

When g^+ and χ^+ are small, we say that the eigenvalues are "well-clustered" and we refer to the sets $\mathcal{G}_k$ as the "clusters".

We expect the eigenvalues of the data covariance matrix to be clustered for data that has variability across different scales. The large scale variations would result in the first (largest eigenvalues') cluster and the smaller scale variations would form the later clusters. This is true for video "textures" such as moving waters or waving trees in a forest. We tested this assumption on some such videos. We describe our conclusions here for three videos - "lake" (video of moving lake waters), "waving-tree" (video consisting of waving trees), and "curtain" (video of window curtains moving due to the wind). For each

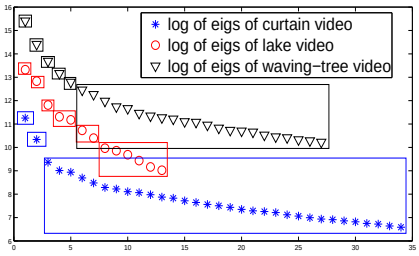

Fig. 1: Eigenvalue clusters of the three low-rankified videos.

video, we first made it low-rank by keeping the eigenvectors corresponding to the smallest number of eigenvalues that contain at least 90% of the total energy and projecting the video onto this subspace. Even for the "low-rankified" versions of these videos, f is large: $f = 74$ for lake, $f = 107$ for curtain, $f = 180$ for waving-tree.

For the lake video, the clustering assumption, Model 3.4, holds with $\vartheta = 6$ clusters, $g^+ = 2.6$ and $\chi^+ = 0.7$. For the waving-tree video, it holds with $\vartheta = 6$, $g^+ = 9.4$ and $\chi^+ = 0.72$. For the curtain video, $\vartheta = 3$, $g^+ = 16.1$ and $\chi^+ = 0.5$. We show the clusters of eigenvalues in Fig. 1.

B. Cluster-EVD algorithm

The cluster-EVD approach is summarized in Algorithm 1. It is related to, but significantly different from, the ones introduced in [2], [4] for the subspace deletion step of an online dynamic RPCA algorithm. The one introduced in [2] assumed that the clusters were known to the algorithm (which is unrealistic). The one studied in [4] has an automatic cluster estimation approach, but, its cluster-estimation step needs averaging over significantly more data points α compared to what Algorithm 1 needs.

The main idea of Algorithm 1 is as follows. We start by computing the empirical covariance matrix of the first set of α observed data points, $\hat{\boldsymbol{D}}_1 := \frac{1}{\alpha}\sum_{t=1}^{\alpha} \boldsymbol{y}_t \boldsymbol{y}_t'$. Let $\hat{\lambda}_i$ denote its i-th largest eigenvalue. To estimate the first cluster of eigenvalues, $\hat{\mathcal{G}}_1$, we start with the index of the first (largest) eigenvalue and keep adding indices of the smaller eigenvalues to it until the ratio of the maximum to the minimum eigenvalue exceeds $\hat{g}$ or until the minimum eigenvalue goes below a "zero threshold", λ_{thresh}. In other words, we find the index $\hat{r}_1$ for which $\frac{\hat{\lambda}_1}{\hat{\lambda}_{\hat{r}_1}} \leq \hat{g}$ and either $\frac{\hat{\lambda}_1}{\hat{\lambda}_{\hat{r}_1+1}} > \hat{g}$ or $\hat{\lambda}_{\hat{r}_1+1} < \lambda_{\text{thresh}}$. We set $\hat{\mathcal{G}}_1 = \{1, 2, \ldots, \hat{r}_1\}$. Then, we estimate the subspace corresponding to the first cluster, $\text{range}(\boldsymbol{G}_1)$ by computing the top $\hat{r}_1$ eigenvectors of $\hat{\boldsymbol{D}}_1$. To get the second cluster and its subspace, we project the next set of α $\boldsymbol{y}_t$'s orthogonal to $\hat{\boldsymbol{G}}_1$ followed by repeating the above procedure. This is repeated for each $k > 1$ until $\hat{\lambda}_{\hat{r}_k+1} < \lambda_{\text{thresh}}$.

Algorithm 1 **Cluster-EVD**

Parameters: α, $\hat{g}$, λ_{thresh}.

Set $\hat{\boldsymbol{G}}_0 \leftarrow [.]$. Set the flag $\text{Stop} \leftarrow 0$. Set $k \leftarrow 1$.

Repeat

1) Let $\hat{\boldsymbol{G}}_{\text{det},k} := [\hat{\boldsymbol{G}}_0, \hat{\boldsymbol{G}}_1, \dots \hat{\boldsymbol{G}}_{k-1}]$ and let $\boldsymbol{\Psi}_k := (\boldsymbol{I} - \hat{\boldsymbol{G}}_{\text{det},k}\hat{\boldsymbol{G}}_{\text{det},k}{}')$. Notice that $\boldsymbol{\Psi}_1 = \boldsymbol{I}$. Compute
$$\hat{\boldsymbol{D}}_k = \boldsymbol{\Psi}_k \left(\frac{1}{\alpha} \sum_{t=(k-1)\alpha+1}^{k\alpha} \boldsymbol{y}_t \boldsymbol{y}_t{}' \right) \boldsymbol{\Psi}_k$$
2) Find the k-th cluster, $\hat{\mathcal{G}}_k$: let $\hat{\lambda}_i = \lambda_i(\hat{\boldsymbol{D}}_k)$;
   a) find the smallest index $\hat{r}_k$ for which $\frac{\hat{\lambda}_1}{\hat{\lambda}_{\hat{r}_k}} \leq \hat{g}$ and either $\frac{\hat{\lambda}_1}{\hat{\lambda}_{\hat{r}_k+1}} > \hat{g}$ or $\hat{\lambda}_{\hat{r}_k+1} < \lambda_{\text{thresh}}$;
   b) set $\hat{\mathcal{G}}_k = \{\hat{r}_* + 1, \hat{r}_* + 2, \dots, \hat{r}_* + \hat{r}_k\}$ where $\hat{r}_* := \sum_{j=1}^{k-1} \hat{r}_j$;
   c) if $\hat{\lambda}_{\hat{r}_k+1} < \lambda_{\text{thresh}}$, update the flag $\text{Stop} \leftarrow 1$
3) Compute $\hat{\boldsymbol{G}}_k \leftarrow \text{eigenvectors}(\hat{\boldsymbol{D}}_k, \hat{r}_k)$; increment k

Until $\text{Stop} == 1$.

Set $\hat{\vartheta} \leftarrow k$. Output $\hat{\boldsymbol{P}} \leftarrow [\hat{\boldsymbol{G}}_1 \cdots \hat{\boldsymbol{G}}_{\hat{\vartheta}}]$.

$\text{eigenvectors}(\boldsymbol{\mathcal{M}}, r)$ returns a basis matrix for the span of the top r eigenvectors of $\boldsymbol{\mathcal{M}}$.

C. Main result

We give the performance guarantee for Algorithm 1 here. Its parameters are set as follows. We set $\hat{g}$ to a value that is a little larger than g. This is needed to allow for the fact that $\hat{\lambda}_i$ is not equal to the i-th eigenvalue of $\boldsymbol{\Lambda}$ but is within a small margin of it. For the same reason, we need to also use a nonzero "zeroing" threshold, λ_{thresh}, that is larger than zero but smaller than λ^-. We set α large enough to ensure that $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \leq r\zeta$ holds with a high enough probability.

Theorem 3.5 (cluster-EVD result)**.** *Consider Algorithm 1. Pick a ζ so that $r^2\zeta \leq 0.0001$, and $r^2\zeta f \leq 0.01$. Suppose that $\boldsymbol{y}_t$'s satisfy* (2) *and the following hold.*

1) Model 1.1 and Model 3.4 on $\boldsymbol{\ell}_t$ hold with χ^+ satisfying $\chi^+ \leq \min(1 - r\zeta - \frac{0.08}{0.25}, \frac{g^+ - 0.0001}{1.01g^+ + 0.0001} - 0.0001)$. Define
$$\alpha_0 := C\eta^2 \frac{r^2(11\log n + \log\vartheta)}{(r\zeta)^2} \max(g^+, qg^+,$$
$$q^2 f, q(r\zeta)f, (r\zeta)^2 f, q\sqrt{fg^+}, (r\zeta)\sqrt{fg^+})^2,\ C := \frac{32 \cdot 16}{0.01^2}.$$

2) Model 1.2 on $\boldsymbol{M}_t$ holds with $\alpha \geq \alpha_0$ and with β satisfying

$$\frac{\beta}{\alpha} \leq \left(\frac{(1-r\zeta-\chi^+)}{2}\right)^2 \min\left(\frac{(r_k\zeta)^2}{4.1(qg^+)^2}, \frac{(r_k\zeta)}{q^2 f}\right).$$

3) Set algorithm parameters $\hat{g} = 1.01g^+ + 0.0001$, $\lambda_{\text{thresh}} = 0.95\lambda^-$ and $\alpha \geq \alpha_0$.

Then, with probability at least $1 - 12n^{-10}$, $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \leq r\zeta$.

Proof: The proof is given in Section VI and a key lemma for it is proved in Appendix A.

Remark 3.6. *In the above result, for simplicity, we have given one specific value for $\hat{g}$ and λ_{thresh}. Note, however, that one can set $\hat{g}$ to be anything that satisfies* (13) *given later and one can set λ_{thresh} to be anything satisfying $5r\zeta\lambda^- \leq \lambda_{\text{thresh}} \leq 0.95\lambda^-$.*

With some changes to the proof (needed to apply the matrix Azuma inequality instead of the matrix Hoeffding), it is also possible to replace the mutual independence assumption on $\boldsymbol{\ell}_t$'s with an autoregressive model assumption. This will be studied in future work.

Remark 3.7. *We can also get corollaries for PCA-missing and PCA-SDDC for cluster-EVD by using Lemma 1.4.*

D. Added uncorrelated noise case

So far, we only stated results for when $\boldsymbol{y}_t$ satisfies (2). If, instead $\boldsymbol{y}_t$ satisfies (3), i.e., if $\boldsymbol{y}_t = \boldsymbol{\ell}_t + \boldsymbol{M}_t\boldsymbol{\ell}_t + \boldsymbol{\nu}_t$ where $\boldsymbol{\nu}_t$ is independent of $\boldsymbol{\ell}_t$, then our results change as follows.

Corollary 3.8 (cluster-EVD, with $\boldsymbol{\nu}_t$)**.** *Assume that $\boldsymbol{y}_t$ satisfies* (3)*. Let $b_{\boldsymbol{\nu}} = 0.01$. If $\boldsymbol{\nu}_t$ is independent of $\boldsymbol{\ell}_t$ and satisfies $\|\mathbb{E}[\boldsymbol{\nu}_t\boldsymbol{\nu}_t{}']\| \leq b_{\boldsymbol{\nu}} r_k\zeta\lambda^-$ and $\|\boldsymbol{\nu}_t\|^2 \leq rq\lambda^-$; if the following tighter bounds on χ^+ and on β/α hold:*

1) $\chi^+ \leq 1 - \frac{(b_\nu+0.08)}{0.25} - (1+b_{\boldsymbol{\nu}})r\zeta$, and

2) $\beta \leq \left(\frac{(1-(1+b_{\boldsymbol{\nu}})r\zeta-\chi^+)}{2}\right)^2 \min\left(\frac{(r_k\zeta)^2}{4.1q^2g^2}, \frac{(r_k\zeta)}{q^2 f}\right)\alpha$;

and if everything else in Theorem 3.5 holds, then its conclusions hold.

Proof. This follows in the same fashion as Theorem 3.5. The claims in the first two items of Lemma 6.6 will change to incorporate the new noise term. See Appendix B-3 (Supplementary document) for the changed lemma. ⊠

Corollary 3.9 (EVD, with $\boldsymbol{\nu}_t$)**.** *The corresponding result for EVD is Corollary 3.8 with $\vartheta = 1, r_k = r, \chi^+ = 0, g = f$.*

IV. DISCUSSION

1) Comparison of simple-EVD and cluster-EVD (c-EVD) results: Consider the lower bounds on α. In the c-EVD result (Theorem 3.5), if q is small enough (e.g., if $q \leq 1/\sqrt{f}$), and if $(r^2\zeta)f \leq 0.01$, it is clear that the maximum in the $\max(.,.,.,.)$ expression is achieved by g^2. Thus, in this regime, c-EVD needs $\alpha \geq C\frac{r^2(11\log n+\log\vartheta)}{(r\zeta)^2}g^2$ and its sample complexity is $\vartheta\alpha$. In the EVD result (Theorem 2.1), g gets replaced by f and ϑ by 1, and so, its sample complexity, $\alpha \geq C\frac{r^2 11\log n}{(r\zeta)^2}f^2$. In situations where the condition number f is very large but g is much smaller and ϑ is small (the clustering assumption holds well), the sample complexity of c-EVD will be much smaller than that of simple-EVD. However, notice that, the lower bound on α for simple-EVD holds for any $q < 1$ and for any ζ with $r\zeta < 0.01$ while the c-EVD lower bound given above holds only when q is small enough, e.g., $q = O(1/\sqrt{f})$, and ζ is small enough, e.g., $r\zeta = O(1/f)$. This tighter bound on ζ is needed because the error of the k-th step of c-EVD depends on the errors of the previous steps times f. Secondly, the c-EVD result also needs χ^+ and ϑ to be small (clustering assumption holds well), whereas, for simple-EVD, by definition, $\chi^+ = 0$ and $\vartheta = 1$. Another thing to note is that the constants in both lower bounds are very large with the c-EVD one being even larger.

Consider the upper bounds on β. To compare these, assume that the same α is used by EVD and c-EVD. Suppose that one uses $\alpha = \max(\alpha_0(\text{EVD}), \alpha_0(\text{c-EVD}))$. In this case, as long as r_k is large enough, χ^+ is small enough, and g is small enough, the upper bound on β needed by the c-EVD result is significantly looser. For example, if $\chi^+ = 0.2$, $\vartheta = 2$, $r_k = r/2$, then c-EVD needs $\beta \leq (0.5 \cdot 0.79 \cdot 0.5)^2\frac{(r\zeta)^2}{4.1q^2g^2}\alpha$ while simple-EVD needs $\beta \leq (0.5 \cdot 0.99)^2\frac{(r\zeta)^2}{4.1q^2f^2}\alpha$. If $g = 3$ but $f = 100$, clearly the c-EVD bound is significantly looser.

2) Comparison with other results for PCA-SDDC and PCA-missing: To our knowledge, there is no other result for correlated-PCA. Hence, we provide comparisons of the corollaries given above for the PCA-missing and PCA-SDDC special cases with works that also study these or related problems. An alternative solution for either PCA-missing or PCA-SDDC is to first recover the entire matrix $\boldsymbol{L}$ and then compute its subspace via SVD on the estimated $\boldsymbol{L}$. For the PCA-missing problem, this can be done by using any of the low-rank matrix completion techniques, e.g., nuclear norm minimization (NNM) [14] or alternating minimization (Alt-Min-MC) [17] or greedy techniques [15]. Similarly, for PCA-SDDC, this can be done by solving any of the recent provably correct RPCA techniques such as principal components' pursuit (PCP) [18], [19], [20] or alternating minimization (Alt-Min-RPCA) [21].

However, doing the above will have some disadvantages. The first is that it will be much slower. The difference in speed is most dramatic when solving the matrix-sized convex programs such as NNM or

PCP. As we demonstrate in Sec. VIII for the PCA-SDDC problem, even Alt-Min-RPCA is significantly slower than EVD or c-EVD. See Tables I, III, IV. If we use the time complexity from [21, footnote 3], then finding the span of the top k singular vectors of an $n \times m$ matrix takes $O(nmk)$ time. Thus, if ϑ is a constant both simple-EVD and c-EVD need $O(n\alpha r)$ time, whereas, for each iteration, Alt-Min-RPCA needs $O(n\alpha r^2)$ time [21].

The second disadvantage is that the above methods for MC or RPCA need more assumptions to provably correctly recover $\boldsymbol{L}$. For example, all the above methods need an incoherence assumption on both the left singular vectors, $\boldsymbol{P}$, and the right singular vectors, $\boldsymbol{V}$, of $\boldsymbol{L}$. Of course, it is possible that, if one studies these methods with the goal of only recovering the column space of $\boldsymbol{L}$ correctly, the incoherence assumption on the right singular vectors is not needed. From simulation experiments, the incoherence of the left singular vectors is definitely needed. On the other hand, for the PCA-SDDC problem, simple-EVD or c-EVD do not even need the incoherence assumption on $\boldsymbol{P}$. In Sec. VIII, in Tables I, III, IV, we show the results of simulation experiments in which the matrix $\boldsymbol{P}$ has sparse columns. As can be seen, PCP and Alt-Min-RPCA fail but EVD or c-EVD work well (of course only as long as q is small).

The disadvantage of both EVD and c-EVD compared with PCP or Alt-Min-RPCA is that they will work only when q is small enough (the corrupting noise is small compared to $\boldsymbol{\ell}_t$). Another disadvantage of c-EVD is that it needs the clustered eigenvalues' assumption.

At this point we should mention a result for a problem with a different notion of RPCA studied in [23]. This developed an approach to recover the column space of $\boldsymbol{L}$ when it is corrupted by column-sparse outliers. Its performance guarantee does not need incoherence of the left singular vectors of $\boldsymbol{L}$, but does need incoherence of the right singular vectors.

V. Proof of Theorem 2.1

This result also follows as a corollary of Theorem 3.5. We prove it separately first since its proof is short and and less notation-ally intensive. It will help understand the proof of Theorem 3.5 much more easily. Both results rely on the $\sin\theta$ theorem reviewed next.

A. $\sin\theta$ theorem

Davis and Kahan's $\sin\theta$ theorem [24] studies the rotation of eigenvectors by perturbation.

Theorem 5.1 ($\sin\theta$ theorem [24])**.** *Consider two Hermitian matrices $\boldsymbol{D}$ and $\hat{\boldsymbol{D}}$. Suppose that $\boldsymbol{D}$ can be*

decomposed as

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{E} & \boldsymbol{E}_\perp \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & 0 \\ 0 & \boldsymbol{A}_\perp \end{bmatrix} \begin{bmatrix} \boldsymbol{E}' \\ \boldsymbol{E}_\perp{}' \end{bmatrix}$$

where $[\boldsymbol{E}\ \boldsymbol{E}_\perp]$ *is an orthonormal matrix. Suppose that* $\hat{\boldsymbol{D}}$ *can be decomposed as*

$$\hat{\boldsymbol{D}} = \begin{bmatrix} \boldsymbol{F} & \boldsymbol{F}_\perp \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda} & 0 \\ 0 & \boldsymbol{\Lambda}_\perp \end{bmatrix} \begin{bmatrix} \boldsymbol{F}' \\ \boldsymbol{F}_\perp{}' \end{bmatrix}$$

where $[\boldsymbol{F}\ \boldsymbol{F}_\perp]$ *is another orthonormal matrix and is such that* $\operatorname{rank}(\boldsymbol{F}) = \operatorname{rank}(\boldsymbol{E})$. *Let* $\boldsymbol{H} := \hat{\boldsymbol{D}} - \boldsymbol{D}$ *denote the perturbation. If* $\lambda_{\min}(\boldsymbol{A}) > \lambda_{\max}(\boldsymbol{\Lambda}_\perp)$, *then*

$$\|(\boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}')\boldsymbol{E}\| \leq \frac{\|\boldsymbol{H}\|}{\lambda_{\min}(\boldsymbol{A}) - \lambda_{\max}(\Lambda_\perp)}.$$

Let $r = \operatorname{rank}(E)$. Suppose that $\boldsymbol{F}$ is the matrix of top r eigenvectors of $\hat{\boldsymbol{D}}$. Then $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_\perp$ are diagonal and $\lambda_{\max}(\boldsymbol{\Lambda}_\perp) = \lambda_{r+1}(\hat{\boldsymbol{D}}) \leq \lambda_{r+1}(\boldsymbol{D}) + \|\boldsymbol{H}\|$. The inequality follows using Weyl's inequality. Suppose also that $\lambda_{\min}(\boldsymbol{A}) > \lambda_{\max}(\boldsymbol{A}_\perp)$. Then, (i) $\lambda_r(\boldsymbol{D}) = \lambda_{\min}(\boldsymbol{A})$ and $\lambda_{r+1}(\boldsymbol{D}) = \lambda_{\max}(\boldsymbol{A}_\perp)$ and (ii) $\operatorname{range}(\boldsymbol{E})$ is equal to the span of the top r eigenvectors of $\boldsymbol{D}$. Thus, $\lambda_{\max}(\boldsymbol{\Lambda}_\perp) \leq \lambda_{\max}(\boldsymbol{A}_\perp) + \|\boldsymbol{H}\|$. With this we have the following corollary.

Corollary 5.2. *Consider a Hermitian matrix* $\boldsymbol{D}$ *and its perturbed version* $\hat{\boldsymbol{D}}$. *Suppose that* $\boldsymbol{D}$ *can be decomposed as*

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{E} & \boldsymbol{E}_\perp \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & 0 \\ 0 & \boldsymbol{A}_\perp \end{bmatrix} \begin{bmatrix} \boldsymbol{E}' \\ \boldsymbol{E}_\perp{}' \end{bmatrix}$$

where $\boldsymbol{E}$ *is a basis matrix. Let* $\boldsymbol{F}$ *denote the matrix containing the top* $\operatorname{rank}(\boldsymbol{E})$ *eigenvectors of* $\hat{\boldsymbol{D}}$. *Let* $\boldsymbol{H} := \hat{\boldsymbol{D}} - \boldsymbol{D}$ *denote the perturbation. If* $\lambda_{\min}(\boldsymbol{A}) - \lambda_{\max}(\boldsymbol{A}_\perp) - \|\boldsymbol{H}\| > 0$, *then*

$$\|(\boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}')\boldsymbol{E}\| \leq \frac{\|\boldsymbol{H}\|}{\lambda_{\min}(\boldsymbol{A}) - \lambda_{\max}(\boldsymbol{A}_\perp) - \|\boldsymbol{H}\|}.$$

and $\operatorname{range}(\boldsymbol{E})$ *is equal to the span of the top* $\operatorname{rank}(\boldsymbol{E})$ *eigenvectors of* $\boldsymbol{D}$.

B. Proof of Theorem 2.1

We use the $\sin\theta$ theorem [24] from Corollary 5.2. Apply it with $\hat{\boldsymbol{D}} = \frac{1}{\alpha}\sum_t \boldsymbol{y}_t\boldsymbol{y}_t{}'$ and $\boldsymbol{D} = \frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'$. Thus, $\boldsymbol{F} = \hat{\boldsymbol{P}}$. Let $\boldsymbol{a}_t := \boldsymbol{P}'\boldsymbol{\ell}_t$. Then, $\boldsymbol{D}$ can be decomposed as $\boldsymbol{P}(\frac{1}{\alpha}\sum_t \boldsymbol{a}_t\boldsymbol{a}_t{}')\boldsymbol{P}' + \boldsymbol{P}_\perp\boldsymbol{0}\boldsymbol{P}_\perp'$, and so we have $\boldsymbol{E} = \boldsymbol{P}$, $\boldsymbol{A} = \frac{1}{\alpha}\sum_t \boldsymbol{a}_t\boldsymbol{a}_t{}'$ and $\boldsymbol{A}_\perp = \boldsymbol{0}$. Moreover, it is easy to see that the perturbation $\boldsymbol{H} := \frac{1}{\alpha}\sum_t \boldsymbol{y}_t\boldsymbol{y}_t{}' - \frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'$ satisfies

$$\boldsymbol{H} = \frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t' + \frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{\ell}_t' + \frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t'. \tag{8}$$

Thus,

$$\begin{aligned}&\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})\\ &\leq \frac{2\|\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t \boldsymbol{w}_t'\| + \|\frac{1}{\alpha}\sum_t \boldsymbol{w}_t \boldsymbol{w}_t'\|}{\lambda_r(\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t \boldsymbol{\ell}_t') - (2\|\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t \boldsymbol{w}_t'\| + \|\frac{1}{\alpha}\sum_t \boldsymbol{w}_t \boldsymbol{w}_t'\|)}\end{aligned}$$

if the denominator is positive.

Remark 5.3. *Because $\boldsymbol{w}_t$ is correlated with $\boldsymbol{\ell}_t$, the $\boldsymbol{\ell}_t \boldsymbol{w}_t'$ terms are the dominant ones in the perturbation expression given in* (8)*. If they were uncorrelated, these two terms would be close to zero whp due to law of large numbers and the $\boldsymbol{w}_t \boldsymbol{w}_t'$ term would be the dominant one.*

In the next lemma, we bound the terms in the bound on $\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$ using the matrix Hoeffding inequality [22].

Lemma 5.4. *Let* $\epsilon = 0.01 r\zeta\lambda^-$.

1) With probability at least $1 - 2n\exp\left(-\alpha\frac{\epsilon^2}{32(\eta r q \lambda^+)^2}\right)$,

$$\|\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t \boldsymbol{w}_t{}'\| \leq q\lambda^+\sqrt{\frac{\beta}{\alpha}} + \epsilon = [qf\sqrt{\frac{\beta}{\alpha}} + 0.01 r\zeta]\lambda^-$$

2) With probability at least $1 - 2n\exp(-\frac{\alpha\epsilon^2}{32(\eta r q^2 \lambda^+)^2})$,

$$\|\frac{1}{\alpha}\sum_t \boldsymbol{w}_t \boldsymbol{w}_t{}'\| \leq \frac{\beta}{\alpha} q^2\lambda^+ + \epsilon = [q^2 f\frac{\beta}{\alpha} + 0.01 r\zeta]\lambda^-$$

3) With probability at least $1 - 2n\exp(-\frac{\alpha\epsilon^2}{32(\eta r \lambda^+)^2})$,

$$\lambda_r(\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t \boldsymbol{\ell}_t') \geq (1 - (r\zeta)^2)\lambda^- - \epsilon$$

Proof. This follows by using Lemma 6.6 given later with $\boldsymbol{G}_{\mathrm{cur}} \equiv \boldsymbol{P}$, $\boldsymbol{G}_{\mathrm{det}} \equiv [.]$, $\boldsymbol{G}_{\mathrm{undet}} \equiv [.]$, $\zeta_{\mathrm{det}} \equiv 0$, $r\zeta \equiv 0$, $r_{\mathrm{cur}} = r$, $g \equiv f$, $\chi \equiv 0$, $\vartheta \equiv 1$. ⊠

Using this lemma to bound the subspace error terms, followed by using the bounds on β/α and ζ, we conclude the following: w.p. at least $1 - 2n\exp\left(-\alpha\frac{\epsilon^2}{32(\eta r q \lambda^+)^2}\right) - 2n\exp(-\frac{\alpha\epsilon^2}{32(\eta r q^2 \lambda^+)^2}) - 2n\exp(-\frac{\alpha\epsilon^2}{32(\eta r \lambda^+)^2})$,

$$\begin{aligned}&\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})\\ &\leq \frac{2qf\sqrt{\frac{\beta}{\alpha}} + q^2 f\frac{\beta}{\alpha} + 0.03 r\zeta}{1 - (r\zeta)^2 - 0.01 r\zeta - (2qf\sqrt{\frac{\beta}{\alpha}} + q^2 f\frac{\beta}{\alpha} + 0.03 r\zeta)}\\ &\leq \frac{0.75(1 - r\zeta) r\zeta + 0.03 r\zeta}{1 - r\zeta} < r\zeta\end{aligned}$$

Using the bound $\alpha \geq \alpha_0$ from the theorem, the probability of the above event is at least $1 - 6n^{-10}$. We get this by bounding each of the three negative terms in the probability expression by $-2n^{-10}$. We work this out for the first term: $\alpha \frac{\epsilon^2}{32(\eta r q \lambda^+)^2} \geq \frac{32 \cdot 11}{(0.01)^2} \frac{\eta^2 r^2 (\log n)}{(r\zeta)^2} (qf)^2 \frac{(0.01 r \zeta \lambda^-)^2}{32 \eta^2 r^2 q^2 \lambda^{+2}} = 11 \log n$. Thus, $2n \exp\left(-\alpha \frac{\epsilon^2}{32(\eta r q \lambda^+)^2}\right) \leq 2n \exp(-11 \log n) \leq 2n^{-10}$.

VI. Proof of Theorem 3.5

We explain the overall idea of the proof next. In Sec. VI-B, we give a sequence of lemmas in generalized form (so that they can apply to various other problems). The proof of Theorem 3.5 is given in Sec. VI-C and follows easily by applying these. One of the lemmas of Sec. VI-B is proved in Appendix A while the others are proved there itself.

A. Overall idea

We need to bound $\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$. From Algorithm 1, $\hat{\boldsymbol{P}} = [\hat{\boldsymbol{G}}_1, \hat{\boldsymbol{G}}_2, \ldots, \hat{\boldsymbol{G}}_\vartheta]$ where $\hat{\boldsymbol{G}}_k$ is the matrix of top $\hat{r}_k$ eigenvectors of $\hat{\boldsymbol{D}}_k$ defined in Algorithm 1. Also, $\boldsymbol{P} = [\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_\vartheta]$ where $\boldsymbol{G}_k$ is a basis matrix with r_k columns.

Definition 6.1. *Define* $\zeta_k := \mathrm{SE}([\hat{\boldsymbol{G}}_1, \hat{\boldsymbol{G}}_2, \ldots, \hat{\boldsymbol{G}}_k], \boldsymbol{G}_k)$ *and* $\zeta_0 = 0$. *Define* $\zeta_k^+ := r_k \zeta$. *Let* $r_0 = 0$.

It is easy to see that

$$\begin{aligned} \mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) &\leq \sum_{k=1}^{\vartheta} \mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{G}_k) \\ &\leq \sum_{k=1}^{\vartheta} \mathrm{SE}([\hat{\boldsymbol{G}}_1, \hat{\boldsymbol{G}}_2, \ldots, \hat{\boldsymbol{G}}_k], \boldsymbol{G}_k) = \sum_{k=1}^{\vartheta} \zeta_k \end{aligned} \tag{9}$$

The first inequality is triangle inequality, the second follows because $[\hat{\boldsymbol{G}}_1, \hat{\boldsymbol{G}}_2, \ldots, \hat{\boldsymbol{G}}_k]$ is orthogonal to $[\hat{\boldsymbol{G}}_{k+1}, \ldots \boldsymbol{G}_\vartheta]$. Since $r = \sum_k r_k$, if we can show that $\zeta_k \leq \zeta_k^+ = r_k \zeta$ for all k we will be done.

We bound ζ_k using induction. The base case is easy and follows just from the definition, $\zeta_0 = \mathrm{SE}([.], [.]) = 0 = r_0 \zeta$. For bounding ζ_k, assume that for all $i = 1, 2, \ldots, k-1$, $\zeta_i \leq r_i \zeta$. This implies that

$$\begin{aligned} &\mathrm{SE}([\hat{\boldsymbol{G}}_1, \hat{\boldsymbol{G}}_2, \ldots, \hat{\boldsymbol{G}}_{k-1}], [\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_{k-1}]) \\ &\leq \sum_{i=1}^{k-1} \mathrm{SE}([\hat{\boldsymbol{G}}_1, \hat{\boldsymbol{G}}_2, \ldots, \hat{\boldsymbol{G}}_{k-1}], \boldsymbol{G}_i) \\ &\leq \sum_{i=1}^{k-1} \zeta_i \leq \sum_{i=1}^{k-1} r_i \zeta \leq r\zeta \end{aligned} \tag{10}$$

Using this, we will first show that $\hat{r}_k = r_k$, and then we will use this and the $\sin\theta$ result to bound ζ_k.

Before proceeding further, we simplify notation.

Definition 6.2.

1) Let

$$\boldsymbol{G}_{\rm det} := [\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_{k-1}],\ \boldsymbol{G}_{\rm cur} := \boldsymbol{G}_k,$$

$$\boldsymbol{G}_{\rm undet} := [\hat{\boldsymbol{G}}_{k+1}, \ldots \boldsymbol{G}_{\vartheta}]$$

2) Similarly, let $\hat{\boldsymbol{G}}_{\rm det} := [\hat{\boldsymbol{G}}_1, \hat{\boldsymbol{G}}_2, \ldots, \hat{\boldsymbol{G}}_{k-1}]$, $\hat{\boldsymbol{G}}_{\rm cur} := \hat{\boldsymbol{G}}_k$.

3) Let $\mathcal{G}_{\rm det} := \mathcal{G}_1 \cup \mathcal{G}_2 \cdots \cup \mathcal{G}_{k-1}$ *and* $\mathcal{G}_{\rm cur} = \mathcal{G}_k$.

4) Let $r_{\rm cur} := r_k = \mathrm{rank}(\boldsymbol{G}_k)$ *and* $\hat{r}_{\rm cur} := \hat{r}_k$.

5) Let $\lambda^+_{\rm cur} := \lambda^+_k,\ \lambda^-_{\rm cur} := \lambda^-_k,\ \lambda^+_{\rm undet} := \lambda^+_{k+1}$

6) Let $t_* = k\alpha$.

B. Main lemmas - generalized form

In this section, we give a sequence of lemmas that apply to a generic problem where $\boldsymbol{y}_t = \boldsymbol{\ell}_t + \boldsymbol{w}_t = \boldsymbol{\ell}_t + \boldsymbol{M}_t\boldsymbol{\ell}_t$ with $\boldsymbol{\ell}_t$ satisfying Model 1.1; $\boldsymbol{M}_t$ satisfying Model 1.2; and with $\boldsymbol{P}$ split into three parts as $\boldsymbol{P} = [\boldsymbol{G}_{\rm det}, \boldsymbol{G}_{\rm cur}, \boldsymbol{G}_{\rm undet}]$. We can correspondingly split $\boldsymbol{\Lambda}$ as $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\Lambda}_{\rm det}, \boldsymbol{\Lambda}_{\rm cur}, \boldsymbol{\Lambda}_{\rm undet})$.

We are given $\hat{\boldsymbol{G}}_{\rm det}$ that was computed using (some or all) $\boldsymbol{y}_t$'s for $t \le t_*$ and that satisfies $\zeta_{\rm det} \le r\zeta$. The goal is to estimate $\mathrm{range}(\boldsymbol{G}_{\rm cur})$ and bound the estimation error. This is done by first estimating $\hat{r}_{\rm cur}$ and then computing $\hat{\boldsymbol{G}}_{\rm cur}$ as the top $\hat{r}_{\rm cur}$ eigenvectors of

$$\hat{\boldsymbol{D}} := \frac{1}{\alpha}\sum_{t=t_*+1}^{t_*+\alpha} \boldsymbol{\Psi}\boldsymbol{y}_t\boldsymbol{y}_t{}'\boldsymbol{\Psi}. \tag{11}$$

To bound the estimation error, we first show that, whp, $\hat{r}_{\rm cur} = r_{\rm cur}$ and so $\hat{\mathcal{G}}_{\rm cur} = \mathcal{G}_{\rm cur}$; and then we use this to show that $\zeta_{\rm cur} \le r_{\rm cur}\zeta$.

Definition 6.3.

1) Define $\boldsymbol{\Psi} := \boldsymbol{I} - \hat{\boldsymbol{G}}_{\rm det}\hat{\boldsymbol{G}}_{\rm det}{}'$.

2) Define $\zeta_{\rm det} := \mathrm{SE}(\hat{\boldsymbol{G}}_{\rm det}, \boldsymbol{G}_{\rm det}) = \|\boldsymbol{\Psi}\boldsymbol{G}_{\rm det}\|$ *and* $\zeta^+_{\rm det} = r\zeta$

3) Define $\zeta_{\rm cur} := \mathrm{SE}([\hat{\boldsymbol{G}}_{\rm det}, \hat{\boldsymbol{G}}_{\rm cur}], \boldsymbol{G}_{\rm cur})$.

4) Let $(\boldsymbol{\Psi}\boldsymbol{G}_{\rm cur}) \overset{\rm QR}{=} \boldsymbol{E}_{\rm cur}\boldsymbol{R}_{\rm cur}$ *denote its reduced QR decomposition. Thus* $\boldsymbol{E}_{\rm cur}$ *is a basis matrix whose span equals that of* $(\boldsymbol{\Psi}\boldsymbol{G}_{\rm cur})$ *and* $\boldsymbol{R}_{\rm cur}$ *is a square upper triangular matrix with* $\|\boldsymbol{R}_{\rm cur}\| = \|\boldsymbol{\Psi}\boldsymbol{G}_{\rm cur}\| \le 1$.

5) Let $\lambda^+_{\rm cur} = \lambda_{\max}(\boldsymbol{\Lambda}_{\rm cur})$, $\lambda^-_{\rm cur} = \lambda_{\min}(\boldsymbol{\Lambda}_{\rm cur})$, $\lambda^+_{\rm undet} = \lambda_{\max}(\boldsymbol{\Lambda}_{\rm undet})$.

6) Let $r_{\rm cur} = \mathrm{rank}(\boldsymbol{G}_{\rm cur})$. *Clearly,* $r_{\rm cur} \le r$.

Remark 6.4. *In special cases,* $\boldsymbol{G}_{\rm det}$ *(and hence* $\hat{\boldsymbol{G}}_{\rm det}$*) could be empty; and/or* $\boldsymbol{G}_{\rm undet}$ *could be empty.*

- Since $\boldsymbol{\Lambda}$ contains eigenvalues in decreasing order, when $\boldsymbol{G}_{\rm undet}$ is not empty, $\lambda^- \le \lambda^+_{\rm undet} \le \lambda^-_{\rm cur} \le \lambda^+_{\rm cur} \le \lambda^+$.
- When $\boldsymbol{G}_{\rm undet}$ is empty, $\lambda^+_{\rm undet} = 0$ and $\lambda^- \le \lambda^-_{\rm cur} \le \lambda^+_{\rm cur} \le \lambda^+$.

Using $\|\boldsymbol{R}_{\rm cur}\| = \|\boldsymbol{\Psi}\boldsymbol{G}_{\rm cur}\| \le 1$,

$$\begin{aligned}\zeta_{\rm cur} &= \|(\boldsymbol{I} - \hat{\boldsymbol{G}}_{\rm cur}\hat{\boldsymbol{G}}_{\rm cur}{}')\boldsymbol{\Psi}\boldsymbol{G}_{\rm cur}\| \\ &= \|(\boldsymbol{I} - \hat{\boldsymbol{G}}_{\rm cur}\hat{\boldsymbol{G}}_{\rm cur}{}')\boldsymbol{E}_{\rm cur}\boldsymbol{R}_{\rm cur}\| \\ &\le \|(\boldsymbol{I} - \hat{\boldsymbol{G}}_{\rm cur}\hat{\boldsymbol{G}}_{\rm cur}{}')\boldsymbol{E}_{\rm cur}\| = \mathrm{SE}(\hat{\boldsymbol{G}}_{\rm cur}, \boldsymbol{E}_{\rm cur}).\end{aligned}$$

Thus, to bound $\zeta_{\rm cur}$ we need to bound $\mathrm{SE}(\hat{\boldsymbol{G}}_{\rm cur}, \boldsymbol{E}_{\rm cur})$. $\hat{\boldsymbol{G}}_{\rm cur}$ is the matrix of top $\hat{r}_{\rm cur}$ eigenvectors of $\hat{\boldsymbol{D}}$. From its definition, $\boldsymbol{E}_{\rm cur}$ is a basis matrix with $r_{\rm cur}$ columns. Suppose for a moment that $\hat{r}_{\rm cur} = r_{\rm cur}$. Then, in order to bound $\mathrm{SE}(\hat{\boldsymbol{G}}_{\rm cur}, \boldsymbol{E}_{\rm cur})$, we can use the $\sin\theta$ result, Corollary 5.2. To do this, we need to define a matrix $\boldsymbol{D}$ so that, under appropriate assumptions, the span of its top $r_{\rm cur}$ eigenvectors equals $\mathrm{range}(\boldsymbol{E}_{\rm cur})$. For the simple EVD proof, we used $\frac{1}{\alpha}\sum_{t=t_*+1}^{t_*+\alpha} \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{\Psi}$ as the matrix $\boldsymbol{D}$. However, this will not work now since $\boldsymbol{E}_{\rm cur}$ is not orthonormal to $\boldsymbol{\Psi}\boldsymbol{G}_{\rm det}$ or to $\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}$. But, instead we can use

$$\begin{aligned}&\boldsymbol{D} = \boldsymbol{E}_{\rm cur}\boldsymbol{A}\boldsymbol{E}_{\rm cur}{}' + \boldsymbol{E}_{{\rm cur},\perp}\boldsymbol{A}_\perp\boldsymbol{E}_{{\rm cur},\perp}{}', \text{ where} \\ &\boldsymbol{A} := \boldsymbol{E}_{\rm cur}{}'(\frac{1}{\alpha}\sum_{t=t_*+1}^{t_*+\alpha} \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{\Psi})\boldsymbol{E}_{\rm cur} \text{ and} \\ &\boldsymbol{A}_\perp := \boldsymbol{E}_{{\rm cur},\perp}{}'(\frac{1}{\alpha}\sum_{t=t_*+1}^{t_*+\alpha} \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{\Psi})\boldsymbol{E}_{{\rm cur},\perp}\end{aligned} \tag{12}$$

Now, by construction, $\boldsymbol{D}$ is in the desired form.

With the above choice of $\boldsymbol{D}$, $\boldsymbol{H} := \hat{\boldsymbol{D}} - \boldsymbol{D}$ satisfies[2] $\boldsymbol{H} = \mathrm{term1} + \mathrm{term1}' + \mathrm{term2} + \mathrm{term3} + \mathrm{term3}'$ where $\mathrm{term1} := \frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{w}_t'$, $\mathrm{term2} := \frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{w}_t\boldsymbol{w}_t'$ and $\mathrm{term3} = \boldsymbol{E}_{\rm cur}\boldsymbol{E}_{\rm cur}{}'(\frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{\Psi})\boldsymbol{E}_{{\rm cur},\perp}\boldsymbol{E}_{{\rm cur},\perp}{}'$.

Thus, using the above along with Corollary 5.2, we can conclude the following.

Fact 6.5.

[2]This follows easily by writing $\boldsymbol{H} = (\hat{\boldsymbol{D}} - \frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{\Psi}) + (\frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{\Psi} - \boldsymbol{D})$ and using the fact that $\boldsymbol{M} = (\boldsymbol{E}\boldsymbol{E}' + \boldsymbol{E}_\perp\boldsymbol{E}_\perp{}')\boldsymbol{M}(\boldsymbol{E}\boldsymbol{E}' + \boldsymbol{E}_\perp\boldsymbol{E}_\perp{}')$ for $\frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{\Psi}$.

1) If $\hat{r}_{\rm cur} = r_{\rm cur}$*, and* $\lambda_{\min}(\boldsymbol{A}) - \lambda_{\max}(\boldsymbol{A}_\perp) - \|\boldsymbol{H}\| > 0$*,*

$$\zeta_{\rm cur} \leq {\rm SE}(\hat{\boldsymbol{G}}_{\rm cur}, \boldsymbol{E}_{\rm cur}) \leq \frac{\|\boldsymbol{H}\|}{\lambda_{\min}(\boldsymbol{A}) - \lambda_{\max}(\boldsymbol{A}_\perp) - \|\boldsymbol{H}\|}.$$

2) Let $\boldsymbol{Q} := \boldsymbol{E}_{\rm cur}\boldsymbol{E}_{\rm cur}{}'(\frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{\Psi})\boldsymbol{E}_{{\rm cur},\perp}\boldsymbol{E}_{{\rm cur},\perp}{}'$*. We have*

$$\|\boldsymbol{H}\| \leq 2\|\frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{w}_t'\| + \|\frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t'\| + 2\|\boldsymbol{Q}\|.$$

The next lemma bounds the RHS terms in the above lemma and a few other quantities needed for showing $\hat{r}_{\rm cur} = r_{\rm cur}$.

Lemma 6.6. *(1) Assume that* $\boldsymbol{y}_t = \boldsymbol{\ell}_t + \boldsymbol{w}_t = \boldsymbol{\ell}_t + \boldsymbol{M}_t\boldsymbol{\ell}_t$ *with* $\boldsymbol{\ell}_t$ *satisfying Model 1.1 and* $\boldsymbol{M}_t$ *satisfying Model 1.2.*

(2) Assume that we are given $\hat{\boldsymbol{G}}_{\rm det}$ *that was computed using (some or all)* $\boldsymbol{y}_t$*'s for* $t \leq t_*$ *and that satisfies* $\zeta_{\rm det} \leq r\zeta$*.*

Define $g := \lambda^+_{\rm cur}/\lambda^-_{\rm cur}$*,* $\chi := \lambda^+_{\rm undet}/\lambda^-_{\rm cur}$*. Set* $\epsilon := 0.01 r_{\rm cur}\zeta\lambda^-_{\rm cur}$*.*

Then, the following hold:

1) Let $p_1 := 2n\exp(-\frac{\alpha\epsilon^2}{32b^2_{prob}}$ *where* $b_{prob} := \eta r q((r\zeta)\lambda^+ + \lambda^+_{\rm cur} + (r\zeta)\sqrt{\lambda^+\lambda^+_{\rm cur}} + \sqrt{\lambda^+\lambda^+_{\rm cur}})$*. Conditioned on* $\{\zeta_{\rm det} \leq r\zeta\}$*, with probability at least* $1 - p_1$

$$\begin{aligned}\|\frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{w}_t{}'\| &\leq q((r\zeta)\lambda^+ + \lambda^+_{\rm cur})\sqrt{\frac{\beta}{\alpha}} + \epsilon \\ &\leq [q(r\zeta)f\sqrt{\frac{\beta}{\alpha}} + qg\sqrt{\frac{\beta}{\alpha}} + 0.01 r_{\rm cur}\zeta]\lambda^-_{\rm cur}.\end{aligned}$$

2) Let $p_2 := 2n\exp(-\frac{\alpha\epsilon^2}{32(q^2\eta r\lambda^+)^2})$*. Conditioned on* $\{\zeta_{\rm det} \leq r\zeta\}$*, with probability (w.p.) at least* $1 - p_2$*,*

$$\|\frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t{}'\| \leq \frac{\beta}{\alpha}q^2\lambda^+ + \epsilon \leq [\frac{\beta}{\alpha}q^2 f + 0.01 r_{\rm cur}\zeta]\lambda^-_{\rm cur}.$$

3) Let $p_3 := 2n\exp(-\frac{\alpha\epsilon^2}{32b^2_{prob}})$ *with* $b_{prob} := \eta r((r\zeta)^2\lambda^+ + \lambda^+_{\rm cur} + 2(r\zeta)\sqrt{\lambda^+\lambda^+_{\rm cur}})$*. Conditioned on* $\{\zeta_{\rm det} \leq r\zeta\}$*, with probability at least* $1 - p_3$*,*

$$\begin{aligned}&\|\boldsymbol{E}_{\rm cur}\boldsymbol{E}_{\rm cur}{}'(\frac{1}{\alpha}\boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{\Psi})\boldsymbol{E}_{{\rm cur},\perp}\boldsymbol{E}_{{\rm cur},\perp}{}'\| \\ &\leq (r\zeta)^2\lambda^+ + \frac{(r\zeta)^2}{\sqrt{1-(r\zeta)^2}}\lambda^+_{\rm undet} + \epsilon \\ &\leq [(r\zeta)^2 f + \frac{(r\zeta)^2}{\sqrt{1-(r\zeta)^2}}\chi + 0.01 r_{\rm cur}\zeta]\lambda^-_{\rm cur}.\end{aligned}$$

4) Conditioned on $\{\zeta_{\rm det} \le r\zeta\}$*, w.p. at least* $1-p_3$*,*

$$\begin{aligned}\lambda_{\min}(\boldsymbol{A}) &\ge (1-(r\zeta)^2)\lambda^-_{\rm cur} - \epsilon \\ &= [1-(r\zeta)^2 - 0.01 r_{\rm cur}\zeta]\lambda^-_{\rm cur}\end{aligned}$$

5) Conditioned on $\{\zeta_{\rm det} \le r\zeta\}$*, w.p. at least* $1-p_3$*,*

$$\begin{aligned}\lambda_{\max}(\boldsymbol{A}_\perp) &\le ((r\zeta)^2\lambda^+ + \lambda^+_{\rm undet}) + \epsilon \\ &\le [(r\zeta)^2 f + \chi + 0.01 r_{\rm cur}\zeta]\lambda^-_{\rm cur}.\end{aligned}$$

6) Conditioned on $\{\zeta_{\rm det} \le r\zeta\}$*, with probability at least* $1-p_3$*,*

$$\lambda_{\max}(\boldsymbol{A}_\perp) \ge (1-(r\zeta)^2 - \frac{(r\zeta)^2}{\sqrt{1-(r\zeta)^2}})\lambda^+_{\rm undet} - \epsilon.$$

7) Conditioned on $\{\zeta_{\rm det} \le r\zeta\}$*, w.p. at least* $1-p_3$*,*

$$\begin{aligned}\lambda_{\max}(\boldsymbol{A}) &\ge (1-(r\zeta)^2)\lambda^+_{\rm cur} - \epsilon \\ &= [(1-(r\zeta)^2)g - 0.01 r_{\rm cur}\zeta]\lambda^-_{\rm cur}.\end{aligned}$$

8) Conditioned on $\{\zeta_{\rm det} \le r\zeta\}$*, w.p. at least* $1-p_3$*,*

$$\begin{aligned}\lambda_{\max}(\boldsymbol{A}) &\le \lambda^+_{\rm cur} + (r\zeta)^2\lambda^+ + \frac{1}{1-r^2\zeta^2}(r\zeta)^2\lambda^+_{\rm undet} + \epsilon \\ &\le [g + (r\zeta)^2 f + \frac{(r\zeta)^2}{1-(r\zeta)^2}\chi + 0.01 r_{\rm cur}\zeta]\lambda^-_{\rm cur}.\end{aligned}$$

Proof. The proof is in Appendix A. ⊠

Corollary 6.7. *Consider the setting of Lemma 6.6. Assume*

1) $r(r\zeta) \le 0.0001$*, and* $r(r\zeta)f \le 0.01$*. Since* $r_{\rm cur} \le r$*, this implies that* $r_{\rm cur}\zeta \le 0.0001$*, and*
2) $\beta \le \left(\frac{(1-r_{\rm cur}\zeta-\chi)}{2}\right)^2 \min\left(\frac{(r_{\rm cur}\zeta)^2}{4.1q^2g^2}, \frac{(r_{\rm cur}\zeta)}{q^2 f}\right)\alpha$*.*

Using these and using $g \ge 1$*,* $g \le f$*,* $\chi \le 1$ *(these hold by definition), with probability at least*

$1 - p_1 - p_2 - 4p_3$,

$$\begin{aligned}
\|\boldsymbol{H}\| &\le [2.02qg\sqrt{\frac{\beta}{\alpha}} + \frac{\beta}{\alpha}q^2 f + 0.08 r_{\rm cur}\zeta]\lambda^-_{\rm cur} \\
&\le [0.75(1 - r\zeta - \chi) r_{\rm cur}\zeta + 0.08 r_{\rm cur}\zeta]\lambda^-_{\rm cur} \\
&\le 0.83 r_{\rm cur}\zeta\lambda^-_{\rm cur}, \\
\lambda_{\max}(\boldsymbol{A}_\perp) &\le [\chi + 0.02 r_{\rm cur}\zeta]\lambda^-_{\rm cur}, \\
\lambda_{\max}(\boldsymbol{A}_\perp) &\ge [\chi - 0.02 r_{\rm cur}\zeta]\lambda^-_{\rm cur}, \\
\lambda_{\min}(\boldsymbol{A}) &\ge [1 - 0.0101 r_{\rm cur}\zeta]\lambda^-_{\rm cur}, \\
\lambda_{\max}(\boldsymbol{A}) &\le [g + 0.0202 r_{\rm cur}\zeta]\lambda^-_{\rm cur}, \\
\lambda_{\max}(\boldsymbol{A}) &\ge [g - 0.02 r_{\rm cur}\zeta]\lambda^-_{\rm cur}.
\end{aligned}$$

Lemma 6.8. *Consider the setting of Corollary 6.7. In addition, also assume that*

1) $\hat{g} = 1.01g + 0.0001$ *and*
2) $\chi \le \min\left(\frac{g - 0.0001}{1.01g + 0.0001} - 0.0001, 1 - r_{\rm cur}\zeta - \frac{0.08}{0.25}\right)$.

Let $\hat{\lambda}_i := \lambda_i(\hat{\boldsymbol{D}})$. *Then, with probability at least* $1 - p_1 - p_2 - 4p_3$, *the following hold.*

1) *When* $\boldsymbol{G}_{\rm undet}$ *is not empty:* $\frac{\hat{\lambda}_1}{\hat{\lambda}_{r_{\rm cur}}} \le \hat{g}$, $\frac{\hat{\lambda}_1}{\hat{\lambda}_{r_{\rm cur}+1}} > \hat{g}$, *and* $\hat{\lambda}_{r_{\rm cur}+1} \ge \lambda_{\rm thresh}$.
2) *When* $\boldsymbol{G}_{\rm undet}$ *is empty:* $\frac{\hat{\lambda}_1}{\hat{\lambda}_{r_{\rm cur}}} \le \hat{g}$ *and* $\hat{\lambda}_{r_{\rm cur}+1} < \lambda_{\rm thresh} < \hat{\lambda}_{r_{\rm cur}}$.
3) *If* $\hat{r}_{\rm cur} = r_{\rm cur}$, *then* $\zeta_{\rm cur} \le \frac{\|\boldsymbol{H}\|}{\lambda_{\min}(\boldsymbol{A}) - \lambda_{\max}(\boldsymbol{A}_\perp) - \|\boldsymbol{H}\|} \le 0.75 r_{\rm cur}\zeta + \frac{0.08 r_{\rm cur}\zeta}{(1 - r_{\rm cur}\zeta - \chi)} \le r_{\rm cur}\zeta$.

Proof.

Fact 6.9. *From the bound on* χ, $\chi \le 1 - 0.0001 \le 1 - r_{\rm cur}\zeta$. *Thus, using Corollary 6.7,* $\lambda_{\min}(\boldsymbol{A}) > \lambda_{\max}(\boldsymbol{A}_\perp)$ *and so* $\lambda_{r_{\rm cur}}(\boldsymbol{D}) = \lambda_{\min}(\boldsymbol{A})$, $\lambda_{r_{\rm cur}+1}(\boldsymbol{D}) = \lambda_{\max}(\boldsymbol{A}_\perp)$, *and* $\lambda_1(\boldsymbol{D}) = \lambda_{\max}(\boldsymbol{A})$. *Recall:* $\lambda_1(.)$ *is the same as* $\lambda_{\max}(.)$.

Proof of item 1. Recall that $\hat{\boldsymbol{D}}$ and $\boldsymbol{D}$ are defined in (11) and (12). Using Weyl's inequality, Fact 6.9, and Corollary 6.7, with the probability given there,

$$\frac{\hat{\lambda}_1}{\hat{\lambda}_{r_{\rm cur}}} \le \frac{\lambda_{\max}(\boldsymbol{A}) + \|\boldsymbol{H}\|}{\lambda_{\min}(\boldsymbol{A}) - \|\boldsymbol{H}\|} \le \frac{g + 0.86 r_{\rm cur}\zeta}{1 - 0.85 r_{\rm cur}\zeta}$$

and

$$\frac{\hat{\lambda}_1}{\hat{\lambda}_{r_{\rm cur}+1}} > \frac{\lambda_{\max}(\boldsymbol{A}) - \|\boldsymbol{H}\|}{\lambda_{\max}(\boldsymbol{A}_\perp) + \|\boldsymbol{H}\|} > \frac{g - 0.85 r_{\rm cur}\zeta}{\chi + 0.85 r_{\rm cur}\zeta}$$

Thus, if

$$\frac{g + 0.85 r_{\rm cur}\zeta}{1 - 0.85 r_{\rm cur}\zeta} \le \hat{g} \le \frac{g - 0.85 r_{\rm cur}\zeta}{\chi + 0.85 r_{\rm cur}\zeta} \tag{13}$$

holds, we will be done. The above requires χ to be small enough so that the lower bound is not larger than the upper bound and it requires $\hat{g}$ to be appropriately set. Both are ensured by the assumptions in the lemma.

Since $\boldsymbol{G}_{\rm undet}$ is not empty, $\lambda^+_{\rm undet} = \chi\lambda^-_{\rm cur} > \lambda^-$ Thus, using Weyl's inequality followed by Corollary 6.7, with the probability given there,

$$\begin{aligned}\hat{\lambda}_{r_{\rm cur}+1} &\geq \lambda_{r_{\rm cur}+1}(\boldsymbol{D}) - \|\boldsymbol{H}\| = \lambda_{\max}(\boldsymbol{A}_\perp) - \|\boldsymbol{H}\| \\ &\geq [\chi - 0.02 r_{\rm cur}\zeta]\lambda^-_{\rm cur} - 0.83 r_{\rm cur}\zeta\lambda^-_{\rm cur} \\ &\geq (1 - 0.85 r_{\rm cur}\zeta)\lambda^- > \lambda_{\rm thresh}\end{aligned}$$

Proof of item 2. Since $\boldsymbol{G}_{\rm undet}$ is empty, $\lambda^+_{\rm undet} = 0$ and so $\chi = 0$. Thus, using Corollary 6.7, with probability given there,

$$\begin{aligned}\hat{\lambda}_{r_{\rm cur}+1} &\leq \lambda_{r_{\rm cur}+1}(\boldsymbol{D}) + \|\boldsymbol{H}\| = \lambda_{\max}(\boldsymbol{A}_\perp) + \|\boldsymbol{H}\| \\ &\leq 0 + 0.02 r_{\rm cur}\zeta\lambda^- + \|\boldsymbol{H}\| \leq 0.85 r_{\rm cur}\zeta\lambda^- \\ &< \lambda_{\rm thresh},\end{aligned}$$

$$\begin{aligned}\hat{\lambda}_{r_{\rm cur}} &\geq \lambda_{r_{\rm cur}}(\boldsymbol{D}) - \|\boldsymbol{H}\| = \lambda_{\min}(\boldsymbol{A}) - \|\boldsymbol{H}\| \\ &\geq \lambda^-_{\rm cur} - 0.085 r_{\rm cur}\zeta\lambda^-_{\rm cur} \geq (1 - 0.85 r_{\rm cur}\zeta)\lambda^- \\ &> \lambda_{\rm thresh},\end{aligned}$$

and

$$\frac{\hat{\lambda}_1}{\hat{\lambda}_{r_{\rm cur}}} \leq \frac{\lambda_{\max}(\boldsymbol{A}) + \|\boldsymbol{H}\|}{\lambda_{\min}(\boldsymbol{A}) - \|\boldsymbol{H}\|} \leq \frac{g + 0.85 r_{\rm cur}\zeta}{1 - 0.85 r_{\rm cur}\zeta} \leq \hat{g}$$

Proof of item 3. Using Fact 6.5 and Corollary 6.7, since $\hat{r}_{\rm cur} = r_{\rm cur}$ is assumed, we get

$$\begin{aligned}\zeta_{\rm cur} &\leq \frac{[0.75(1 - r_{\rm cur}\zeta - \chi) r_{\rm cur}\zeta + 0.08 r_{\rm cur}\zeta]\lambda^-_{\rm cur}}{\lambda^-_{\rm cur}[1 - 0.0101 r_{\rm cur}\zeta - \chi - 0.02 r_{\rm cur}\zeta - 0.83 r\zeta]} \\ &\leq \frac{0.75(1 - r\zeta - \chi) r_{\rm cur}\zeta + 0.08 r_{\rm cur}\zeta}{(1 - r_{\rm cur}\zeta - \chi)} \leq r_{\rm cur}\zeta \qquad (14)\end{aligned}$$

The last inequality used the bound on χ. ⊠

C. Proof of Theorem 3.5

The theorem is a direct consequence of using (10) and applying Lemma 6.8 for each of the k steps with the substitutions given in Definition 6.2; along with picking α appropriately. A detailed proof is in Appendix B-1(Supplementary document).

VII. Extensions: Correlated-PCA with partial subspace knowledge

The three main lemmas given above can also be used to provide a correctness result for the problem of correlated-PCA with partial subspace knowledge (correlated-PCA-partial). Consider the problem given in Sec. I-A, but with the following extra information. Split $\boldsymbol{P}$ as $\boldsymbol{P} = [\boldsymbol{P}_0 \ \boldsymbol{P}_{\rm new}]$ and $\boldsymbol{\Lambda}$ as $\boldsymbol{\Lambda} = {\rm diag}(\boldsymbol{\Lambda}_0, \boldsymbol{\Lambda}_{\rm new})$. Assume that $\hat{\boldsymbol{P}}_0$ is available and is such that ${\rm SE}(\hat{\boldsymbol{P}}_0, \boldsymbol{P}_0) \leq r_0\zeta$ and that $\hat{\boldsymbol{P}}_0$ was computed using $\boldsymbol{y}_t$'s before t_*. The goal is again to estimate ${\rm range}(\boldsymbol{P})$.

Let $r := {\rm rank}(\boldsymbol{P}_0)$, $r_{\rm new} = {\rm rank}(\boldsymbol{P}_{\rm new})$, $\lambda^+_{\rm new} = \lambda_{\max}(\boldsymbol{\Lambda}_{\rm new})$, $\lambda^-_{\rm new} = \lambda_{\min}(\boldsymbol{\Lambda}_{\rm new})$. Let $\boldsymbol{\Psi} := \boldsymbol{I} - \hat{\boldsymbol{P}}_0\hat{\boldsymbol{P}}_0{}'$. Recall that eigenvalues in $\boldsymbol{\Lambda}$ are in non-increasing order and hence $\lambda^- \leq \lambda^-_{\rm new} \leq \lambda^+_{\rm new} \leq \lambda_{\min}(\boldsymbol{\Lambda}_0) \leq \lambda_{\max}(\boldsymbol{\Lambda}_0) \leq \lambda^+$.

We recover ${\rm range}(\boldsymbol{P})$ by recovering ${\rm range}(\boldsymbol{P}_{\rm new})$ using Projection-EVD given in Algorithm 2 and setting $\hat{\boldsymbol{P}} = [\hat{\boldsymbol{P}}_0 \ \hat{\boldsymbol{P}}_{\rm new}]$. Also, instead of assuming $\|\boldsymbol{M}_{1,t}\boldsymbol{P}\| \leq q$, assume the following generalization: $\|\boldsymbol{M}_{1,t}\boldsymbol{P}_0\| \leq q_0$ and $\|\boldsymbol{M}_{1,t}\boldsymbol{P}_{\rm new}\| \leq q_{\rm new}$. As we will see below, this generalized bound helps us analyze dynamic robust PCA [2], [3], where, as we will see q_0 is much smaller than $q_{\rm new}$. We have the following result for correlated-PCA-partial.

Algorithm 2 Projection-EVD

Recover $\hat{\boldsymbol{P}}_{\rm new}$ as the eigenvectors of $\frac{1}{\alpha}\sum_{t=t_*+1}^{t_*+\alpha}(\boldsymbol{I} - \hat{\boldsymbol{P}}_0\hat{\boldsymbol{P}}_0{}')\boldsymbol{y}_t\boldsymbol{y}_t'(\boldsymbol{I} - \hat{\boldsymbol{P}}_0\hat{\boldsymbol{P}}_0{}')$ with eigenvalues larger than $\lambda_{\rm thresh}$.

Theorem 7.1. *Assume that $\hat{\boldsymbol{P}}_0$ is available and is such that ${\rm SE}(\hat{\boldsymbol{P}}_0, \boldsymbol{P}_0) \leq r_0\zeta$ and that $\hat{\boldsymbol{P}}_0$ was computed using $\boldsymbol{y}_t$'s before t_*. Recover $\hat{\boldsymbol{P}}_{\rm new}$ by Algorithm 2 and set $\hat{\boldsymbol{P}} = [\hat{\boldsymbol{P}}_0 \ \hat{\boldsymbol{P}}_{\rm new}]$. Pick a small scalar ζ so that $r(r\zeta) \leq 0.0001$, and $r(r\zeta)f \leq 0.01$. Suppose that $\boldsymbol{y}_t$ satisfies* (2) *and the following hold.*

1) *Model 1.1 on $\boldsymbol{\ell}_t$ holds and $\frac{\lambda^+_{\rm new}}{\lambda^-_{\rm new}} \leq g^+$. Define*

$$\alpha_0 := C\eta^2\frac{r_{\rm new}^2\log n}{(r_{\rm new}\zeta)^2}\max(g, q_{\rm new}g, q_{\rm new}^2g, q_0^2f,$$
$$q_0(r_{\rm new}\zeta)f, (r_{\rm new}\zeta)^2f, q_0\sqrt{fg}, q_{\rm new}(r\zeta)\sqrt{fg})^2,$$
$$C := 32\cdot 11\cdot 9/0.01^2$$

2) *Model 1.2 on $\boldsymbol{M}_t$ holds with the following generalization: $\|\boldsymbol{M}_{1,t}\boldsymbol{P}_0\| \leq q_0$ and $\|\boldsymbol{M}_{1,t}\boldsymbol{P}_{\rm new}\| \leq q_{\rm new}$, for any $\alpha \geq \alpha_0$*
3) *Set algorithm parameters $\lambda_{\rm thresh} = 0.95\lambda^-$ and $\alpha \geq \alpha_0$.*

Let $\text{numer} := 2.02 q_{\text{new}} g\sqrt{\frac{\beta}{\alpha}} + q_{\text{new}}^2 g\frac{\beta}{\alpha} + 2.02 q_0 (r_0\zeta) f\sqrt{\frac{\beta}{\alpha}} + q_0^2 f\frac{\beta}{\alpha} + 0.08 r_{\text{new}}\zeta$. *Then, with probability at least* $1 - 20n^{-10}$,

$$\text{SE}([\hat{\boldsymbol{P}}_0, \hat{\boldsymbol{P}}_{\text{new}}], \boldsymbol{P}_{\text{new}}) \le \frac{\text{numer}}{1 - 0.04 r_{\text{new}}\zeta - \text{numer}}$$

and $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \le r\zeta + \text{SE}([\hat{\boldsymbol{P}}_0, \hat{\boldsymbol{P}}_{\text{new}}], \boldsymbol{P}_{\text{new}})$.

Proof. This follows by generalizing Corollary 6.7 to incorporate the separate bounds on $||\boldsymbol{M}_{1,t}\boldsymbol{P}_0||$ and $||\boldsymbol{M}_{1,t}\boldsymbol{P}_{\text{new}}||$. ⊠

A. Application to dynamic robust PCA

As we explain next, the above result can be used to obtain guarantees for the subspace update step of a recently proposed online dynamic robust PCA algorithm called ReProCS (Recursive Projected Compressed Sensing) [2], [3]. In fact, the entire proof of correctness of ReProCS can be significantly shortened by applying this result. ReProCS assumes that the initial subspace is accurately known, and that the subspace changes over time, albeit slowly. To track the changes, ReProCS first projects the observed data vector $\boldsymbol{y}_t := \boldsymbol{\ell}_t + \boldsymbol{I}_{\mathcal{T}_t}\boldsymbol{x}_t$ orthogonal to the previous subspace estimate. This mostly nullifies $\boldsymbol{\ell}_t$ and gives us projected measurements of the sparse outlier $\boldsymbol{I}_{\mathcal{T}_t}\boldsymbol{x}_t$. It then solves a sparse recovery problem followed by support estimation and least squares estimation to recover $\mathcal{T}_t$ and $\boldsymbol{x}_t$. Finally, $\boldsymbol{I}_{\hat{\mathcal{T}}_t}\hat{\boldsymbol{x}}_t$ is subtracted out from $\boldsymbol{y}_t$ to get an estimate of $\boldsymbol{\ell}_t$, denoted $\hat{\boldsymbol{\ell}}_t$. The current and previous $\hat{\boldsymbol{\ell}}_t$'s are used to update the subspace estimate every α frames. Under simple assumptions, one can argue that $\mathcal{T}_t$ is exactly recovered and $\boldsymbol{x}_t$ is accurately recovered. With this, $\hat{\boldsymbol{\ell}}_t$ can be expressed as

$$\hat{\boldsymbol{\ell}}_t = \boldsymbol{\ell}_t - \boldsymbol{e}_t \text{ where } \boldsymbol{e}_t = \boldsymbol{I}_{\mathcal{T}_t}(\boldsymbol{\Phi}_{\mathcal{T}_t}{}'\boldsymbol{\Phi}_{\mathcal{T}_t})^{-1}\boldsymbol{I}_{\mathcal{T}_t}{}'\boldsymbol{\Phi}\boldsymbol{\ell}_t. \tag{15}$$

We specify $\boldsymbol{\Phi}$ below. If the support of the sparse outlier, $\mathcal{T}_t$, satisfies Model 1.3, then, recovering $\text{range}(\boldsymbol{P})$ from $\hat{\boldsymbol{\ell}}_t$'s satisfying (15) is clearly a correlated-PCA problem. The subspace update step consists of a subspace addition step and a subspace deletion step. In each subspace addition step, we estimate the newly added subspace, denoted $\boldsymbol{P}_{\text{new}}$, and we use $\boldsymbol{P}_0$ to denote the existing subspace. This is assumed to have been estimated accurately with $\text{SE}(\hat{\boldsymbol{P}}_0, \boldsymbol{P}_0) \le r_0\zeta$. We estimate $\boldsymbol{P}_{\text{new}}$ using projection-EVD applied K

	cEVD	EVD	PCP	A-M-RPCA
MEAN ERROR	0.0908	0.0911	1.0000	1.0000
AVERAGE TIME	0.0549	0.0255	0.2361	0.0810

TABLE I: Comparison of $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$ and execution time (in seconds): $n = 500$, $\alpha = 300$. A-M-RPCA: Alt-Min-RPCA.

	cEVD	EVD
MEAN ERROR	0.0189633	0.0189705
AVERAGE TIME	111.9170	52.8464

TABLE II: Table for $n = 10000$ case.

times. In the k-th proj-EVD step, one computes $\hat{\boldsymbol{P}}_{\text{new},k}$ using Algorithm 2 with $t_* = (k-1)\alpha$. It is assumed that we have a (not necessarily very accurate) estimate of $\boldsymbol{P}_{\text{new}}$, denoted $\hat{\boldsymbol{P}}_{\text{new},k-1}$. It can be shown that (a) $\boldsymbol{\Phi} = \boldsymbol{I} - \hat{\boldsymbol{P}}_0\hat{\boldsymbol{P}}_0{}' - \hat{\boldsymbol{P}}_{\text{new},k-1}\hat{\boldsymbol{P}}_{\text{new},k-1}{}'$; (b) $\hat{\boldsymbol{P}}_{\text{new},k-1}$ satisfies $\text{SE}([\hat{\boldsymbol{P}}_0, \hat{\boldsymbol{P}}_{\text{new},k-1}], \boldsymbol{P}_{\text{new}}) \le \zeta^+_{\text{new},k-1}$ with $\zeta^+_{\text{new},0} = 1$; (c) $\boldsymbol{\Phi}$ satisfies the restricted isometry property and $\|(\boldsymbol{\Phi}_{\mathcal{T}_t}{}'\boldsymbol{\Phi}_{\mathcal{T}_t})^{-1}\| \le \phi^+ = 1.2$ (this follows using a denseness assumption on the columns of $\boldsymbol{P}$ and $|\mathcal{T}_t| \le s$). With these facts, we are in the setting of Theorem 7.1. We can apply it with $\boldsymbol{y}_t \equiv \hat{\boldsymbol{\ell}}_t$, $\boldsymbol{w}_t \equiv \boldsymbol{e}_t$, $\boldsymbol{M}_{2,t} = \boldsymbol{I}_{\mathcal{T}_t}$, $\boldsymbol{M}_{1,t} = (\boldsymbol{\Phi}_{\mathcal{T}_t}{}'\boldsymbol{\Phi}_{\mathcal{T}_t})^{-1}\boldsymbol{I}_{\mathcal{T}_t}{}'\boldsymbol{\Phi}$ and with Model 1.2 replaced by Model 1.3. From the above facts, we have $q_0 = r_0\zeta \cdot \phi^+$, $q_{\text{new}} = \zeta^+_{\text{new},k-1} \cdot \phi^+$. Moreover, if β is such that $\sqrt{\frac{\beta}{\alpha}} \le \frac{0.1}{2.02g}$, then, using Theorem 7.1, it can be shown that $\text{SE}([\hat{\boldsymbol{P}}_0, \hat{\boldsymbol{P}}_{\text{new},1}], \boldsymbol{P}_{\text{new}}) \le \zeta^+_{\text{new},1} = 0.6$; that $\text{SE}([\hat{\boldsymbol{P}}_0, \hat{\boldsymbol{P}}_{\text{new},k}], \boldsymbol{P}_{\text{new}}) \le \zeta^+_{\text{new},k} \le 0.6$ and that $\zeta^+_{\text{new},k} \le 1.1 \cdot 0.11 \cdot \zeta^+_{\text{new},k-1} + 1.1 \cdot 0.06 r_{\text{new}}\zeta$. Thus, after $K = \lceil \frac{\log(1-0.066)r_{\text{new}}\zeta}{\log(0.6)} \rceil$ projection-EVD steps, the subspace error $\text{SE}([\hat{\boldsymbol{P}}_0, \hat{\boldsymbol{P}}_{\text{new},K}], \boldsymbol{P}_{\text{new}}) \le \zeta^+_{\text{new},K} \le r_{\text{new}}\zeta$ and hence $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \le (r_0 + r_{\text{new}})\zeta$.

In the subspace deletion step, one re-estimates the entire subspace in order to remove the deleted directions. This is done using either EVD or cluster-EVD. At this time $q_{\text{new}} = \zeta^+_K \cdot \phi^+ \le r_{\text{new}}\zeta \cdot \phi^+$ and $q_0 = r_0\zeta \cdot \phi^+$ are both of the same order. So one can apply Theorem 2.1 or Theorem 3.5 with $q = (r_0 + r_{\text{new}})\zeta \cdot \phi^+$ to obtain a guarantee for this step.

VIII. Numerical experiments

We use the PCA-SDDO problem as our case study example. Thus, $\boldsymbol{y}_t$ satisfied (7). We compare EVD and cluster-EVD with PCP [19], solved using [25], and with Alt-Min-RPCA [21] (implemented using code from the authors' webpage). For both PCP and Alt-Min-RPCA, $\hat{\boldsymbol{P}}$ is recovered as the top r eigenvectors of of the estimated $\boldsymbol{L}$.To show the advantage of EVD or cluster-EVD, we let $\boldsymbol{\ell}_t = \boldsymbol{P}\boldsymbol{a}_t$ with columns of $\boldsymbol{P}$ being sparse. These were chosen as the first $r = 5$ columns of the identity matrix. We generate $\boldsymbol{a}_t$'s iid uniformly with zero mean and covariance matrix $\boldsymbol{\Lambda} = diag(100, 100, 100, 0.1, 0.1)$. Thus the

	cEVD	EVD	PCP	A-M-RPCA
MEAN ERROR	0.1615	0.1618	0.9932	1.0000
AVERAGE TIME	0.0541	0.0254	1.0705	0.6591

TABLE III: Table for case with added $\boldsymbol{\nu}_t$. A-M-RPCA: Alt-Min-RPCA.

Fig. 2: A low-rankified escalator video overlaid with a moving object (shown with a blue arrow). Frames 2, 50, 100 are shown.

condition number $f = 1000$. The clustering assumption holds with $\vartheta = 2$, $g^+ = 1$ and $\chi^+ = 0.001$. The noise $\boldsymbol{w}_t$ is generated as $\boldsymbol{w}_t = \boldsymbol{I}_{\mathcal{T}_t} \boldsymbol{M}_{s,t} \boldsymbol{\ell}_t$ with $\mathcal{T}_t$ generated to satisfy Model 1.3 with $s = 5$, $\rho = 2$, and $\tilde{\beta} = 1$; and the entries of $\boldsymbol{M}_{s,t}$ being iid $\mathcal{N}(0, q^2)$ with $q = 0.01$.

For our first experiment, $n = 500$. EVD and c-EVD (Algorithm 1) were implemented with $\alpha = 300$, $\lambda_{\text{thresh}} = 0.095$, $\hat{g} = 3$. 10000-time Monte Carlo averaged values of $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$ and execution time are shown in Table I. Since the columns of $\boldsymbol{P}$ are sparse, neither of PCP or Alt-Min-RPCA work. Both have average SE close to one whereas the average SE of c-EVD and EVD is 0.0908 and 0.0911 respectively. Also, both EVD and c-EVD are much faster than even Alt-Min-RPCA, which is known to be a fast algorithm for RPCA. In our second experiment, we used $n = 10000$ and $\alpha = 5000$. Everything else is the same as above. The results are shown in Table II.

We also did an experiment with the settings of the first expt, but $\boldsymbol{P}$ dense. In this case, EVD and c-EVD errors were similar, but PCP and Alt-Min-RPCA errors were less than 10^{-5}.

In the third experiment, we generate data using (3) with $\boldsymbol{\ell}_t$ and $\boldsymbol{w}_t$ as above and with $\boldsymbol{\nu}_t$ being iid zero mean uniform with variance 0.001. The results are in Table III. Once again, PCP and Alt-Min-RPCA fail because $\boldsymbol{P}$ is sparse.

We do our last experiment, with a low-rankified real video sequence. We chose the escalator sequence from http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html since the video changes are only in the region where the escalator moves (and hence can be modeled as being sparse). We made it exactly low-rank by retaining its top 5 eigenvectors and projecting onto their subspace. This resulted in a data matrix $\boldsymbol{L}$ of size $n \times r$ with $n = 20800$ and $r = 5$. We overlaid a simulated moving foreground block on it. The intensity of the moving block was controlled to ensure that q is not too large. Three frames of this video are shown in the rows in Fig. 2. We estimated $\hat{\boldsymbol{P}}$ using EVD, c-EVD, PCP and Alt-Min-RPCA. We let $\boldsymbol{P}$ as the eigenvectors of the low-rankified video with nonzero eigenvalues and computed $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$. The errors are displayed in Table IV. Since n is very large, the difference in speed is most apparent in this case.

	cEVD	EVD	PCP	A-M-RPCA
LEVEL-1 ERROR	0.3626	0.3821	0.4970	0.4846
LEVEL-2 ERROR	0.6204	0.6319	0.4973	0.6942
EXEC. TIME	0.0613	0.0223	1.6784	5.5144

TABLE IV: SE comparison for real video data, with two settings of the moving object's intensity. A-M-RPCA: Alt-Min-RPCA.

From all the above experiments, cluster-EVD outperforms EVD. The advantage in averaged error is not as much as our theorems predict. As explained in Sec. IV, one reason is that the constant in the required lower bounds on α is very large. It is hard to pick an α that is this large and still only $O(\log n)$ (it will need n to be extremely large). Secondly, both guarantees are only sufficient conditions.

IX. Conclusions

We studied the problem of PCA in corrupting noise that is correlated with the data (data-dependent noise). We showed that, under simple assumptions on the data-dependency (or data-noise correlation), for a fixed desired subspace error level, the sample complexity of the simple-EVD based solution to PCA scales as $f^2 r^2 \log n$ where f is the condition number of the true data's covariance matrix and r is its rank. We developed and analyzed a generalization of EVD, called cluster-EVD. Under a clustered eigenvalues' assumption, we argued that its guarantee has a much weaker dependence on f. To our knowledge, there is no other result on this problem. Hence, we provided a detailed comparison of the two results with other approaches to solving the example applications of this problem - PCA in missing data and PCA with sparse data-dependent corruptions. We also obtained guarantees for correlated-PCA with partial subspace knowledge; and we briefly explained how this result can be used to significantly simplify the correctness proof of the ReProCS algorithm for online robust PCA given in [3].

References

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

[2] C. Qiu, N. Vaswani, B. Lois, and L. Hogben, "Recursive robust pca or recursive sparse recovery in large but structured noise," *IEEE Trans. Info. Th.*, vol. 60, no. 8, pp. 5007–5039, August 2014.

[3] B. Lois and N. Vaswani, "Online matrix completion and online robust pca," in *IEEE Intl. Symp. Info. Th. (ISIT)*, 2015.

[4] J. Zhan, B. Lois, and N. Vaswani, "Online (and Offline) Robust PCA: Novel Algorithms and Performance Guarantees," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016, also at ArXiv: 1601.07985.

[5] A. H. Bentbib and A. Kanber, "Block power method for svd decomposition," *Analele Stiintifice Ale Universitatii Ovidius Constanta-Seria Matematica*, vol. 23, no. 2, pp. 45–58, 2015.

[6] G. H. Golub and H. A. Van der Vorst, "Eigenvalue computation in the 20th century," *Journal of Computational and Applied Mathematics*, vol. 123, no. 1, pp. 35–65, 2000.

[7] R. Arora, A. Cotter, and N. Srebro, "Stochastic optimization of pca with capped msg," in *Adv. Neural Info. Proc. Sys. (NIPS)*, 2013, pp. 1815–1823.

[8] O. Shamir, "A stochastic pca and svd algorithm with an exponential convergence rate," *arXiv:1409.2848*, 2014.

[9] Christos Boutsidis, Dan Garber, Zohar Karnin, and Edo Liberty, "Online principal components analysis," in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2015, pp. 887–901.

[10] A. Balsubramani, S. Dasgupta, and Y. Freund, "The fast convergence of incremental pca," in *Adv. Neural Info. Proc. Sys. (NIPS)*, 2013, pp. 3174–3182.

[11] Z. Karnin and booktitle=Proceedings of the 28th Annual Conference on Computational Learning Theory (COLT) pages=505–509 year=2015 Liberty, E., "Online pca with spectral bounds," .

[12] I. Mitliagkas, C. Caramanis, and P. Jain, "Memory limited, streaming pca," in *Adv. Neural Info. Proc. Sys. (NIPS)*, 2013, pp. 2886–2894.

[13] M. Fazel, "Matrix rank minimization with applications," *PhD thesis, Stanford University*, 2002.

[14] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Found. of Comput. Math*, , no. 9, pp. 717–772, 2008.

[15] K. Lee and Y. Bresler, "Admira: Atomic decomposition for minimum rank approximation," *IEEE Transactions on Information Theory*, vol. 56, no. 9, September 2010.

[16] R.H. Keshavan, A. Montanari, and Sewoong Oh, "Matrix completion from a few entries," *IEEE Trans. Info. Th.*, vol. 56, no. 6, pp. 2980–2998, 2010.

[17] P. Netrapalli, P. Jain, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Symposium on Theory of Computing (STOC)*, 2013.

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

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

[20] D. Hsu, S.M. Kakade, and T. Zhang, "Robust matrix decomposition with sparse corruptions," *IEEE Trans. Info. Th.*, Nov. 2011.

[21] P. Netrapalli, U N Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust pca," in *Adv. Neural Info. Proc. Sys. (NIPS)*, 2014.

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

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

[24] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. iii," *SIAM J. Numer. Anal.*, vol. 7, pp. 1–46, Mar. 1970.

[25] Z. Lin, M. Chen, and Y. Ma, "Alternating direction algorithms for l1 problems in compressive sensing," Tech. Rep., University of Illinois at Urbana-Champaign, November 2009.

APPENDIX A
PROOF OF HOEFFDING LEMMA, LEMMA 6.6

The following lemma, which is a modification of [2, Lemma 8.15], will be used in our proof. It is proved in Appendix B-2 (Supplementary document). The proof uses [2, Lemma 2.10].

Lemma A.1. *Given* $\zeta_{\rm det} \leq r\zeta$.

1) $\|\boldsymbol{\Psi}\boldsymbol{G}_{\rm det}\| \leq r\zeta$ *and* $\|\boldsymbol{\Psi}\boldsymbol{G}_{\rm cur}\| \leq 1$.
2) $\sqrt{1-(r\zeta)^2} \leq \sigma_i(\boldsymbol{R}_{\rm cur}) = \sigma_i(\boldsymbol{\Psi}\boldsymbol{G}_{\rm cur}) \leq 1$ *and* $\sqrt{1-(r\zeta)^2} \leq \sigma_i(\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}) \leq 1$
3) $\|\boldsymbol{E}_{\rm cur}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}\| \leq \dfrac{(r\zeta)^2}{\sqrt{1-(r\zeta)^2}}$
4)

$$\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi} = [\boldsymbol{\Psi}\boldsymbol{G}_{\rm det}\ \boldsymbol{\Psi}\boldsymbol{G}_{\rm cur}\ \boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}] \begin{bmatrix} \boldsymbol{\Lambda}_{\rm det} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_{\rm cur} & \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Lambda}_{\rm undet} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Psi}\boldsymbol{G}_{\rm det} \\ \boldsymbol{\Psi}\boldsymbol{G}_{\rm cur} \\ \boldsymbol{\Psi}\boldsymbol{G}_{\rm undet} \end{bmatrix}'$$

with $\lambda_{\max}(\boldsymbol{\Lambda}_{\rm det}) \leq \lambda^+$, $\lambda^-_{\rm cur} \leq \lambda_{\min}(\boldsymbol{\Lambda}_{\rm cur}) \leq \lambda_{\max}(\boldsymbol{\Lambda}_{\rm cur}) \leq \lambda^+_{\rm cur}$, $\lambda_{\max}(\boldsymbol{\Lambda}_{\rm undet}) \leq \lambda^+_{\rm undet}$.

5) *Using the first four claims, it is easy to see that*
   a) $\|\boldsymbol{E}_{{\rm cur},\perp}{}'\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}\boldsymbol{E}_{{\rm cur},\perp}\| \leq (r\zeta)^2\lambda^+ + \lambda^+_{\rm undet}$
   b) $\|\boldsymbol{E}_{{\rm cur},\perp}{}'\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}\boldsymbol{E}_{\rm cur}\| \leq (r\zeta)^2\lambda^+ + \frac{(r\zeta)^2}{\sqrt{1-(r\zeta)^2}}\lambda^+_{\rm undet}$
   c) $\|\boldsymbol{\Psi}\boldsymbol{\Sigma}\| \leq (r\zeta)\lambda^+ + \lambda^+_{\rm cur}$ *and* $\|\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{M}_{1,t}{}'\| \leq q((r\zeta)\lambda^+ + \lambda^+_{\rm cur})$
   d) $\|\boldsymbol{M}_{1,t}\boldsymbol{\Sigma}\| \leq q\lambda^+$ *and* $\|\boldsymbol{M}_{1,t}\boldsymbol{\Sigma}\boldsymbol{M}_{1,t}{}'\| \leq q^2\lambda^+$

   If $\hat{\boldsymbol{G}}_{\rm det} = \boldsymbol{G}_{\rm det} = [.]$, *then all the terms containing* $(r\zeta)$ *disappear.*
6) $\lambda_{\min}(A+B) \geq \lambda_{\min}(A) + \lambda_{\min}(B)$
7) *Let* $\boldsymbol{a}_t := \boldsymbol{P}'\boldsymbol{\ell}_t$, $\boldsymbol{a}_{t,\rm det} := \boldsymbol{G}_{\rm det}{}'\boldsymbol{\ell}_t$, $\boldsymbol{a}_{t,\rm cur} := \boldsymbol{G}_{\rm cur}{}'\boldsymbol{\ell}_t$ *and* $\boldsymbol{a}_{t,\rm undet} := \boldsymbol{G}_{\rm undet}{}'\boldsymbol{\ell}_t$. *Also let* $\boldsymbol{a}_{t,\rm rest} := [\boldsymbol{a}_{t,\rm cur}{}', \boldsymbol{a}_{t,\rm undet}{}']'$. *Then* $\|\boldsymbol{a}_{t,\rm rest}\|^2 \leq r\eta\lambda^+_{\rm cur}$ *and* $\|\boldsymbol{a}_{t,\rm det}\|^2 \leq \|\boldsymbol{a}_t\|^2 \leq r\eta\lambda^+$.
8) $\sigma_{\min}(\boldsymbol{E}_{{\rm cur},\perp}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet})^2 \geq 1-(r\zeta)^2 - \frac{(r\zeta)^2}{\sqrt{1-(r\zeta)^2}}$.

The following corollaries of the matrix Hoeffding inequality [22], proved in [2], will be used in the proof.

Corollary A.2. *Given an* α*-length sequence* $\{\boldsymbol{Z}_t\}$ *of random Hermitian matrices of size* $n \times n$, *a r.v.* X, *and a set* $\mathcal{C}$ *of values that* X *can take. For all* $X \in \mathcal{C}$, *(i)* $\boldsymbol{Z}_t$*'s are conditionally independent given* X;

(ii) $\mathbb{P}(b_1\boldsymbol{I} \preceq \boldsymbol{Z}_t \preceq b_2\boldsymbol{I}|X) = 1$ *and (iii)* $b_3\boldsymbol{I} \preceq \mathbb{E}[\frac{1}{\alpha}\sum_t \boldsymbol{Z}_t|X] \preceq b_4\boldsymbol{I}$*. For any* $\epsilon > 0$*, for all* $X \in \mathcal{C}$*,*

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{\alpha}\sum_t \boldsymbol{Z}_t\right) \leq b_4 + \epsilon \middle| X\right) \geq 1 - n\exp\left(\frac{-\alpha\epsilon^2}{8(b_2-b_1)^2}\right),$$
$$\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{\alpha}\sum_t \boldsymbol{Z}_t\right) \geq b_3 - \epsilon \middle| X\right) \geq 1 - n\exp\left(\frac{-\alpha\epsilon^2}{8(b_2-b_1)^2}\right).$$

Corollary A.3. *Given an* α*-length sequence* $\{\boldsymbol{Z}_t\}$ *of random matrices of size* $n_1 \times n_2$*. For all* $X \in \mathcal{C}$*, (i)* $\boldsymbol{Z}_t$*'s are conditionally independent given* X*; (ii)* $\mathbb{P}(\|\boldsymbol{Z}_t\| \leq b_1|X) = 1$ *and (iii)* $\|\mathbb{E}[\frac{1}{\alpha}\sum_t \boldsymbol{Z}_t|X]\| \leq b_2$*. For any* $\epsilon > 0$*, for all* $X \in \mathcal{C}$*,*

$$\mathbb{P}\left(\left\|\frac{1}{\alpha}\sum_t \boldsymbol{Z}_t\right\| \leq b_2 + \epsilon \middle| X\right) \geq 1 - (n_1+n_2)\exp\left(\frac{-\alpha\epsilon^2}{32{b_1}^2}\right).$$

Proof of Lemma 6.6. Recall that we are given $\hat{\boldsymbol{G}}_{\text{det}}$ that was computed using (some or all) $\boldsymbol{y}_t$'s for $t \leq t_*$ and that satisfies $\zeta_{\text{det}} \leq r\zeta$. From (2), $\boldsymbol{y}_t$ is a linear function of $\boldsymbol{\ell}_t$. Thus, we can let $X := \{\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots \boldsymbol{\ell}_{t_*}\}$ denote all the random variables on which the event $\{\zeta_{\text{det}} \leq r\zeta\}$ depends. In each item of this proof, we need to lower bound the probability of the desired event conditioned on $\zeta_{\text{det}} \leq r\zeta$. To do this, we first lower bound the probability of the event conditioned on X that is such that $X \in \{\zeta_{\text{det}} \leq r\zeta\}$. We get a lower bound that does not depend on X as long as $X \in \{\zeta_{\text{det}} \leq r\zeta\}$. Thus, the same probability lower bound holds conditioned on $\{\zeta_{\text{det}} \leq r\zeta\}$.

Fact A.4. *For an event* $\mathcal{E}$ *and random variable* X*,* $\mathbb{P}(\mathcal{E}|X) \geq p$ *for all* $X \in \mathcal{C}$ *implies that* $\mathbb{P}(\mathcal{E}|X \in \mathcal{C}) \geq p$*.*

Proof of Lemma 6.6, item 1. Let

$$\text{term} := \frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{w}_t{}' = \frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t'\boldsymbol{M}_{1,t}{}'\boldsymbol{M}_{2,t}{}'$$

Since $\boldsymbol{\Psi}$ is a function of X, since $\boldsymbol{\ell}_t$'s used in the summation above are independent of X and $\mathbb{E}[\boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'] = \boldsymbol{\Sigma}$,

$$\mathbb{E}[\text{term}|X] = \frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{M}_{1,t}{}'\boldsymbol{M}_{2,t}{}'$$

Next, we use Cauchy-Schwartz for matrices:

$$\left\|\sum_{t=1}^{\alpha} \boldsymbol{X}_t\boldsymbol{Y}_t{}'\right\|^2 \leq \lambda_{\max}\left(\sum_{t=1}^{\alpha} \boldsymbol{X}_t\boldsymbol{X}_t{}'\right)\lambda_{\max}\left(\sum_{t=1}^{\alpha} \boldsymbol{Y}_t\boldsymbol{Y}_t{}'\right) \tag{16}$$

Using (16), with $\boldsymbol{X}_t = \boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{M}_{1,t}{}'$ and $\boldsymbol{Y}_t = \boldsymbol{M}_{2,t}$, followed by using $\sqrt{\|\frac{1}{\alpha}\sum_t \boldsymbol{X}_t\boldsymbol{X}_t'\|} \leq \max_t \|\boldsymbol{X}_t\|$, Model 1.2 with $\boldsymbol{A}_t \equiv \boldsymbol{I}$, and Lemma A.1,

$$\begin{aligned}\|\mathbb{E}[\mathrm{term}|X]\| &\leq \max_t \|\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{M}_{1,t}{}'\|\sqrt{\frac{\beta}{\alpha}} \\ &\leq q((r\zeta)\lambda^+ + \lambda^+_{\mathrm{cur}})\sqrt{\frac{\beta}{\alpha}}\end{aligned}$$

for all $X \in \{\zeta_{\mathrm{det}} \leq r\zeta\}$. To bound $\|\boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{w}_t'\|$, rewrite it as $\boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{w}_t' = [\boldsymbol{\Psi}\boldsymbol{G}_{\mathrm{det}}a_{t,\mathrm{det}} + \boldsymbol{\Psi}\boldsymbol{G}_{\mathrm{rest}}a_{t,\mathrm{rest}}][a'_{t,\mathrm{det}}\boldsymbol{G}'_{\mathrm{det}} + a'_{t,\mathrm{rest}}\boldsymbol{G}'_{\mathrm{rest}}]\boldsymbol{M}'_{1,t}\boldsymbol{M}'_{2,t}$. Thus, using $\|\boldsymbol{M}_{2,t}\| \leq 1$, $\|\boldsymbol{M}_{1,t}\boldsymbol{P}\| \leq q < 1$, and Lemma A.1,

$$\|\boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{w}_t'\| \leq qr\eta((r\zeta)\lambda^+ + \lambda^+_{\mathrm{cur}} + (r\zeta)\sqrt{\lambda^+\lambda^+_{\mathrm{cur}}} + \sqrt{\lambda^+\lambda^+_{\mathrm{cur}}})$$

holds w.p. one when $\{\zeta_{\mathrm{det}} \leq r\zeta\}$.

Finally, conditioned on X, the individual summands in term are conditionally independent. Using matrix Hoeffding, Corollary A.3, followed by Fact A.4, the result follows.

Proof of Lemma 6.6, item 2.

$$\mathbb{E}[\frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t'|X] = \frac{1}{\alpha}\sum_t \boldsymbol{M}_{2,t}\boldsymbol{M}_{1,t}\boldsymbol{\Sigma}\boldsymbol{M}_{1,t}{}'\boldsymbol{M}_{2,t}{}'$$

By Lemma A.1, $\|\boldsymbol{M}_{1,t}\boldsymbol{\Sigma}\boldsymbol{M}_{1,t}{}'\| \leq q^2\lambda^+$. Thus, using Model 1.2 with $\boldsymbol{A}_t \equiv \boldsymbol{M}_{1,t}\boldsymbol{\Sigma}\boldsymbol{M}_{1,t}{}'$,

$$\|\mathbb{E}[\frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t'|X]\| \leq \frac{\beta}{\alpha}q^2\lambda^+.$$

Using Model 1.2 and Lemma A.1,

$$\|\boldsymbol{w}_t\boldsymbol{w}_t'\| = \|\boldsymbol{M}_{2,t}\boldsymbol{M}_{1,t}\boldsymbol{P}\boldsymbol{a}_t\|^2 \leq q^2\eta r\lambda^+.$$

Conditional independence of the summands holds as before. Thus, using Corollary A.3 and Fact A.4, the result follows.

Proof of Lemma 6.6, item 3.

$$\begin{aligned}&\mathbb{E}[\frac{1}{\alpha}\sum_t \boldsymbol{E}_{\mathrm{cur}}\boldsymbol{E}_{\mathrm{cur}}{}'\boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'\boldsymbol{\Psi}\boldsymbol{E}_{\mathrm{cur},\perp}\boldsymbol{E}_{\mathrm{cur},\perp}{}'\|\|X] \\ &= \boldsymbol{E}_{\mathrm{cur}}\boldsymbol{E}_{\mathrm{cur}}{}'\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}\boldsymbol{E}_{\mathrm{cur},\perp}\boldsymbol{E}_{\mathrm{cur},\perp}{}'\end{aligned}$$

Using Lemma A.1, $\|\boldsymbol{E}_{\mathrm{cur}}\boldsymbol{E}_{\mathrm{cur}}{}'\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}\boldsymbol{E}_{\mathrm{cur},\perp}\boldsymbol{E}_{\mathrm{cur},\perp}{}'\| \leq (r\zeta)^2\lambda^+ + \frac{(r\zeta)^2}{\sqrt{1-(r\zeta)^2}}\lambda^+_{\mathrm{undet}}$ when $\{\zeta_{\mathrm{det}} \leq r\zeta\}$. Also, $\|\boldsymbol{E}_{\mathrm{cur}}{}'\boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'\boldsymbol{\Psi}\boldsymbol{E}_{\mathrm{cur},\perp}\| \leq \|\boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'\boldsymbol{\Psi}\| \leq \eta r((r\zeta)^2\lambda^+ + \lambda^+_{\mathrm{cur}} + 2(r\zeta)\sqrt{\lambda^+\lambda^+_{\mathrm{cur}}}) := b_{prob}$ holds w.p. one when $\{\zeta_{\mathrm{det}} \leq r\zeta\}$. In the above bound, the first inequality is used to get a loose bound, but one that

will also apply for the proofs of the later items given below. The rest is the same as in the proofs of the earlier parts.

Proof of Lemma 6.6, item 4. Using Ostrowski's theorem,

$$\begin{aligned}\lambda_{\min}(\mathbb{E}[\boldsymbol{A}|X]) &= \lambda_{\min}(\boldsymbol{E}_{\rm cur}{}'\boldsymbol{\Psi}(\boldsymbol{\Sigma})\boldsymbol{\Psi}\boldsymbol{E}_{\rm cur}) \\ &\geq \lambda_{\min}(\boldsymbol{E}_{\rm cur}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm cur}\boldsymbol{\Lambda}_{\rm cur}\boldsymbol{G}_{c}ur'\boldsymbol{\Psi}\boldsymbol{E}_{\rm cur}) \\ &= \lambda_{\min}(\boldsymbol{R}_{\rm cur}\boldsymbol{\Lambda}_{\rm cur}\boldsymbol{R}_{\rm cur}{}') \\ &\geq \lambda_{\min}(\boldsymbol{R}_{\rm cur}\boldsymbol{R}_{\rm cur}{}')\lambda_{\min}(\boldsymbol{\Lambda}_{\rm cur}) \geq (1-(r\zeta)^2)\lambda^{-}_{\rm cur}\end{aligned}$$

for all $X \in \{\zeta_{\rm det} \leq r\zeta\}$. Ostrowski's theorem is used to get the second-last inequality, while Lemma A.1 helps get the last one.

As in the proof of item 3, $\|\boldsymbol{E}_{\rm cur}{}'\boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'\boldsymbol{\Psi}\boldsymbol{E}_{\rm cur}\| \leq \|\boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'\boldsymbol{\Psi}\| \leq b_{prob}$ holds w.p. one when $\{\zeta_{\rm det} \leq r\zeta\}$. Conditional independence of the summands holds as before. Thus, by matrix Hoeffding, Corollary A.2, the result follows.

Proof of Lemma 6.6, item 5. By Lemma A.1,

$$\begin{aligned}\lambda_{\max}(\mathbb{E}[\boldsymbol{A}_{\perp}|X]) &= \lambda_{\max}(\boldsymbol{E}_{{\rm cur},\perp}{}'\boldsymbol{\Psi}\boldsymbol{\Sigma}\boldsymbol{\Psi}\boldsymbol{E}_{{\rm cur},\perp}) \\ &\leq ((r\zeta)^2\lambda^{+} + \lambda^{+}_{\rm undet})\end{aligned}$$

when $\{\zeta_{\rm det} \leq r\zeta\}$. The rest of the proof is the same as that of the previous part.

Proof of Lemma 6.6, item 6. Using Ostrowski's theorem, $\lambda_{\max}(\mathbb{E}[\boldsymbol{A}_{\perp}|X]) \geq \lambda_{\max}(\boldsymbol{E}_{{\rm cur},\perp}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}\boldsymbol{\Lambda}_{\rm undet}\boldsymbol{G}_{\rm undet}{}'\boldsymbol{\Psi}\boldsymbol{E}_{{\rm cur},\perp}) \geq \lambda_{\min}(\boldsymbol{E}_{{\rm cur},\perp}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}\boldsymbol{G}_{\rm undet}{}'\boldsymbol{\Psi}\boldsymbol{E}_{{\rm cur},\perp})\lambda_{\max}(\boldsymbol{\Lambda}_{\rm undet})$. By definition, $\lambda_{\max}(\boldsymbol{\Lambda}_{\rm undet}) = \lambda^{+}_{\rm undet}$. By Lemma A.1, $\lambda_{\min}(\boldsymbol{E}_{{\rm cur},\perp}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}\boldsymbol{G}_{\rm undet}{}'\boldsymbol{\Psi}\boldsymbol{E}_{{\rm cur},\perp}) = \sigma_{\min}(\boldsymbol{E}_{{\rm cur},\perp}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet})^2 \geq (1-(r\zeta)^2 - \frac{(r\zeta)^2}{\sqrt{1-(r\zeta)^2}})$ when $\{\zeta_{\rm det} \leq r\zeta\}$. The rest of the proof is the same as above.

Proof of Lemma 6.6, item 7. Using Ostrowski's theorem and Lemma A.1, $\lambda_{\max}(\mathbb{E}[\boldsymbol{A}|X]) \geq \lambda_{\max}(\boldsymbol{E}_{\rm cur}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm cur}\boldsymbol{\Lambda}_{\rm cur}\boldsymbol{G}_{\rm cur}{}'\boldsymbol{\Psi}\boldsymbol{E}_{\rm cur}) \geq \lambda_{\min}(\boldsymbol{R}_{\rm cur}\boldsymbol{R}_{\rm cur}{}')\lambda_{\max}(\boldsymbol{\Lambda}_{\rm cur}) \geq (1-(r\zeta)^2)\lambda^{+}_{\rm cur}$ when $\{\zeta_{\rm det} \leq r\zeta\}$. The rest of the proof is the same as above. ⊠

APPENDIX B
SUPPLEMENTARY MATERIAL

1) Detailed Proof of Theorem 3.5: Recall that we need to show that $\zeta_k \leq r_k\zeta$. Assume the substitutions given in Definition 6.2. We will use induction.

Consider a $k < \vartheta$. For the k-th step, assume that $\zeta_i \leq r_i\zeta$ for $i = 1, 2, \ldots, k-1$. Thus, using (10), $\zeta_{\rm det} \leq r\zeta$ and so Lemma 6.8 is applicable. We first show that $\hat{r}_k = r_k$ and that Algorithm 1 does not stop (proceeds to $(k+1)$-th step). From Algorithm 1, $\hat{r}_k = r_k$ if $\frac{\hat{\lambda}_1}{\hat{\lambda}_{r_k}} \leq \hat{g}$, and $\frac{\hat{\lambda}_1}{\hat{\lambda}_{r_k+1}} > \hat{g}$. Also it will not stop if $\hat{\lambda}_{r_k+1} \geq \lambda_{\rm thresh}$. Since $k < \vartheta$, $\boldsymbol{G}_{\rm undet}$ is not empty. Thus, item 1 of Lemma 6.8 shows that all these hold. Hence $\hat{r}_k = r_k$ and algorithm does not stop w.p. at least $1 - p_1 - p_2 - 4p_3$. Thus, by item 3 of the same lemma, with the same probability, $\zeta_k \leq r_k\zeta$.

Now consider $k = \vartheta$. We first show $\hat{r}_k = r_k$ and that Algorithm 1 does stop, i.e., $\hat{\vartheta} = \vartheta$. This will be true if $\frac{\hat{\lambda}_1}{\hat{\lambda}_{r_k}} \leq \hat{g}$ and $\hat{\lambda}_{r_k+1} < \lambda_{\rm thresh}$. For $k = \vartheta$, $\boldsymbol{G}_{\rm undet}$ is empty. Thus, item 2 of Lemma 6.8 shows that this holds w.p. at least $1 - p_1 - p_2 - 4p_3$. Thus, by item 3 of the same lemma, with the same probability, $\zeta_k \leq r_k\zeta$.

Thus, using the union bound, w.p. at least $1 - \vartheta(p_1 + p_2 + 4p_3)$, $\hat{r}_k = r_k$ and $\zeta_k \leq r_k\zeta$ for all k. Using (9), this implies that ${\rm SE} \leq r\zeta$ with the same probability.

Finally, the choice $\alpha \geq \alpha_0$, implies that $p_1 \leq \frac{1}{\vartheta}2n^{-10}$, $p_2 \leq \frac{1}{\vartheta}2n^{-10}$, $p_3 \leq \frac{1}{\vartheta}2n^{-10}$. Hence ${\rm SE} \leq r\zeta$ w.p. at least $1 - 12n^{-10}$. We work this out for p_1 below. The others follow similarly.

Recall that $p_1 = 2n\exp(-\alpha\frac{\epsilon^2}{32b_{prob}^2})$, $\epsilon = 0.01(r\zeta)\lambda^-$ and $b_{prob} = \eta rq((r\zeta)\lambda^+ + \lambda^+_{\rm cur} + (r\zeta)\sqrt{\lambda^+\lambda^+_{\rm cur}} + \sqrt{\lambda^+\lambda^+_{\rm cur}})$. Thus, $\frac{b_{prob}^2}{(\lambda^-)^2} \leq (4\eta r\max(q(r\zeta)f, qg, q\sqrt{fg}, q(r\zeta)\sqrt{fg}))^2 \leq 16\eta^2r^2\max(q(r\zeta)f, qg, q\sqrt{fg})^2$

Thus, $\alpha\frac{\epsilon^2}{32b_{prob}^2} \geq \frac{32\cdot16}{(0.01)^2}\frac{\eta^2r^2(11\log n + \log\vartheta)}{(r\zeta)^2}\max(q(r\zeta)f, qg, q\sqrt{fg})\frac{(0.01(r\zeta))^2}{32\cdot16\eta^2r^2\max(q(r\zeta)f, qg, q\sqrt{fg})^2} \geq 11\log n + \log\vartheta$. Thus, $p_1 \leq \frac{1}{\vartheta}2n^{-10}$.

2) Proof of Lemma A.1: The first claim is obvious. The next two claims follow using the following lemma:

Lemma B.1 ([2], Lemma 2.10)**.** *Suppose that $\boldsymbol{P}$, $\hat{\boldsymbol{P}}$ and $\boldsymbol{Q}$ are three basis matrices. Also, $\boldsymbol{P}$ and $\hat{\boldsymbol{P}}$ are of the same size, $\boldsymbol{Q}'\boldsymbol{P} = \boldsymbol{0}$ and $\|(\boldsymbol{I} - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{P}\| = \zeta_*$. Then,*

1) $\|(\boldsymbol{I} - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{P}\boldsymbol{P}'\| = \|(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}')\hat{\boldsymbol{P}}\hat{\boldsymbol{P}}'\| = \|(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}')\hat{\boldsymbol{P}}\| = \|(\boldsymbol{I} - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{P}\| = \zeta_*$
2) $\|\boldsymbol{P}\boldsymbol{P}' - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}'\| \leq 2\|(\boldsymbol{I} - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{P}\| = 2\zeta_*$
3) $\|\hat{\boldsymbol{P}}'\boldsymbol{Q}\| \leq \zeta_*$
4) $\sqrt{1-\zeta_*^2} \leq \sigma_i\left((\boldsymbol{I} - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{Q}\right) \leq 1$

Use item 4 of Lemma B.1 and the fact that $\boldsymbol{G}_{\rm det}{}'\boldsymbol{G}_{\rm cur} = \boldsymbol{0}$ and $\boldsymbol{G}_{\rm det}{}'\boldsymbol{G}_{\rm undet} = \boldsymbol{0}$ to get the second claim.

For the third claim, notice that $\boldsymbol{E}_{\rm cur}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet} = \boldsymbol{R}_{\rm cur}^{-1}\boldsymbol{G}_{\rm cur}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet} = \boldsymbol{R}_{\rm cur}^{-1}\boldsymbol{G}_{\rm cur}{}'\hat{\boldsymbol{G}}_{\rm det}\hat{\boldsymbol{G}}_{\rm det}{}'\boldsymbol{G}_{\rm undet}$. since $\boldsymbol{\Psi}^2 = \boldsymbol{\Psi}$ and $\boldsymbol{G}_{\rm cur}{}'\boldsymbol{G}_{\rm undet} = 0$. Using the second claim, $\|\boldsymbol{R}_{\rm cur}^{-1}\| \le \frac{1}{\sigma_{\min}(\boldsymbol{R}_{\rm cur})} \le \frac{1}{1-(r\zeta)^2}$. Use item 3 of Lemma B.1 and the facts that $\boldsymbol{G}_{\rm cur}{}'\boldsymbol{G}_{\rm det} = 0$ and $\boldsymbol{G}_{\rm undet}{}'\boldsymbol{G}_{\rm det} = 0$ to bound $\|\boldsymbol{G}_{\rm cur}{}'\hat{\boldsymbol{G}}_{\rm det}\|$ and $\|\hat{\boldsymbol{G}}_{\rm det}{}'\boldsymbol{G}_{\rm undet}\|$ respectively.

The fourth claim just uses the definitions. The fifth claim uses the previous claims and the assumptions on $\boldsymbol{M}_t$ from Model 1.2. The sixth claim follows using Weyl's inequality.

The second last claim: We show how to bound $\boldsymbol{a}_{t,\rm rest}$: $\|a_{t,\rm rest}\|^2 = \|\boldsymbol{a}_{t,\rm cur}\|^2 + \|\boldsymbol{a}_{t,\rm undet}\|^2 \le \sum_{j\in\mathcal{G}_{\rm cur}} \eta\lambda_j + \sum_{j\in\mathcal{G}_{\rm undet}} \eta\lambda_j \le r\eta\lambda_{\rm cur}^+$ (since $\lambda_j \le \lambda_{\rm cur}^+$ for all the j's being summed over). The other bounds follow similarly.

Last claim:

$$
\begin{aligned}
&\sigma_{\min}(\boldsymbol{E}_{\rm cur,\perp}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet})^2 \\
&= \lambda_{\min}(\boldsymbol{G}_{\rm undet}{}'\boldsymbol{\Psi}\boldsymbol{E}_{\rm cur,\perp}\boldsymbol{E}_{\rm cur,\perp}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}) \\
&= \lambda_{\min}(\boldsymbol{G}_{\rm undet}{}'\boldsymbol{\Psi}(\boldsymbol{I} - \boldsymbol{E}_{\rm cur}\boldsymbol{E}_{\rm cur}{}')\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}) \\
&\ge \lambda_{\min}(\boldsymbol{G}_{\rm undet}{}'\boldsymbol{\Psi}\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}) - \\
&\lambda_{\max}(\boldsymbol{G}_{\rm undet}{}'\boldsymbol{\Psi}\boldsymbol{E}_{\rm cur}\boldsymbol{E}_{\rm cur}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}) \\
&= \sigma_{\min}(\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet})^2 - \|\boldsymbol{E}_{\rm cur}{}'\boldsymbol{\Psi}\boldsymbol{G}_{\rm undet}\| \\
&\ge 1 - (r\zeta)^2 - \frac{(r\zeta)^2}{\sqrt{1-(r\zeta)^2}}.
\end{aligned}
$$

The last inequality follows using the second and the third claim.

3) Proof of Corollary 3.8: Corollary 3.8 follows in exactly the same fashion as Theorem 3.5 with the following lemma being used to replace Lemma 6.6.

Lemma B.2. *(1) Assume that $\boldsymbol{y}_t = \boldsymbol{\ell}_t + \tilde{\boldsymbol{w}}_t$, where $\tilde{\boldsymbol{w}}_t = \boldsymbol{w}_t + \nu_t = \boldsymbol{M}_t\boldsymbol{\ell}_t + \nu_t$ with $\boldsymbol{\ell}_t$ satisfying Model 1.1, $\boldsymbol{M}_t$ satisfying Model 1.2, and ν_t independent of $\boldsymbol{\ell}_t$ and satisfying $\|\mathbb{E}[\nu_t\nu_t{}']\| \le b_\nu r_{\rm cur}\zeta\lambda^-$ and $\|\nu_t\|^2 \le b2_\nu r\lambda^-$*

(2) Assume that we are given $\hat{\boldsymbol{G}}_{\rm det}$ that was computed using (some or all) $\boldsymbol{y}_t$'s for $t \le t_*$ and that satisfies $\zeta_{\rm det} \le r\zeta$.

Define $g := \lambda_{\rm cur}^+/\lambda_{\rm cur}^-$, $\chi := \lambda_{\rm undet}^+/\lambda_{\rm cur}^-$ and $\epsilon := 0.01 r_{\rm cur}\zeta\lambda_{\rm cur}^-$. Let $\hat{\boldsymbol{D}}$ be as defined in (11) *and let $\boldsymbol{D}$, $\boldsymbol{A}$, $\boldsymbol{A}_\perp$ be as defined in* (12) *and below it. Also, let $\boldsymbol{H} := \hat{\boldsymbol{D}} - \boldsymbol{D}$. Then,*

1) Let $p_1 := 2n\exp\left(-\alpha\frac{\epsilon^2}{32(2qr\eta((r\zeta)\lambda^+ + \lambda^+_{\rm cur}) + \sqrt{r\eta((r\zeta)\lambda^+ + \lambda^+_{\rm cur})rb2_\nu\lambda^-})^2}\right) \leq 2n\exp\left(-\alpha\frac{\epsilon^2}{32(2qr\eta((r\zeta)f+g)+2\sqrt{b2_\nu(r\zeta)f}+2\sqrt{b2_\nu g})\lambda^-_{\rm cur})^2}\right)$. *Conditioned on* $\{\zeta_{\rm det} \leq r\zeta\}$*, with probability at least* $1-p_1$*,*

$$\|\frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\tilde{\boldsymbol{w}}_t{}'\| \leq q((r\zeta)\lambda^+ + \lambda^+_{\rm cur})\sqrt{\frac{\beta}{\alpha}} + 2\epsilon \leq [q(r\zeta)f\sqrt{\frac{\beta}{\alpha}} + qg\sqrt{\frac{\beta}{\alpha}} + 0.02r_{\rm cur}\zeta]\lambda^-_{\rm cur}$$

2) Let $p_2 := 2n\exp(-\frac{\alpha\epsilon^2}{32(q^2\eta r\lambda^+ + b2_\nu r\lambda^-)^2}) \leq 2n\exp(-\frac{\alpha\epsilon^2}{32((q^2\eta rf + b2_\nu r)\lambda^-_{\rm cur})^2})$. *Conditioned on* $\{\zeta_{\rm det} \leq r\zeta\}$*, with probability at least* $1-p_2$*,*

$$\|\frac{1}{\alpha}\sum_t \tilde{\boldsymbol{w}}_t\tilde{\boldsymbol{w}}_t{}'\| \leq \frac{\beta}{\alpha}q^2\lambda^+ + b_\nu r_{\rm cur}\zeta\lambda^- + 3\epsilon \leq [\frac{\beta}{\alpha}q^2 f + b_\nu r_{\rm cur}\zeta + 0.03r_{\rm cur}\zeta]\lambda^-_{\rm cur}$$

In the lower bound on α, in the $\max(.)$ term we will also have $\max(., \sqrt{b2_\nu(r\zeta)f}, \sqrt{b2_\nu g}, b2_\nu)$. If $b2_\nu \leq q$, all these will be smaller than terms already in the $\max(.)$. And hence the bound on α will not get affected.

The bound on $\|\boldsymbol{H}\|$ changes as follows. If the bounds on ζ hold and if $\beta \leq \left(\frac{(1-(1+b_\nu)r_{\rm cur}\zeta-\chi)}{2}\right)^2 \min\left(\frac{(r_{\rm cur}\zeta)^2}{4.1q^2g^2}, \frac{(r_{\rm cur}\zeta)}{q^2f}\right)\alpha$, then with probability given earlier,

$$\begin{aligned}\|\boldsymbol{H}\| &\leq [2.02qg\sqrt{\frac{\beta}{\alpha}} + \frac{\beta}{\alpha}q^2 f + b_\nu r_{\rm cur}\zeta + 0.05r_{\rm cur}\zeta + 0.03\zeta]\lambda^-_{\rm cur} \\ &\leq [0.75(1-(1+b_\nu)r\zeta-\chi)r_{\rm cur}\zeta + (b_\nu+0.08)r_{\rm cur}\zeta]\lambda^-_{\rm cur} \leq (b_\nu+0.83)r_{\rm cur}\zeta\lambda^-_{\rm cur} \quad (17)\end{aligned}$$

Thus, $\zeta_{\rm cur} \leq \frac{\|\boldsymbol{H}\|}{\lambda_{\min}(\boldsymbol{A})-\lambda_{\max}(\boldsymbol{A}_\perp)-\|\boldsymbol{H}\|} \leq 0.75r_{\rm cur}\zeta + \frac{(b_\nu+0.08)r_{\rm cur}\zeta}{(1-(1+b_\nu)r_{\rm cur}\zeta-\chi)} \leq r_{\rm cur}\zeta$ if $\beta \leq \left(\frac{(1-(1+b_\nu)r_{\rm cur}\zeta-\chi)}{2}\right)^2 \min\left(\frac{(r_{\rm cur}\zeta)^2}{4.1q^2g^2}, \frac{(r_{\rm cur}\zeta)}{q^2f}\right)\alpha$, and $\chi \leq 1 - \frac{(b_\nu+0.08)}{0.25} - (1+b_\nu)r_{\rm cur}\zeta$.

Notice from above that the most stringent requirement is the bound on χ. It shows that b_ν cannot be much more than $0.25-0.08$ and, in fact, to allow χ to be large, b_ν should be much smaller than $0.25-0.08$. To keep things simple, we assume $b_\nu = 0.01$. Also, we let $\boldsymbol{b}2_\nu = q$ since that works as well as any other small value. It does not change the lower bound on α.

Thus, with our assumption that ν_t independent of $\boldsymbol{\ell}_t$ and satisfying $\|\mathbb{E}[\nu_t\nu_t{}']\| \leq 0.01r_{\rm cur}\zeta\lambda^-$ and $\|\nu_t\|^2 \leq rq\lambda^-$, the result follows in a fashion exactly analogous to Theorem 3.5.

4) Proof of Theorem 7.1: Theorem 7.1 follows in exactly the same fashion as Theorem 3.5 with the following changes to Lemma 6.6.

Instead of $\|\boldsymbol{M}_{1,t}\boldsymbol{P}\| \leq q$ assume the following generalized version: $\|\boldsymbol{M}_{1,t}\boldsymbol{G}_{\rm det}\| \leq q_0$ and $\|\boldsymbol{M}_{1,t}\boldsymbol{G}_{\rm rest}\| \leq q_1$. In applications where we use this generalized form, q_0 will be much smaller than q_1. Then the first two items of Lemma 6.6 change as follows.

1) Let $p_1 := 2n\exp\left(-\alpha\frac{\epsilon^2}{32(2r\eta(q_0(r\zeta)f+q_1g)\lambda_{\rm cur}^-+(2\sqrt{(r\zeta)f}+2\sqrt{g})\lambda_{\rm cur}^-)^2}\right)$. Conditioned on $\{\zeta_{\rm det} \leq r\zeta\}$, with probability at least $1-p_1$,

$$\|\frac{1}{\alpha}\sum_t \boldsymbol{\Psi}\boldsymbol{\ell}_t\boldsymbol{w}_t{}'\| \leq (q_0(r\zeta)\lambda^+ + q_1\lambda_{\rm cur}^+)\sqrt{\frac{\beta}{\alpha}} + 2\epsilon \leq [q_0(r\zeta)f\sqrt{\frac{\beta}{\alpha}} + q_1g\sqrt{\frac{\beta}{\alpha}} + 0.02r_{\rm cur}\zeta]\lambda_{\rm cur}^-$$

2) Let $p_2 := 2n\exp(-\frac{\alpha\epsilon^2}{32(q_0^2\eta r\lambda^+ + q_1^2\eta r\lambda_{\rm cur}^+)^2}) \leq 2n\exp(-\frac{\alpha\epsilon^2}{32((\eta r q_0^2 f+\eta r q_1^2 g)\lambda_{\rm cur}^-)^2})$. Conditioned on $\{\zeta_{\rm det} \leq r\zeta\}$, with probability at least $1-p_2$,

$$\|\frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t{}'\| \leq \frac{\beta}{\alpha}q_0^2\lambda^+ + \frac{\beta}{\alpha}q_1^2\lambda_{\rm cur}^+ + 3\epsilon \leq [\frac{\beta}{\alpha}q_0^2 f + \frac{\beta}{\alpha}q_1^2 g + 0.03r_{\rm cur}\zeta]\lambda_{\rm cur}^-$$

Recall that $p_3 := 2n\exp(-\frac{\alpha\epsilon^2}{32b_{prob}^2})$ with $b_{prob} := \eta r((r\zeta)^2\lambda^+ + \lambda_{\rm cur}^+) \leq 2\eta r((r\zeta)^2 f + g)\lambda_{\rm cur}^-$.

Thus, in the lower bound on α, we need

$$\max(g, (r\zeta)^2 f,\ q_0^2 f, q_1^2 g,\ q_0(r\zeta)f, q_1 g) = \max(g, q_1 g, q_1^2 g, q_0^2 f, q_0(r\zeta)f, (r\zeta)^2 f)$$

We also need $\beta \leq \left(\frac{(1-r_{\rm cur}\zeta-\chi)}{2}\right)^2 \min\left(\frac{(r_{\rm cur}\zeta)^2}{4.1(q_1 g+q_0(r\zeta)f)^2}, \frac{(r_{\rm cur}\zeta)}{q_0^2 f+q_1^2 g}\right)\alpha$, and $\chi \leq 1 - \frac{0.08}{0.25} - (1+b_\nu)r_{\rm cur}\zeta$.

Apply this lemma with $q_0 \equiv q_0$ and $q_1 \equiv q_{\rm new}$, $r \equiv r_0$ to get the theorem.