Abstract—We study Principal Component Analysis (PCA) in a setting where a part of the corrupting noise is data-dependent and, hence, the noise and the true data are correlated. Under a bounded-ness assumption on both the true data and noise, and a few assumptions on the data-noise correlation, we obtain a sample complexity bound for the most common PCA solution, singular value decomposition (SVD). This bound, which is within a logarithmic factor of the best achievable, significantly improves upon our bound from recent work (NIPS 2016) where we first studied this “correlated-PCA” problem.

I. INTRODUCTION

We study Principal Component Analysis (PCA) in a setting where a part of the corrupting noise is data-dependent in such a way that the corrupting noise and the true data are correlated. We first studied this problem in [1] where we called it “correlated-PCA”. Under a bounded-ness assumption on both the true data and noise, and a few assumptions on the data-noise correlation (Assumption 2 given later), we obtain nearly optimal sample complexity guarantees for the most commonly used PCA solution, singular value decomposition (SVD) on the observed data matrix. Henceforth we refer to this strategy as simple SVD or just SVD. Observe that SVD is equivalent to eigenvalue decomposition (EVD) of the sample covariance matrix, and hence some authors would refer to this as simple EVD or just EVD. The bounded-ness assumption is, in fact, a more practical one than Gaussianity, since data acquisition devices typically have finite power.

Problem Setting. For $t = 1, 2, \ldots$, we are given $n$-length data vectors, $y_t$, that satisfy

$$y_t = \ell_t + w_t + v_t,$$

$P$ is an $n \times r$ matrix with orthonormal columns and $r \ll n$; $\ell_t$ is the true data vector that lies in a low ($r$) dimensional subspace of $\mathbb{R}^n$, range($P$); $\alpha_t$ is its projection into this subspace; $w_t$ is the data-dependent (correlated) noise component; and $v_t$ is the uncorrelated noise component, i.e., it satisfies $\mathbb{E}[\ell_t, v_t^\top] = 0$. The matrices $M_t$ are unknown and such that $\mathbb{E}[\ell_t w_t^\top] \neq 0$ (holds if $M_t \neq 0$, or $\mathbb{E}[M_t] \neq 0$ in case $M_t$ is random). The goal is to estimate range($P$). Since the matrices $M_t$ are time-varying, observe that, in general, the $w_t$’s do not lie in a lower dimensional subspace of $\mathbb{R}^n$.

Examples. A motivating example for this study is the problem of PCA in the presence of additive sparse outliers (“robust PCA” [2]) when the corrupting sparse outlier values are data-dependent. To be precise, let $T_t$ denote the outlier support at time $t$. Then, robust PCA with data-dependent outlier values involves PCA from observed data $y_t := \ell_t + I_{T_t} s_t + v_t$ where $s_t = M_{s,t} \ell_t$ with $M_{s,t}$ being a $|T_t| \times n$ matrix. Here $I_{T_t} M_{s,t}$ is the data-dependency matrix. This model is often a valid one for video analytics applications, where $\ell_t$ is the background layer of image frame $t$, $T_t$ is the foreground support of frame $t$, and $s_t$ is the difference between foreground and background intensities on $T_t$. Another example where data-dependent noise occurs is the subspace update step of the Recursive Projected Compressive Sensing (ReProCS) solution to dynamic robust PCA [3], [4]. Dynamic robust PCA refers to the robust PCA problem when the true data lies in a slowly changing subspace and the goal is to track this changing subspace over time. The assumptions made in the current work are inspired by this example. Finally, as explained in [5], data-dependent noise also often occurs in molecular biology applications when the noise affects the measurement levels through the very same process as the interesting signal.

Contributions. In recent work [1], we studied the correlated-PCA problem described above. Our new result given here (Theorem [3]) addresses three key limitations of [1]. (1) It gives a significantly improved sample complexity bound and one that is within a logarithmic factor of the best achievable. (2) We generalize the observed data model to also include an uncorrelated noise term. This is a more practically valid noise model since the noise/corruption is usually not fully data-dependent. (3) We provide a provably correct method for automatic subspace dimension estimation that does not use knowledge of any model parameter (see Corollary [6]).

To our best knowledge, most existing finite sample guarantees for the simple SVD solution to PCA, other than [1], assume that the true data and the corrupting noise are independent, or, at least uncorrelated, e.g., see [6], [7] and references therein, and also see the summary of existing batch PCA guarantees given in [8] Section 1. This is valid in practice often, but not always. There are, of course, a large number of works on robust PCA that assume nothing about the dependence between the outlier magnitudes and the true data, e.g., [2], [9], [10], [11], [12]. In particular, these allow the outlier values to be dependent on (correlated with) the true data. However, these works focus on large magnitude sparse outliers and hence (i) need more expensive solutions than simple SVD; and (ii) need the columns of $P$ to be dense (not sparse). On the other hand, the simple SVD solution is faster and does not require denseness of columns of $P$; however it, of course, only works for small magnitude outliers. This point is demonstrated experimentally in Table [1]. We should mention
that there are very recent works on fast robust PCA (13), (14) that have the same order of computational complexity as simple SVD. However, these still require denseness of columns of $P$, and will be slower than SVD in practice (their initialization step itself involves an SVD).

II. ASSUMPTIONS AND MAIN RESULT

We assume the following about the true data $\ell_t$ and the data-dependency matrix $M_t$.

**Assumption 1.** The $\ell_t$’s satisfy $\ell_t = P a_t$ with $a_t$’s being zero mean, mutually independent, and bounded r.v.’s, with diagonal covariance matrix, $\Lambda$.

Define $\lambda^− := \lambda_{\min}(\Lambda)$, $\lambda^+ := \lambda_{\max}(\Lambda)$ and $f := \lambda^+/\lambda^−$. Since the $a_t$’s are bounded, there exists a finite constant, $\eta$, such that, $\max_{j=1,2,\ldots, r} \max_t (\frac{a_{tj}}{\lambda^−}) \leq \eta$. Observe that $\eta$ bounds the ratio of the square of the maximum magnitude of $a_t$ over $t$ in any direction to its variance in that direction. For most bounded distributions, it is a little more than one, e.g., if the $a_t$’s are iid uniform, then $\eta = 3$.

**Assumption 2.** The data-dependency matrices $M_t$ can be split as $M_t = M_{2,t}M_{1,t}$ with $M_{2,t}$, $M_{1,t}$ satisfying the following. For a $q < 1$, a $b_0 < 1$, and a positive integer $\alpha$,

$$0 < \|M_{1,t}P\|_2 \leq q < 1, \quad \|M_{2,t}\|_2 \leq 1, \quad and$$

$$\left\|\sum_{t=1}^{\alpha} M_{2,t}A_tM_{2,t}\right\|_2 \leq b_0 \max_{t\in[1,\alpha]} \|A_t\|.$$

for any $\alpha$-length sequence of positive semi-definite Hermitian matrices, $A_t$.

Assumption 1 just states mutual independence and boundedness of the $\ell_t$’s. The first part of Assumption 2 bounds the instantaneous noise-to-signal ratio of the correlated (data-dependent) component of the noise, $\nu_t$; using it, $\|\nu_t\|_2 \leq q\|a_t\|_2$ and $\|E[\nu_t\nu_t^\prime]\|_2 \leq q^2\|E[\ell_t\ell_t^\prime]\|_2$. The second part can be understood as one way to reduce the time-averaged power of $\nu_t$. Observe that $\|E[\nu_t\nu_t^\prime]\|_2 \leq q^2\lambda^+$, whereas $\|\sum_{t=1}^{\alpha} E[\nu_t\nu_t^\prime]\|_2 \leq b_0q^2\lambda^+$. Thus, when $b_0$ is small, the expected value of the time-averaged correlated noise power is much smaller than the instantaneous one. This is useful because it helps to reduce the time-averaged signal-noise correlation: using Cauchy-Schwarz, it is not hard to see that $\|\sum_{t=1}^{\alpha} E[\nu_t\nu_t^\prime]\|_2 \leq \sqrt{b_0}q\lambda^+$. One example where Assumption 2 holds is when $\nu_t$ is sparse with time-varying support sets, denoted $T_t$. In this case, $M_{2,t} = I_{T_t}$. If all the sets $T_t$ are mutually disjoint, the matrix on the LHS of (2) is either block-diagonal, or is permutation-similar to a block-diagonal matrix, with blocks $A_t$. Thus, in this case, (2) holds with $b_0 = 1/\alpha$. This example can be generalized to also allow the support sets to change every so often, and to not even be mutually disjoint; see (14), (11).

With the above assumptions, we study Algorithm 1. We bound the subspace error recovery,

$$\text{SE}(P, P) := \|(I - PP')P\|_2,$$

of its output$^1$ For simplicity, we first study this simple algorithm that assumes $r$ known. We give corollaries for the $r$ unknown case later (see Corollary 3 and 4).

**Algorithm 1** Simple SVD (or EVD)

Let $\hat{P}$ be the matrix of top $r$ singular vectors of $[y_1, y_2, \ldots, y_n]$. Equivalently, $P$ is the matrix of top $r$ eigenvectors of $\frac{1}{n} \sum_{t=1}^n y_t y_t^\prime$.

**Theorem 3.** Assume that $v_t$ satisfies $\|E[v_t v_t^\prime]\|_2 \leq \lambda^+_v$ and $\|v_t\|_2 \leq \eta r \lambda^+_v$. For an $\varepsilon_{SE} < 1$, define $d := \max(1, \frac{\eta r \log 9 + 10 \log n}{\varepsilon_{SE}})$ and

$$\alpha_0 := C\eta d \left(\log n\right) \max(\eta r^2 q^2, w_0^2 \eta \lambda^+), \max(\eta r, r)F \frac{\lambda^+}{\eta^2 \lambda^+} \right).$$

For an $\alpha \geq \alpha_0$, let $\hat{P}$ be as defined in Algorithm 1. Assume that Assumptions 1 and 2 hold with this $\alpha$.

If $3.3\sqrt{\log f} + 1.3\frac{9^2 + 10 \log n}{\varepsilon_{SE}} = 0.49\varepsilon_{SE}$, then, with probability (w.p.) at least $1 - 6n^{-10}$,

$$\text{SE}(\hat{P}, P) \leq \varepsilon_{SE}$$

A corollary of this result for the $v_t = 0$ case is as follows.

**Corollary 4.** Assume that $v_t = 0$. For an $\varepsilon_{SE} < 1$, define $d := \max(1, \frac{\eta r \log 9 + 10 \log n}{\varepsilon_{SE}})$ and

$$\alpha_0 := C\eta d \left(\log n\right) r f^2 q^2 \frac{\eta^2}{\varepsilon_{SE}}.$$

For an $\alpha \geq \alpha_0$, let $\hat{P}$ be as defined in Algorithm 1. Assume that Assumptions 1 and 2 hold with this $\alpha$. If $3.3\sqrt{\log f} = 0.49\varepsilon_{SE}$, then, w.p. at least $1 - 6n^{-10}$,

$$\text{SE}(\hat{P}, P) \leq \varepsilon_{SE}$$

III. DISCUSSION

Effect of correlated noise. To compare the effects of correlated and uncorrelated noises, consider corollaries of the above result when only one type of noise is present. For a head-to-head comparison, equate the time-averaged correlated noise power bound and the uncorrelated noise power bound, and also equate the bounds on $\|\nu_t\|_2$ and $\|\nu_t\|_2$. Thus, suppose that $\lambda^+ = b_0q^2\lambda^+$ and $\eta^2 \lambda^+ = \eta^2 q^2 \lambda^+$. Then, in the only correlated-noise case ($v_t = 0$), we need $3.3\sqrt{\log f} \leq 0.49\varepsilon_{SE}$, and $\alpha \geq C\eta d \left(\log n\right) r f^2 q^2 \frac{\eta\lambda^+}{\varepsilon_{SE}}$. In the $w_t = 0$ case, we need $1.3b_0q^2 f \leq 0.49\varepsilon_{SE}$ and $\alpha \geq C\eta d \frac{(\log n)}{\varepsilon_{SE}} \left(\log (n) \sqrt{f}\right)^2$. Thus the $\alpha$ required in both cases is the same. However, the upper bound on $f$ needed in the correlated noise case is stronger. For example, when $\varepsilon_{SE} = q/2$, in the only correlated noise case, one needs $f < 0.074/\sqrt{\gamma}$, while, in the only uncorrelated noise case, one needs $f < 0.188/(b_0q)$. Since $b_0 < 1$ and $q < 1$, the latter bound is looser.

$^1\varepsilon_{SE}(P, P)$ quantifies the principal angle between the column spans of $\hat{P}$ and $P$ (this is a valid definition when $\hat{P}$ and $P$ have orthonormal columns).
The reason that the correlated noise case is harder is because the bound on \( \text{SE}(\mathbf{P}, \mathbf{P}) \) is governed by the ratio between the spectral norm of the perturbation matrix, \( \mathbf{H} := \frac{1}{\alpha} \sum \ell_i \ell_i' - \frac{1}{\alpha} \sum \ell_i \ell_i' \), and the minimum eigenvalue along the principal subspace, \( \lambda^- \). This follows using the Davis-Kahan sin \( \theta \) theorem [15]. In the correlated noise case, the dominant terms in \( \mathbf{H} \) are the signal-noise correlation terms, \( \frac{1}{\alpha} \sum \ell_i \ell_i' \) and its transpose. Since the noise is smaller than signal (\( q < 1 \)), these terms are larger than the noise power terms \( \frac{1}{\alpha} \sum \ell_i \ell_i' \) or \( \frac{1}{\alpha} \sum \ell_i \ell_i' \). On the other hand, in the only uncorrelated noise case, with high probability (whp), the only non-negligible term is \( \frac{1}{\alpha} \sum \ell_i \ell_i' \).

We should mention here that there is work in linear algebra on studying the effect of multiplicative perturbations of Hermitian matrices on their principal subspaces, e.g., see [16] and references therein. This line of work provides a tighter bound than Davis-Kahan for the subspace error between principal subspaces of a Hermitian matrix \( \mathbf{A} \) and of its perturbed version \( D'AD \) for a non-singular matrix \( D \). However, such results are not applicable for our problem since \( M \) is time-varying.

**Comparison with [17]**. The result of [1] assumed that \( v_t = 0 \). Thus, to compare with it, let \( v_t = 0 \) so that \( \beta^+ = 0 \) and \( \gamma = 0 \) in Theorem 3. First consider the case where \( d = 1 \). In this case, the sample complexity, \( \alpha \), is lower bounded by \( Cf^2 r(\log n) \frac{g}{\varepsilon^2} \). Thus, to get the subspace error to below \( \varepsilon_{SE} = q/4 \), we need \( \alpha \geq 16Cr(\log n)^2 \) samples. This is much better than our earlier sample complexity bound [11] of \( Cr^2(\log n)^2 \frac{f}{\varepsilon_{SE}} \) which implies that we need \( \alpha \geq 16Cr^2(\log n)^2 \) to achieve the above subspace error level. This inverse dependence on noise level, \( q \), of our earlier bound is counter-intuitive; we should not need more samples when \( q \) is smaller. Moreover, our current bound replaces \( r^2(\log n) \) by \( r(\log n) \). Both results also need \( \sqrt{n} q f \leq 0.14 \varepsilon_{SE} \).

We get the first improvement by bounding the \( r \)-th eigenvalue of \( \sum \ell_i \ell_i' = \mathbf{P}' \sum_a a_i a_i' \mathbf{P} \) by using a result of Vershynin [17] Theorem 5.39] to bound the minimum eigenvalue of \( \sum_a a_i a_i' \). In [11], we had used matrix Hoeffding for doing this. We get the second improvement by using matrix Bernstein to replace matrix Hoeffding to get high probability bounds on time-averaged signal-noise correlation and noise power.

If \( \varepsilon_{SE} \) is larger than \( q \), and \( n \) and \( r \) are small enough, \( d < 1 \). In this case, just \( \alpha \geq Cf^2 r \log 9 + 10 \log n \) suffices.

**Matching lower bound**. The minimum number of samples required to estimate the subspace range(\( \mathbf{P} \)) is \( r \). Thus, if \( f = O(1) \), up to constants, a sample complexity of \( \alpha \geq Cf^2 r(\log n) \) is only \( (\log n) \) times larger than the best achievable. We need to have some dependence on \( n \) because, in general, the \( w_i \)'s lie in \( \mathbb{R}^n \) (and not in a lower dimensional subspace of it).

**Logarithmic dependence on signal dimension** \( n \). The reason that we get a logarithmic dependence on \( n \) is because of the bounded-ness assumption on both \( \ell_i \) and \( w_i \). If we were removed, our guarantees would require \( O(n) \) samples. This sample complexity would then be similar to that of existing results for the uncorrelated (or independent) noise cases, e.g., [6] (finite sample guarantee for \( r = 1 \) dimensional PCA), [7], or [8] (finite sample guarantee for memory-limited streaming PCA), all of which assume Gaussian noise. Since the latter is a memory-limited streaming algorithm, it, in fact, needs \( O(n \log n) \) samples. We should point out that there is a large amount of other literature on both streaming PCA and online PCA which we do not cite or discuss here since those are not the problems studied in this work.

**Automatically estimating \( r \)**. There are two easy and commonly used ways to automatically estimate \( r \). As the next two corollaries show, both will return the correct estimate \( r \) with the probability stated in Theorem 3. The first is as done in [1]. This computes \( \hat{r} \) as the smallest index \( j \) for which

\[
\lambda_j \left( \sum_{i=1}^{\alpha} y_i y_i' \right) \geq 0.5 \alpha \lambda^-
\]

and thus requires knowledge of \( \lambda^- \). We have the following corollary.

**Corollary 5**. In the setting of Theorem 3 if \( \varepsilon_{SE} < 1/2, \) then, with probability \( \geq 1 - 10n^{-10} \),

1. \( \lambda_j \left( \sum_{i=1}^{\alpha} y_i y_i' \right) \geq \lambda^- \left( 0.98 - \varepsilon_{SE}/2 \right) \geq 0.73 \lambda^-, \) and 2. \( \lambda_j \left( \sum_{i=1}^{\alpha} y_i y_i' \right) \leq \varepsilon_{SE}/2 \lambda^- < 0.25 \lambda^-, \) and thus, the above approach returns \( \hat{r} = r \), with this probability.

An alternate way to estimate \( r \) is as

\[
\hat{r} := \arg \max_j \left[ \lambda_j \left( \sum_{i=1}^{\alpha} y_i y_i' \right) - \lambda_{j+1} \left( \sum_{i=1}^{\alpha} y_i y_i' \right) \right].
\]

This does not require knowledge of \( \lambda^- \). But, it is more expensive (needs all eigenvalues), and, as we see below, it needs one extra assumption.

**Corollary 6**. In the setting of Theorem 3 let \( \varepsilon_{SE} < 1/4 \). Assume also that \( \lambda_j(A) - \lambda_{j+1}(A) \leq 0.45 \lambda^- \) for all \( j = 1, 2, \ldots, r \). Then, with probability \( \geq 1 - 10n^{-10} \)

1. for \( j < r \), \( \lambda_j \left( \sum_{i=1}^{\alpha} y_i y_i' \right) - \lambda_{j+1} \left( \sum_{i=1}^{\alpha} y_i y_i' \right) \leq 0.45 \lambda^- + 2(\varepsilon_{SE}/2) \lambda^- < 0.73 \lambda^- \), and 2. for \( j > r \), \( \lambda_j \left( \sum_{i=1}^{\alpha} y_i y_i' \right) - \lambda_{j+1} \left( \sum_{i=1}^{\alpha} y_i y_i' \right) \leq 2(\varepsilon_{SE}/2) \lambda^- < 0.25 \lambda^- \), and 3. for \( j = r \), \( \lambda_j \left( \sum_{i=1}^{\alpha} y_i y_i' \right) - \lambda_{j+1} \left( \sum_{i=1}^{\alpha} y_i y_i' \right) \geq \lambda^- \left( 0.98 - 2(\varepsilon_{SE}/2) \right) > 0.73 \lambda^- \), and thus, the above approach returns \( \hat{r} = r \), with this probability.

**IV. PROOF OF THEOREM 3**

To start with a simple proof, we prove Corollary 4 (\( v_t = 0 \) case) first. In a few places in this proof, we have missed the subscript 2, but everywhere the norm used is the spectral norm (induced l2-norm) only.

**Proof of Corollary 4**. Using the Davis-Kahan sin \( \theta \) theorem [15] followed by Weyl’s inequality (see [1]),

\[
\text{SE}(\mathbf{P}, \mathbf{P}) \leq \frac{2 \| \frac{1}{\alpha} \sum \ell_i w_i' \|_2 + \| \frac{1}{\alpha} \sum \ell_i w_i' \|_2}{\lambda_1 \left( \frac{1}{\alpha} \sum \ell_i \ell_i' \right) - \left( \frac{1}{\alpha} \sum \ell_i w_i' \|_2 + \| \frac{1}{\alpha} \sum \ell_i w_i' \|_2 \right)}
\]
if the denominator is positive. The two numerator terms can be bounded using the matrix Bernstein inequality [18]. Observe that 
\[ \lambda \left( \sum_{t} \ell_t \ell'_t \right) = \lambda_{\text{min}} \left( \sum_{t} a_t a'_t \right) \]. We can bound \( \lambda_{\text{min}} \left( \sum_{t} a_t a'_t \right) \) using Theorem 5.39 of [17]. Since the \( a_t \)'s are bounded, they are sub-Gaussian with sub-Gaussian norm bounded by \( \sqrt{\eta \lambda^+} \). Because the \( a_t \)'s are \( r \)-length vectors, the Vershynin theorem gives a much higher concentration probability than if we use matrix Bernstein for this term.

Matrix Bernstein for rectangular matrices, Theorem 1.6 of [18] says the following. For a finite sequence of \( d_1 \times d_2 \) zero mean independent matrices \( Z_t \) with 
\[
\|Z_t\|_2 \leq R, \max(\|\sum_k E[Z_k Z_k']\|_2, \|\sum_k E[Z_k Z_k']\|_2) \leq \sigma^2,\]
we have \( \mathbb{P}(\|\sum_k Z_k\|_2 \geq s) \leq (d_1 + d_2) \exp \left( -\frac{s^2}{2 \sigma^2 + ks^4} \right) \).

Let \( Z_t := E[Z_t] \) with \( s = \epsilon \alpha \). To get the values of \( R \) and \( \sigma^2 \) in a simple fashion, we use the facts that (i) if \( \|Z_t\|_2 \leq R \), then \( \|Z_t\|_2 \leq 2 \epsilon R \); and (ii) \( \sum_t E[Z_t Z_t'] \preceq \sum_t E[Z_t Z_t'] \). Thus, we can set \( R \) to be the two times the bound on \( \|Z_t\|_2 \) and we can set \( \sigma^2 \) as the maximum of the bounds on \( \|\sum_t E[Z_t Z_t']\|_2 \) and \( \|\sum_t E[Z_t Z_t']\|_2 \).

It is easy to see that \( R = 2 \sqrt{\eta \lambda^+} \sqrt{\eta q^2 \lambda^+} = 2 \sqrt{\eta q^2 \lambda^+} \).

To get \( \sigma^2 \), observe that
\[
\left\| \sum_t E[w_t \ell_t \ell_t' \ell_t' w_t'] \right\|_2 \leq \alpha \max(\|\ell_t\|_2, \|\ell_t\|_2) \cdot \|E[w_t w_t']\| \leq \alpha \eta \lambda^+ \cdot q^2 \lambda^+ = \alpha \eta \lambda^+ \cdot q^2 \lambda^+.
\]

Repeating the above steps, we get the same bound on \( \|\sum_t E[Z_t Z_t']\|_2 \). Thus, \( \sigma^2 = \alpha \eta \lambda^+ \cdot q^2 \lambda^+ \).

Thus, we conclude that
\[
\left\| \sum_t \ell_t w_t - E[\sum_t \ell_t w_t] \right\|_2 \geq \epsilon \alpha
\]
w.p. at most \( 2n \exp \left( -\frac{\epsilon^2 \alpha^2}{2 \sigma^2 + \epsilon^2 \sigma^2} \right) \).

If \( \epsilon < q \lambda^+ \), the above probability is bounded by \( 2n \exp \left( -\frac{\epsilon^2 \sigma^2}{4 \eta \sigma^2 q^2 \lambda^+} \right) \).

Thus, with probability at least \( 1 - 2n \exp \left( -\frac{\epsilon^2 \sigma^2}{4 \eta \sigma^2 q^2 \lambda^+} \right) \), we have
\[
\left\| \sum_t \ell_t w_t \right\|_2 \leq \left\| E[\sum_t \ell_t w_t] \right\|_2 + \epsilon \leq \sqrt{b_0 q^2 \lambda^+} + \epsilon
\]
as long as \( \epsilon < q \lambda^+ \). Set \( \epsilon = \epsilon_0 \lambda^+ \), then we get: as long as \( \epsilon_0 < q \lambda^+ \), with probability at least \( 1 - 2n \exp \left( -\frac{\epsilon^2 \sigma^2}{4 \eta \sigma^2 q^2 \lambda^+} \right) \), we have
\[
\left\| \sum_t \ell_t w_t \right\|_2 \leq \sqrt{b_0 q^2 \lambda^+} + \epsilon = \sqrt{b_0 q \lambda^+} + \epsilon_0 \lambda^+
\]
Thus, the above event holds w.p. at least \( 1 - 2n^{-10} \) if
\[
\alpha \geq \alpha_0 = (11 \log n) 4 \eta q^2 \frac{q^2 f^2}{\epsilon_0^2} = 44 \eta q^2 \log n \frac{q^2 f^2}{\epsilon_0^2}
\]
and \( \epsilon_0 \leq q \lambda^+ \).

Consider the second term. Proceeding as above, we get
\[
R = 2 \eta q^2 \frac{q^2 f^2}{\epsilon_0^2} \lambda^+ + \sigma^2 = \alpha \sigma^2 = \eta q^4 (\lambda^+)^2.
\]
The probability at least \( 1 - 2n \exp \left( -\frac{\epsilon^2 \sigma^2}{4 \eta \sigma^2 q^2 \lambda^+} \right) \), \( \left\| \frac{1}{\alpha} \sum_t \ell_t w_t \right\|_2 \leq \left\| \frac{1}{\alpha} E[\sum_t \ell_t w_t] \right\|_2 + \epsilon_2 \lambda^- \leq \frac{b_0 \epsilon_2 q^2 f + c_2}{\lambda^-}
\]
Thus, the above event holds w.p. at least \( 1 - 2n^{-10} \) if
\[
\alpha \geq \alpha_2 = 44 \eta (\log n) \max \left( \frac{q^2 f^2}{\epsilon_0^2}, \frac{q^2 f}{\epsilon_2} \right)
\]
Using Theorem 5.39 of [17] applied to \( \frac{1}{n} \sum_{t} a_t a'_t \), and using the fact that the \( a_t \)'s are \( r \)-length independent sub-Gaussian vectors with sub-Gaussian norm bounded by \( \sqrt{\eta \lambda^+} \), we get the following: with probability at least \( 1 - 2n \exp \left( -\frac{\epsilon^2 \sigma^2}{4 \eta \sigma^2 q^2 \lambda^+} \right) \), \( \lambda_r \left( \frac{1}{\alpha} \sum_t \ell_t \ell' t \right) \leq \lambda_{\text{min}} \left( \frac{1}{\alpha} \sum_t a_t a'_t \right) \geq \lambda^-(1 - \epsilon_1)
\]
Thus, the above event holds w.p. at least \( 1 - 2n^{-10} \) if
\[
\alpha \geq \alpha_1 = \frac{(r \log 9 + 10 \log n) \cdot 16 q^2 f^2}{\epsilon_1} = 16 q^2 c (r \log 9 + 10 \log n) \frac{f^2}{\epsilon_1^2}
\]
Thus, we have the following result.

**Theorem 7.** For an \( \alpha \geq \max(\alpha_0, \alpha_1, \alpha_2) \), let \( \tilde{P} \) be the matrix of top \( r \) eigenvectors of \( \frac{1}{n} \sum_{t} y_t y_t' \). Assume that Assumptions 7 and 2 hold for the chosen \( \alpha \). Then w.p. \( 1 - 6n^{-10} \),

\[
\text{SE}(\tilde{P}, P) \leq \frac{2q f \sqrt{b_0} + q^2 f b_0 + 2 \epsilon_0 + \epsilon_2}{1 - \epsilon_1 - \text{num}}
\]
as long as numerator \( < 1 - \epsilon_1, \epsilon_0 < q f \). Here numerator refers to the numerator term.

Set \( \epsilon_2 = \epsilon_0 = 0.1 \sqrt{b_0} q f \) and \( \epsilon_1 = 0.02 \). Then, \( \alpha_2 < \alpha_0 = 44 \eta q^4 (\log n) \frac{100}{b_0} = 4400 \eta (\log n) \frac{1}{b_0}, \alpha_1 = 16 q^2 c (r \log 9 + 10 \log n) (25 f^2), \) and

\[
\text{SE}(\tilde{P}, P) \leq \frac{3.3 q f \sqrt{b_0}}{0.98 - 3.3 q f \sqrt{b_0}}
\]
if denominator is positive. To bound the RHS by an \( \varepsilon_{\text{SE}} < 1 \), it suffices to have
\[
3.3 q f \sqrt{b_0} \leq 0.98 \frac{\varepsilon_{\text{SE}}}{2} = 0.49 \varepsilon_{\text{SE}}
\]
This means that we need
\[
b_0 \leq \frac{0.49^2 \varepsilon_{\text{SE}}^2}{3.3^2 q^2 f^2} = 0.022 \frac{\varepsilon_{\text{SE}}^2}{q^2 f^2}
\]
With \( b_0 \) set equal to its upper bound,
\[
\alpha_0 = 4400 \eta (\log n) \frac{1}{b_0} = C \eta (\log n) \frac{q^2 f^2}{\varepsilon_{\text{SE}}^2}
\]
Recall that \( \alpha_1 = C \eta^2 (r \log 9 + 10 \log n) f^2 \). Define
\[
d = \max \left( \frac{1}{\eta} \frac{r \log 9 + 10 \log n \varepsilon_{\text{SE}}}{q^2 \log q^2} \right)
\]
Thus, if
\[ \alpha \geq C d \eta \ (\log n) \frac{q^2 f^2}{\varepsilon_{SE}} \]
and if \( b_0 = 0.022 \frac{\varepsilon_{SE}}{q^2 f^2} \), then
\[ \text{SE}(\hat{P}, P) \leq \varepsilon_{SE} \]

Now consider the general case \( \nu_t \neq 0 \). We get the final result for this case by also using the following lemma (which again follows by matrix Bernstein).

**Lemma 8.** Pick an \( \epsilon_{0, \nu} > 0 \).

1. With probability at least \( 1 - \frac{2n \exp\left(-\alpha(\epsilon_{0, \nu} \lambda^+)\right)}{2n \exp\left(-\alpha(\epsilon_{0, \nu} \lambda^-)\right)} \),
\[ \left\| \frac{1}{\alpha} \sum_t \ell_t \nu_t \right\|_2 \leq \epsilon_{0, \nu} \lambda^- \]
2. With probability at least \( 1 - \frac{2n \exp\left(-\alpha(\epsilon_{0, \nu} \lambda^-)\right)}{2n \exp\left(-\alpha(\epsilon_{0, \nu} \lambda^+)\right)} \),
\[ \left\| \frac{1}{\alpha} \sum_t \nu_t \nu_t^T \right\|_2 \leq \lambda^+ + \epsilon_{0, \nu} \lambda^- \]

V. NUMERICAL EXPERIMENTS

Our first experiment fixed \( n = 100 \) and numerically computed the sample complexity as a function of \( r \) for a few different values of \( \varepsilon_{SE} \). We generated \( \ell_t = P a_t \) where \( (a_t)_i \)'s were iid \( unif((-c_r, c_r)) \) with \( c_1 = c_2 = \ldots c_{r-1} = \sqrt{3r} \) and \( c_r = \sqrt{3} \). With this, \( \lambda^+ = f, \lambda^- = 1 \) (and so \( f \) is the condition number) and \( \eta = 3 \). The subspace basis matrix \( P \) was generated by orthormalizing an \( n \times r \) iid Gaussian matrix. We generated \( \nu_t \) as \( \nu_t = I_r M_{s,t} \ell_t \frac{q}{\|M_{s,t} P\|} \) where \( T_r \) followed Model 2.3 of [4] with \( s = 5, \rho = 1, \beta = 0.25 \alpha \). We generated each entry of \( M_{s,t} \) as the absolute value of a standard Gaussian random variable (taking the absolute value ensures that \( \|\mathbb{E}[M_{s,t}]\| > 0 \)). Thus, if we let \( M_{2,t} = I_r \) and \( M_{1,t} = M_{s,t} \frac{q}{\|M_{s,t} P\|} \), using [4] Lemma 5.2, 5.3, Assumption [2] holds with \( b_0 = 0.25 \) and \( q = q \). We set \( y_t = \ell_t + \nu_t \).

In our first experiment, we fixed \( f = 1 \) and \( q = 0.1 \). We used \( n = 100 \) and four values of \( r: r = 2, 5, 15, 25 \). For each \( r \), we varied \( \alpha \) in the range \( 5 \) to \( 300 \) (selected 100 uniformly spaced values of \( \alpha \) in this range) and computed the simple EVD estimate \( \hat{P} \). For each \( r, \alpha \) pair, we repeated the experiment 5000 times and computed \( \text{SE}(\hat{P}, P) \) each time. For an error threshold \( \varepsilon_{SE} \), for each \( r, \alpha \), we counted the percentage of times \( \text{SE}(\hat{P}, P) \leq \varepsilon_{SE} \). For a given \( r \), we computed \( \alpha_0 \) as the smallest value of \( \alpha \) for which this percentage was at least \( 1 - n^{-1.5} \) for all \( \alpha \)’s greater than this one. In Fig. [1] we plot \( \alpha_0 \) versus \( r \) for three different values of \( \varepsilon_{SE}: \varepsilon_{SE} = 0.75q, q, 10q \). As can be seen, the plots are close to linear.

Our second experiment compares the correlated and uncorrelated noise cases. We generated \( \ell_t \) and \( \nu_t \) as explained above with \( n = 1000, r = 10, s = 10, \rho = 1, b_0 = 0.25 \) and varying values of \( f \) and \( q \). For the uncorrelated noise case, we set \( y_t = \ell_t + \nu_t \) where \( \nu_t \) was generated as explained earlier to allow for a head-to-head comparison. Thus, we used \( \lambda^+ = b_0 q^2 \lambda^+ \) and \( r_v = r/b_0 = 4r \). This ensures that the bounds match. However, the actual subspace error values also depend on all the noise eigenvalues. To make the comparison fair, we used a simple approach to ensure that the trace of \( \frac{1}{\alpha} \sum_{t=1}^{\alpha} \nu_t \nu_t^T \) is approximately equal to that of \( \frac{1}{\alpha} \sum_{t=1}^{\alpha} \nu_t \nu_t^T \) for \( \alpha = 200 \), and each value of \( f \), we plot the average SE versus \( q \) for the two cases in Fig. [2] As can be seen, the correlated noise case results in much higher errors.

In our third experiment, we compare the average SE and time taken by simple SVD (implemented using the svds command in MATLAB) with two popular robust PCA solutions - PCP [2], [9] and AltProj [11]. For this comparison, we generated \( P \) in two ways. First, to use a sparse \( P \) (a case for which the identifiability assumption for robust PCA fails), we let \( P \) be the first \( r \) columns of the identity matrix. We also generated a dense \( P \) as described above to simulate case where robust PCA methods also work. We used \( n = 1000, r = 10, q = 0.01, s = 10, \rho = 1, b_0 = 0.25 \) and \( \alpha = 500 \). We show the average SE and time taken in Table [3] As can be seen, simple SVD is always faster. Also, when \( P \) is sparse, AltProj and PCP fail while simple EVD still works. When \( P \) is dense, all three methods work, but, the robust PCA methods
have much smaller errors than just SVD. In all cases, the robust PCA methods are much slower. We should reiterate that SVD works in these examples only because we used a small $q$.

| $P$ Sparse | SE$(P, P)$ | IALM | AltProg | Simple-EVD |
|------------|------------|------|--------|------------|
| run time/sec | 1.000 | 1.000 | 0.0031 | 0.0149 |
| $P$ Dense | SE$(P, P)$ | 3.3 × 10$^{-5}$ | 6.9 × 10$^{-7}$ | 0.0888 |

**TABLE 1: Comparison with RPCA algorithms.** IALM is the solver for PCP.

## VI. Conclusions and Extensions

In this work, we studied the PCA problem when a part of the noise is data-dependent and, hence, the noise and data are correlated. We also assumed bounded-ness of both data and noise, which is, in fact, a more practical assumption than Gaussianity, since most sensors or other data collection devices have finite power. We showed that, with as few as $\alpha = C r (\log n) f^2$ samples, one can achieve subspace recovery error that is below a fraction of $\varepsilon$ whp. Recall that $q$ bounds the noise-to-signal ratio. If the condition number $f$ is $O(1)$, then, up to constants, our sample complexity bound is only $(\log n)$ times the minimum required which would be $r$.

Further improvements. The result given here assumes that the $\ell_i$’s are bounded and mutually independent random variables. Both assumptions can be relaxed. Mutual independence can be replaced by an autoregressive (AR) model on the $\ell_i$’s. As long as the AR parameter is not too large, it is possible to get a result that is slightly weaker than the one above by using the matrix Azuma inequality \[18\] (the approach will be similar to that used to analyze the subspace deletion step of ReProCS in \[19\]). We can also replace the bounded-ness assumption by a sub-Gaussianity assumption, but this would need $\alpha \geq C f^2 n$. Thus, in the unbounded case, one would need $O(n)$ samples for correlated-PCA; this is similar to the complexity of various other PCA results for uncorrelated or independent Gaussian noise, e.g., \[6\], \[8\].

Ongoing and Future Work. In ongoing work, we are studying the problem of correlated-PCA with partial subspace knowledge, and its application to analyzing the subspace update (addition) step of ReProCS for dynamic robust PCA \[4\]. This approach has helped to significantly simplify the ReProCS correctness proof. This simplification has also helped to significantly improve the original correctness guarantee given in \[4\]. A useful open question is how to generalize algorithms for streaming PCA, e.g., the block-stochastic power method studied in \[8\], or Oja’s algorithm, in the correlated-PCA setting. In recent literature \[8\], \[20\], strong guarantees for both have appeared in the general uncorrelated noise setting.

A. Extensions - cluster-EVD (cluster-SVD)

In \[1\], we introduced an improvement of simple SVD (simple EVD) called cluster-EVD. This assumes that the eigenvalues of $\Lambda$ are clustered, i.e., there exists a partition of the index set \{1, 2, ..., $r$\} into subsets $G_1, G_2, ..., G_K$ so that $\lambda^+_k := \max_{i \in G_k} \lambda_i(\Lambda)$ and $\lambda^-_k := \min_{i \in G_k} \lambda_i(\Lambda)$ satisfy the following: $\lambda^+_{k+1} < \lambda^-_k$, $\lambda^+_{k}/\lambda^-_k \leq \epsilon < \chi < 1$. In words, the clusters are arranged in decreasing order of eigenvalues; the condition number within a cluster is at most $\epsilon$, and the normalized gap between consecutive clusters’ eigenvalues is at least $1 - \chi$. We say that the eigenvalues are well-clustered when $\epsilon \ll f$ and $\chi < 1$.

To understand the basic idea of the cluster-EVD algorithm, suppose that the clusters are known\[1\]. Thus $r_k := |G_k|$ is also known. Let $G_k := \{P \cap G_k\}$ and let $G_k$ denote its estimate. Cluster-SVD computes $\hat{G}_1$ as the top $r_1$ eigenvectors of $\sum_{i=1}^n y_i y_i'$. For each $k > 1$, it computes $\hat{G}_k$ as the top $r_k$ eigenvectors of $\Phi \sum_{i=1}^n y_i y_i' \Phi'$ where $\Phi := \mathbf{I} - \hat{G}_1 \hat{G}_1' - \hat{G}_2 \hat{G}_2' - ... - \hat{G}_{k-1} \hat{G}_{k-1}'$. After $K$ such steps, it sets $\hat{P} = \{\hat{G}_1, \hat{G}_2, ..., \hat{G}_K\}$.

By using the Vershynin result or matrix Bernstein to replace matrix Hoeffding at various places in the cluster-EVD proof of \[1\], it is possible to show that, to get $\text{SE}(\hat{P}, P) \leq K \epsilon$, cluster-EVD needs

$$\alpha \geq C \eta d r (\log n + \log K)^2 \sqrt{(gf)^2}$$

This will hold as long as $\sqrt{bo} \eta g = 0.15 \epsilon$, and $\chi < 0.4$. Simple SVD needs $\sqrt{bo} f g = 0.15 \epsilon$ and $\alpha \geq C \eta d r (\log n)^2 \epsilon^2 f^2$. Both are stronger requirements since $f > g$. However, the dependence of any sample complexity on $f$ or $g$ is not tight (in numerical experiments, the sample complexity of EVD does not increase as $f^2$ for example), and thus, it is hard to see the difference between cluster-EVD and EVD performance via simulations. This explains the results of the numerical experiments of \[1\].

## References

1. N. Vaswani and H. Guo, “Correlated-pca: Principal components’ analysis when data and noise are correlated,” in *Adv. Neural Inf. Proc. Sys. (NIPS)*, 2016.
2. E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *Journal of ACM*, vol. 58, no. 3, 2011.
3. C. Qiu, N. Vaswani, B. Lois, and L. Høgben, “Recursive robust pca or recursive sparse recovery in large but structured noise,” *IEEE Trans. Info. Th.*, pp. 5007–5039, August 2014.
4. B. Lois and N. Vaswani, “Online matrix completion and online robust pca,” in *IEEE Intl. Symp. Info. Th. (ISIT)*, 2015.
5. Jussi Gillberg, Pekka Marttinen, Matti Pirinen, Antti J Kangas, Pasi Soininen, Mehreen Ali, Aki S Havulinna, Marjo-Riitta Marjo-Riitta Jarvelin, Mika Ala-Korpela, and Samuel Kaski, “Multiple output regression with latent noise,” *Journal of Machine Learning Research*, 2016.
6. B. Nadler, “Finite sample approximation results for principal component analysis: A matrix perturbation approach,” *The Annals of Statistics*, vol. 36, no. 6, 2008.
7. R. Vershynin, “How close is the sample covariance matrix to the actual covariance matrix?,” *Journal of Theoretical Probability*, pp. 1–32, 2010.
8. I. Mitliagkas, C. Caramanis, and P. Jain, “Memory limited, streaming pca,” in *Adv. Neural Inf. Proc. Sys. (NIPS)*, 2013, pp. 2886–2894.
9. V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Rank-sparsity incoherence for matrix decomposition,” *SIAM Journal on Optimization*, vol. 21, 2011.
10. D. Hsu, S.M. Kakade, and T. Zhang, “Robust matrix decomposition with sparse corruptions,” *IEEE Trans. Info. Th.*, Nov. 2011.

3As explained in \[1\], these can be estimated automatically also.
[11] P. Netrapalli, U N Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, “Non-convex robust pca,” in Neural Info. Proc. Sys. (NIPS), 2014.
[12] Huan Xu, Constantine Caramanis, and Shie Mannor, “Outlier-robust pca: the high-dimensional case,” IEEE Trans. Info. Th., vol. 59, no. 1, pp. 546–572, 2013.
[13] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis, “Fast algorithms for robust pca via gradient descent,” in Neural Info. Proc. Sys. (NIPS), 2016.
[14] Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain, “Nearly-optimal robust matrix completion,” arXiv preprint arXiv:1606.07315, 2016.
[15] 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.
[16] Ren-Cang Li, “Relative perturbation theory: ii. eigenspace and singular subspace variations,” SIAM Journal on Matrix Analysis and Applications, vol. 20, no. 2, pp. 471–492, 1998.
[17] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” Compressed sensing, pp. 210–268, 2012.
[18] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” Found. Comput. Math., vol. 12, no. 4, 2012.
[19] J. Zhan, B. Lois, H. Guo, and N. Vaswani, “Online (and Offline) Robust PCA: Novel Algorithms and Performance Guarantees,” in Intl. Conf. Artif. Intell. and Stat. (AISTATS), 2016.
[20] Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford, “Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for ojas algorithm,” in 29th Annual Conference on Learning Theory, 2016, pp. 1147–1164.