PCA in Data-Dependent Noise (Correlated-PCA): Improved Finite Sample Guarantees

Namrata Vaswani and Praneeth Narayanamurthy
Dept of ECE, Iowa State University, Ames IA 50010, USA

***Abstract*—We study Principal Component Analysis (PCA) in a setting where a part of the corrupting noise is data-dependent, and, as a result, the corrupting noise and the true data are correlated. We provide a nearly optimal sample complexity bound for the most common PCA solution, simple singular value decomposition (SVD). Our bound, which holds under a bounded-ness and mutual independence assumption on the true data and a few assumptions on the data-noise correlation, is within a logarithmic factor of the best achievable. We first studied this problem in recent work (NIPS 2016) where we called it "correlated-PCA".**

I. Introduction

We study Principal Component Analysis (PCA) in a setting where a part of the corrupting noise is data-dependent, and, as a result, the corrupting noise and the true data are correlated (correlated-PCA [1]). Under a bounded-ness and mutual independence assumption on the true data 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*. For the reader who is more familiar with eigenvalue decomposition (EVD), this is equivalent to EVD of the sample covariance matrix of the observed data.

Problem Setting. For $t = 1, 2, \dots,$ we are given n-length data vectors, $\boldsymbol{y}_t$, that satisfy

$$\boldsymbol{y}_t := \boldsymbol{\ell}_t + \boldsymbol{w}_t + \boldsymbol{v}_t, \text{ where } \boldsymbol{\ell}_t = \boldsymbol{P}\boldsymbol{a}_t, \ \boldsymbol{w}_t = \boldsymbol{M}_t\boldsymbol{\ell}_t,$$

$\boldsymbol{P}$ is an $n \times r$ matrix with orthonormal columns and $r \ll n$; $\boldsymbol{\ell}_t$ is the true data vector that lies in a low (r) dimensional subspace of $\mathbb{R}^n$, $\text{range}(\boldsymbol{P})$; $\boldsymbol{a}_t$ is its projection into this subspace; $\boldsymbol{w}_t$ is the data-dependent (correlated) noise component; and $\boldsymbol{v}_t$ is the uncorrelated noise component, i.e., it satisfies $\mathbb{E}[\boldsymbol{\ell}_t \boldsymbol{v}_t{}'] = 0$. The matrices $\boldsymbol{M}_t$ are *unknown* and such that $\mathbb{E}[\boldsymbol{\ell}_t \boldsymbol{w}_t'] \neq 0$ (holds if $\|\mathbb{E}[\boldsymbol{M}_t]\| > 0$). The goal is to estimate $\text{range}(\boldsymbol{P})$. Since the matrices $\boldsymbol{M}_t$ are time-varying, observe that, in general, the $\boldsymbol{w}_t$'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 $\mathcal{T}_t$ denote the outlier support at time t. Then, robust PCA with data-dependent outlier values involves PCA from observed data $\boldsymbol{y}_t := \boldsymbol{\ell}_t + \boldsymbol{I}_{\mathcal{T}_t}\boldsymbol{s}_t + \boldsymbol{v}_t$ where $\boldsymbol{s}_t = \boldsymbol{M}_{s,t}\boldsymbol{\ell}_t$ with $\boldsymbol{M}_{s,t}$ being a $|\mathcal{T}_t| \times n$ matrix. Here $\boldsymbol{I}_{\mathcal{T}_t}\boldsymbol{M}_{s,t}$ is the data-dependency matrix. This model is often a valid one for video analytics applications, where $\boldsymbol{\ell}_t$ is the background layer of image frame t, $\mathcal{T}_t$ is the foreground support of frame t, and $\boldsymbol{s}_t$ is the difference between foreground and background intensities on $\mathcal{T}_t$. Another related example is the subspace update step of the Recursive Projected Compressive Sensing (ReProCS) solution to the dynamic robust PCA problem [3].

As explained in [4], 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 addresses three important limitations of [1]. (1) It gives a significantly improved sample complexity bound and one that is within a logarithmic factor of the best achievable sample complexity. (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 typically not fully data-dependent in most real applications. (3) We provide a method for automatic data dimension estimation that does not use knowledge of any model parameter (see Corollary 5).

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 [5] and references therein, [6], and see the summary of existing batch PCA guarantees given in [7]. 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], [8], [9], [10], [11]. 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 $\boldsymbol{P}$ to be dense (not sparse). On the other hand, the simple SVD solution is faster and does not require denseness of columns of $\boldsymbol{P}$; however it, of course, only works for small magnitude outliers. This point is demonstrated experimentally in Table I. We should mention that there are some very recent works on fast robust PCA methods such as [12], [13] that have the same order of computational complexity as simple SVD. However, these still require denseness of columns of $\boldsymbol{P}$, and will be slower than SVD in practice (since their initialization step itself involves an r-SVD).

II. Assumptions and Main Result

We assume the following about the true data $\boldsymbol{\ell}_t$ and the data-dependency matrix $\boldsymbol{M}_t$.

Asssumption 1. *The $\boldsymbol{\ell}_t$'s satisfy $\boldsymbol{\ell}_t = \boldsymbol{P}\boldsymbol{a}_t$ with $\boldsymbol{a}_t$'s being zero mean, mutually independent, and bounded r.v.'s, with diagonal covariance matrix, $\boldsymbol{\Lambda}$.*

Define $\lambda^- := \lambda_{\min}(\boldsymbol{\Lambda})$, $\lambda^+ := \lambda_{\max}(\boldsymbol{\Lambda})$ and $f := \frac{\lambda^+}{\lambda^-}$. Since the $\boldsymbol{a}_t$'s are bounded, there exists a finite constant, η, such that, $\max_{j=1,2,\ldots r} \max_t \frac{(\boldsymbol{a}_t)_j^2}{\lambda_j} \leq \eta$. Observe that η bounds the ratio of the square of the maximum magnitude of $\boldsymbol{a}_t$ 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 $\boldsymbol{a}_t$'s are i.i.d. uniform, then $\eta = 3$.

Asssumption 2. *The data-dependency matrices $\boldsymbol{M}_t$ can be split as $\boldsymbol{M}_t = \boldsymbol{M}_{2,t}\boldsymbol{M}_{1,t}$ with $\boldsymbol{M}_{2,t}$, $\boldsymbol{M}_{1,t}$ satisfying the following. For a $q < 1$, a $b_0 < 1$, and a positive integer α,*

$$0 < \|\boldsymbol{M}_{1,t}\boldsymbol{P}\|_2 \leq q < 1, \ \|\boldsymbol{M}_{2,t}\|_2 \leq 1, \ \textit{and} \tag{1}$$

$$\left\| \frac{1}{\alpha} \sum_{t=1}^{\alpha} \boldsymbol{M}_{2,t}\boldsymbol{A}_t\boldsymbol{M}_{2,t}{}' \right\|_2 \leq b_0 \max_{t\in[1,\alpha]} \|\boldsymbol{A}_t\|. \tag{2}$$

for any α-length sequence of positive semi-definite Hermitian matrices, $\boldsymbol{A}_t$.

Assumption 1 just states mutual independence and boundedness of the $\boldsymbol{\ell}_t$'s. The first part of Assumption 2 bounds the instantaneous noise-to-signal ratio of the correlated (data-dependent) component of the noise, $\boldsymbol{w}_t$: using it, $\|\boldsymbol{w}_t\|_2 \leq q\|\boldsymbol{a}_t\|_2 = q\|\boldsymbol{\ell}_t\|_2$ and $\|\mathbb{E}[\boldsymbol{w}_t\boldsymbol{w}_t{}']\|_2 \leq q^2\|\mathbb{E}[\boldsymbol{\ell}_t\boldsymbol{\ell}_t{}']\|_2$. The second part can be understood as one way to reduce the time-averaged power of $\boldsymbol{w}_t$. Observe that, $\|\mathbb{E}[\boldsymbol{w}_t\boldsymbol{w}_t{}']\|_2 \leq q^2\lambda^+$, whereas, $\|\frac{1}{\alpha}\sum_{t=1}^{\alpha}\mathbb{E}[\boldsymbol{w}_t\boldsymbol{w}_t{}']\|_2 \leq b_0 q^2\lambda^+$. 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-Schwartz, it is not hard to see that $\|\frac{1}{\alpha}\sum_{t=1}^{\alpha}\mathbb{E}[\boldsymbol{\ell}_t\boldsymbol{w}_t{}']\|_2 \leq \sqrt{b_0}q\lambda^+$.

One example where Assumption 2 holds is when $\boldsymbol{w}_t$ is sparse with time-varying support sets, denoted $\mathcal{T}_t$. In this case, $\boldsymbol{M}_{2,t} = \boldsymbol{I}_{\mathcal{T}_t}$. If all the sets $\mathcal{T}_t$ 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 $\boldsymbol{A}_t$. 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 [1].

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

$$\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) := \|(\boldsymbol{I} - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{P}\|_2,$$

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 4 and 5).

[1] $\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$ quantifies the principal angle between the column spans of $\hat{\boldsymbol{P}}$ and $\boldsymbol{P}$ (this is a valid definition when $\hat{\boldsymbol{P}}$ and $\boldsymbol{P}$ have orthonormal columns).

Algorithm 1 Simple SVD (or EVD)

Let $\hat{\boldsymbol{P}}$ be the matrix of top r singular vectors of $[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_\alpha]$. Equivalently, $\hat{\boldsymbol{P}}$ is the matrix of top r eigenvectors of $\frac{1}{\alpha}\sum_{t=1}^{\alpha}\boldsymbol{y}_t\boldsymbol{y}_t{}'$.

Theorem 3. *Assume that $\boldsymbol{v}_t$ satisfies $\|\mathbb{E}[\boldsymbol{v}_t\boldsymbol{v}_t{}']\|_2 \leq \lambda_{\boldsymbol{v}}^+$ and $\|\boldsymbol{v}_t\|_2^2 \leq \eta r_{\boldsymbol{v}}\lambda_{\boldsymbol{v}}^+$. For an $\varepsilon_{\mathrm{SE}} < 1$, define $d := \max(1, \frac{\eta(r\log 9 + 10\log n)\varepsilon_{\mathrm{SE}}^2}{r(\log n)q^2})$ and*

$$\alpha_0 := C\eta d \frac{(\log n)\max\left(rf^2q^2, r_{\boldsymbol{v}}(\frac{\lambda_{\boldsymbol{v}}^+}{\lambda^-})^2, \max(r_{\boldsymbol{v}}, r)f\frac{\lambda_{\boldsymbol{v}}^+}{\lambda^-}\right)}{\varepsilon_{\mathrm{SE}}^2}.$$

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

If $3.3\sqrt{b_0}qf = \varepsilon_{\mathrm{SE}}/4$ and $1.3\frac{\lambda_{\boldsymbol{v}}^+}{\lambda^-} = \varepsilon_{\mathrm{SE}}/4$, then, with probability at least $1 - 10n^{-10}$,

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

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 $\|\boldsymbol{w}_t\|_2$ and $\|\boldsymbol{v}_t\|_2$. Thus, suppose that $\lambda_{\boldsymbol{v}}^+ = b_0q^2\lambda^+$ and $\eta r_{\boldsymbol{v}}\lambda_{\boldsymbol{v}}^+ = \eta r q^2\lambda^+$. Then, in the only correlated-noise case ($\boldsymbol{v}_t = 0$), we need $3.3\sqrt{b_0}qf < \varepsilon_{\mathrm{SE}}/2$, and $\alpha \geq C\eta d\frac{(\log n)rf^2q^2}{\varepsilon_{\mathrm{SE}}^2}$. In the $\boldsymbol{w}_t = 0$ case, we need $3.3b_0q^2f < \varepsilon_{\mathrm{SE}}/2$ and $\alpha \geq C\eta d\frac{(\log n)\frac{r}{b_0}(b_0q^2f)f}{\varepsilon_{\mathrm{SE}}^2} = C\eta d\frac{(\log n)rf^2q^2}{\varepsilon_{\mathrm{SE}}^2}$. Thus the α required in both cases is the same. However, the upper bound on f needed in the correlated noise case is stronger. For example, say $\varepsilon_{\mathrm{SE}} = q/4$. Then, in the only correlated noise case, one needs $f < 1(25\sqrt{b_0}q)$, while, in the only uncorrelated noise case, one needs $f < 1(25b_0q^2)$.

The reason that the correlated noise case is harder is as follows. The bound on $\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$, given by the Davis-Kahan $\sin\theta$ theorem [14], is governed by the ratio between the spectral norm of the perturbation matrix, $\boldsymbol{H} := \frac{1}{\alpha}\sum_t \boldsymbol{y}_t\boldsymbol{y}_t{}' - \frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'$, and the minimum eigenvalue along the principal subspace, λ^-. In the correlated noise case, the dominant terms in $\boldsymbol{H}$ are the signal-noise correlation terms, $\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t{}'$ and its transpose. Since the noise is smaller than signal ($q < 1$), these terms are larger than the noise power terms $\frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t{}'$ or $\frac{1}{\alpha}\sum_t \boldsymbol{v}_t\boldsymbol{v}_t{}'$. In the only uncorrelated noise case ($\boldsymbol{w}_t = 0$ case), the signal-noise correlation terms are nearly zero with high probability and the only non-negligible term is $\frac{1}{\alpha}\sum_t \boldsymbol{v}_t\boldsymbol{v}_t{}'$.

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 [15] 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 A and of its perturbed version

$D'AD$ for a non-singular matrix D. However, such results are not applicable for our problem since $\boldsymbol{M}_t$ is time-varying.

Comparison with [1]. The result of [1] assumed that $\boldsymbol{v}_t = 0$. Thus, to compare with it let $\boldsymbol{v}_t = 0$ so that $\lambda_{\boldsymbol{v}}^+ = 0$ and $r_{\boldsymbol{v}} = 0$ in Theorem 3. First consider the case where the desired final error is smaller than the noise level, i.e., $\varepsilon_{\text{SE}} < q$. In this case, $d = 1$, and so, Theorem 3 shows that the sample complexity, α, is lower bounded by $Cf^2 r(\log n)\frac{q^2}{\varepsilon_{\text{SE}}^2}$. This bound holds as long as $\sqrt{b_0}qf < \varepsilon_{\text{SE}}/6.6$. Thus, to get the subspace error to below $\varepsilon_{\text{SE}} = q/4$, we need $\sqrt{b_0}f < 1/28$ and $\alpha \geq 16Cr(\log n)f^2$ samples. This is much better than our earlier sample complexity bound [1] of $Cr^2(\log n)\frac{f^2}{\varepsilon_{\text{SE}}^2}$ [1] which implies that we need $\alpha \geq 16Cr^2(\log n)\frac{f^2}{q^2}$ 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 $r(\log n)$. We get the first improvement by bounding the r-th eigenvalue of $\sum_t \boldsymbol{\ell}_t\boldsymbol{\ell}_t{}' = \boldsymbol{P}(\sum_t \boldsymbol{a}_t\boldsymbol{a}_t{}')\boldsymbol{P}'$ by using a result of Vershynin [16, Theorem 5.39] to bound the minimum eigenvalue of $\sum_t \boldsymbol{a}_t\boldsymbol{a}_t{}'$. In [1], 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 $\varepsilon_{\text{SE}} > q$ (this is a useful scenario only when q is small) and n is small enough, $d = \frac{\eta(r\log 9 + 10\log n)\varepsilon_{\text{SE}}^2}{r(\log n)q^2}$ and, so, in this case, our result needs an even smaller α: $\alpha \geq Cf^2(r\log 9 + 10\log n)$ suffices. In fact, in this scenario, if we let the subspace error bound hold with probability only at least $1 - c\exp(-cr)$, we will only need $\alpha \geq Cf^2 r$.

Matching lower bound. The minimum number of samples required to estimate the subspace $\text{range}(\boldsymbol{P})$ 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 get the dependence on n because the $\boldsymbol{w}_t$'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 boundedness assumption on both $\boldsymbol{\ell}_t$ and $\boldsymbol{w}_t$. If this 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., [5] (finite sample guarantee for $r = 1$ dimensional PCA) or [7] (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 note here that there is a large amount of literature on online PCA which we do not cite or discuss here (since it is not a problem this work is solving).

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 $\lambda_j(\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') \geq 0.5\alpha\lambda^-$ and thus requires knowledge of λ^-. We have the following corollary.

Corollary 4. *In the setting of Theorem 3, if $\varepsilon_{\text{SE}} < 1/2$, then, with probability $\geq 1 - 10n^{-10}$,*

1) $\lambda_r(\frac{1}{\alpha}\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') \geq \lambda^-(0.98 - \varepsilon_{\text{SE}}/2) \geq 0.73\lambda^-$, and
2) $\lambda_{r+1}(\frac{1}{\alpha}\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') \leq (\varepsilon_{\text{SE}}/2)\lambda^- < 0.25\lambda^-$,

and thus, the above approach returns $\hat{r} = r$.

An alternate way to estimate r is as $\hat{r} := \arg\max_j[\lambda_j(\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') - \lambda_{j+1}(\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}')]$. This does not require knowledge of λ^-. But, it is more expensive (needs all eigenvalues), and, as we see below, it needs one extra assumption.

Corollary 5. *In the setting of Theorem 3, let $\varepsilon_{\text{SE}} < 1/4$. Assume also that $\lambda_j(\boldsymbol{\Lambda}) - \lambda_{j+1}(\boldsymbol{\Lambda}) \leq 0.45\lambda^-$ for all $j = 1, 2, \ldots, r$. Then, with probability $\geq 1 - 10n^{-10}$,*

1) for a $j < r$, $\lambda_j(\frac{1}{\alpha}\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') - \lambda_{j+1}(\frac{1}{\alpha}\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') \leq 0.45\lambda^- + 2(\varepsilon_{\text{SE}}/2)\lambda^- < 0.7\lambda^-$,
2) for a $j > r$, $\lambda_j(\frac{1}{\alpha}\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') - \lambda_{j+1}(\frac{1}{\alpha}\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') \leq 2(\varepsilon_{\text{SE}}/2)\lambda^- < 0.25\lambda^-$, and
3) for $j = r$, $\lambda_j(\frac{1}{\alpha}\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') - \lambda_{j+1}(\frac{1}{\alpha}\sum_{t=1}^{\alpha} \boldsymbol{y}_t\boldsymbol{y}_t{}') \geq \lambda^-(0.98 - 2(\varepsilon_{\text{SE}}/2)) > 0.73\lambda^-$,

and thus, the above approach returns $\hat{r} = r$.

IV. Proof of Theorem 3

To see a simple proof first, suppose that $\boldsymbol{v}_t = 0$. 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 Theorem 3 with $\boldsymbol{v}_t = 0$. Using the Davis-Kahan $\sin\theta$ theorem [14] followed by Weyl's inequality (see [1]),

$$\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \leq \frac{2\|\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t{}'\|_2 + \|\frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t{}'\|_2}{\lambda_r(\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{\ell}_t{}') - (2\|\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t{}'\|_2 + \|\frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t{}'\|_2)}$$

if the denominator is positive. The two numerator terms can be bounded using the matrix Bernstein inequality [17]. Observe that $\lambda_r(\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{\ell}_t{}') = \lambda_{\min}(\frac{1}{\alpha}\sum_t \boldsymbol{a}_t\boldsymbol{a}_t{}')$. We can bound $\lambda_{\min}(\frac{1}{\alpha}\sum_t \boldsymbol{a}_t\boldsymbol{a}_t{}')$ using Theorem 5.39 of [16]. Since the $\boldsymbol{a}_t$'s are bounded, they are sub-Gaussian with sub-Gaussian norm bounded by $\sqrt{\eta\lambda^+}$. Because the $\boldsymbol{a}_t$'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 [17] says the following. For a finite sequence of $d_1 \times d_2$ zero mean independent matrices $\boldsymbol{Z}_k$ with

$$\|\boldsymbol{Z}_k\|_2 \leq R,\ \max(\|\sum_k \mathbb{E}[\boldsymbol{Z}_k{}'\boldsymbol{Z}_k]\|_2, \|\sum_k \mathbb{E}[\boldsymbol{Z}_k\boldsymbol{Z}_k{}']\|_2) \leq \sigma^2,$$

we have $\mathbb{P}(\|\sum_k \boldsymbol{Z}_k\|_2 \geq s) \leq (d_1 + d_2)\exp\left(-\frac{s^2/2}{\sigma^2 + Rs/3}\right)$.

Let $\boldsymbol{Z}_t := \boldsymbol{\ell}_t\boldsymbol{w}_t{}'$. We apply this result to $\tilde{\boldsymbol{Z}}_t := \boldsymbol{Z}_t - \mathbb{E}[\boldsymbol{Z}_t]$ with $s = \epsilon\alpha$. To get the values of R and σ^2 in a simple fashion, we use the facts that (i) If $\|\boldsymbol{Z}_t\|_2 \leq R_1$, Then $\|\tilde{\boldsymbol{Z}}_t\| \leq 2R_1$; and (ii) $\sum_t \mathbb{E}[\tilde{\boldsymbol{Z}}_t\tilde{\boldsymbol{Z}}_t{}'] \preccurlyeq \sum_t \mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t{}']$. Thus,

we can set R to two times the bound on $\|\boldsymbol{Z}_t\|_2$ and we can set σ^2 as the maximum of the bounds on $\|\sum_t \mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t{}']\|_2$ and $\|\sum_t \mathbb{E}[\boldsymbol{Z}_t{}'\boldsymbol{Z}_t]\|_2$.

It is easy to see that $R = 2\sqrt{\eta r\lambda^+}\sqrt{\eta r q^2\lambda^+} = 2\eta r q\lambda^+$. To get σ^2, observe that

$$\left\|\sum_t \mathbb{E}[\boldsymbol{w}_t\boldsymbol{\ell}_t{}'\boldsymbol{\ell}_t\boldsymbol{w}_t{}']\right\|_2 \leq \alpha(\max_{\boldsymbol{\ell}_t}\|\boldsymbol{\ell}_t\|^2)\cdot\|\mathbb{E}[\boldsymbol{w}_t\boldsymbol{w}_t{}']\| \leq \alpha\eta r\lambda^+\cdot q^2\lambda^+ = \alpha\eta r q^2(\lambda^+)^2.$$

Repeating the above steps, we get the same bound on $\|\sum_t \mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t{}']\|_2$. Thus, $\sigma^2 = \alpha\eta r q^2(\lambda^+)^2$.

Thus, we conclude that

$$\left\|\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t{}' - \mathbb{E}[\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t{}']\right\|_2 \geq \epsilon\alpha$$

w.p. at most $2n\exp\left(-\frac{\epsilon^2\alpha^2/2}{\alpha\eta r q^2(\lambda^+)^2+\eta r q\lambda^+\epsilon\alpha/3}\right)$.

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

Thus, with probability at least $1-2n\exp\left(-\alpha\frac{\epsilon^2}{4\eta r q^2(\lambda^+)^2}\right)$,

$$\left\|\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t{}'\right\|_2 \leq \left\|\frac{1}{\alpha}\mathbb{E}[\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t{}']\right\|_2 + \epsilon \leq \sqrt{b_0}q\lambda^+ + \epsilon$$

as long as $\epsilon < q\lambda^+$. Set $\epsilon = \epsilon_0\lambda^-$, then we get: as long as $\epsilon_0 < qf$, with probability at least $1-2n\exp\left(-\alpha\frac{\epsilon_0^2}{4\eta r q^2 f^2}\right)$,

$$\left\|\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t{}'\right\|_2 \leq \sqrt{b_0}q\lambda^+ + \epsilon = [\sqrt{b_0}qf + \epsilon_0]\lambda^-$$

Thus, the above event holds w.p. at least $1-2n^{-10}$ if

$$\alpha \geq \alpha_0 = (11\log n)4\eta r\frac{q^2f^2}{\epsilon_0^2} = 44\eta r(\log n)\frac{q^2f^2}{\epsilon_0^2}$$

and $\epsilon_0 \leq qf$.

Consider the second term. Proceeding as above, we get $R = 2\eta r q^2\lambda^+$ and $\sigma^2 = \alpha\sigma_1^2$, $\sigma_1^2 = \eta r q^4(\lambda^+)^2$. Thus, with probability at least $1-2n\exp\left(-\alpha\frac{1}{4\frac{\sigma^2}{\epsilon_2^2(\lambda^-)^2}+4\frac{R}{\epsilon_2\lambda^-}}\right)$,

$$\left\|\frac{1}{\alpha}\sum_t \boldsymbol{w}_t\boldsymbol{w}_t{}'\right\|_2 \leq \left\|\frac{1}{\alpha}\mathbb{E}[\sum_t \boldsymbol{w}_t\boldsymbol{w}_t{}']\right\| + \epsilon_2\lambda^- \leq [b_0q^2f + \epsilon_2]\lambda^-$$

Thus, the above event holds w.p. at least $1-2n^{-10}$ if

$$\alpha \geq \alpha_2 = 44\eta r(\log n)\max(\frac{q^4f^2}{\epsilon_2^2}, \frac{2q^2f}{\epsilon_2})$$

Using Theorem 5.39 of [16] applied to $\frac{1}{\alpha}\sum_t \boldsymbol{a}_t\boldsymbol{a}_t{}'$, and using the fact that the $\boldsymbol{a}_t$'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 $1-2\exp\left(r\log 9 - \alpha\frac{c(\epsilon_1\lambda^-)^2}{(4\eta\lambda^+)^2}\right) = 1-2\exp\left(r\log 9 - \alpha\frac{c\epsilon_1^2}{16\eta^2f^2}\right)$,

$$\lambda_r\left(\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{\ell}_t'\right) = \lambda_{\min}\left(\frac{1}{\alpha}\sum_t \boldsymbol{a}_t\boldsymbol{a}_t'\right) \geq \lambda^-(1-\epsilon_1)$$

Thus, the above event holds w.p. at least $1-2n^{-10}$ if

$$\alpha \geq \alpha_1 = \frac{(r\log 9 + 10\log n)\cdot 16\eta^2f^2}{\epsilon_1^2} = 16\eta^2c(r\log 9 + 10\log n)\frac{f^2}{\epsilon_1^2}$$

Thus, we have the following result.

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

$$\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \leq \frac{2qf\sqrt{b_0} + q^2fb_0 + 2\epsilon_0 + \epsilon_2}{1-\epsilon_1-\mathrm{numer}}$$

as long as numer $< 1-\epsilon_1$, $\epsilon_0 < qf$. *Here* numer *refers to the numerator term.*

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

$$\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \leq \frac{3.3qf\sqrt{b_0}}{0.98 - 3.3qf\sqrt{b_0}}$$

if denominator is positive. To set the RHS equal to an $\varepsilon_{\mathrm{SE}} < 1$, it suffices to let

$$3.3qf\sqrt{b_0} = 0.98\frac{\varepsilon_{\mathrm{SE}}}{2} = 0.49\varepsilon_{\mathrm{SE}}$$

This means,

$$b_0 = \frac{0.49^2\varepsilon_{\mathrm{SE}}^2}{3.3^2q^2f^2} = 0.022\frac{\varepsilon_{\mathrm{SE}}^2}{q^2f^2}$$

and so

$$\frac{1}{b_0} = \frac{3.3^2q^2f^2}{0.49^2\varepsilon_{\mathrm{SE}}^2} = 45.36\frac{q^2f^2}{\varepsilon_{\mathrm{SE}}^2}$$

With this,

$$\alpha_0 = 4400\eta r(\log n)\frac{1}{b_0} = C\eta r(\log n)\frac{q^2f^2}{\varepsilon_{\mathrm{SE}}^2}$$

Recall that

$$\alpha_1 = C\eta^2(r\log 9 + 10\log n)f^2$$

Define

$$d = \max\left(1, \eta\frac{(r\log 9 + 10\log n)\varepsilon_{\mathrm{SE}}^2}{r(\log n)q^2}\right)$$

Thus, if

$$\alpha \geq Cd\eta\ r(\log n)\frac{q^2f^2}{\varepsilon_{\mathrm{SE}}^2}$$

and if $b_0 = 0.022\frac{\varepsilon_{\mathrm{SE}}^2}{q^2f^2}$, then

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

⊠

Now consider the general case $\boldsymbol{v}_t \neq 0$. We get the final result for this case by also using the following lemma (which again follows by matrix Bernstein).

Lemma 7. *Pick an $\epsilon_{0,\boldsymbol{v}} > 0$.*

1) *With probability at least* $1 - 2n\exp\left(-\alpha\frac{(\epsilon_{0,\boldsymbol{v}}\lambda^-)^2}{4\eta\max(r_{\boldsymbol{v}},r)\lambda^+\lambda^+_{\boldsymbol{v}}}\right)$,

$$\left\|\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{v}_t{}'\right\|_2 \le \epsilon_{0,\boldsymbol{v}}\lambda^-$$

2) *With probability at least* $1 - 2n\exp(-\frac{\alpha(\epsilon_{0,\boldsymbol{v}}\lambda^-)^2}{4\eta r_{\boldsymbol{v}}(\lambda^+_{\boldsymbol{v}})^2})$,

$$\left\|\frac{1}{\alpha}\sum_t \boldsymbol{v}_t\boldsymbol{v}_t{}'\right\|_2 \le \lambda^+_{\boldsymbol{v}} + \epsilon_{0,\boldsymbol{v}}\lambda^-$$

V. Experiments

We repeat the experiments from [1] here. The results of three experiments are shown in Table I. **Experiment 1:** We generated data with $n = 500$. 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 condition number $f = 1000$. The data-dependent 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 so that Assumption 2 holds with $\alpha = 300$ and $b_0 = 4/\alpha$ (the sets $\mathcal{T}_t$ follow Assumption 1.3 of [1] with $s = 5$, $\rho = 2$, and $\tilde{\beta} = 1$). The entries of $\boldsymbol{M}_{s,t}$ were iid $\mathcal{N}(0, q^2)$ with $q = 0.01$. The uncorrelated noise $\boldsymbol{v}_t = 0$. Observe that, since the columns of $\boldsymbol{P}$ are sparse, both PCP (Principal Components' Pursuit [2]) and AltMinRPCA [10] fail. Both have average $\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$ close to one whereas the average SE of SVD is 0.0911. Moreover, both of these are much slower than SVD as well. **Experiment 2:** Data was generated as above, but columns of $\boldsymbol{P}$ were dense. In this case, of course the robust PCA solutions PCP and A-M-RPCA outperform simple SVD. However, they are still much slower than simple SVD.

Experiment 3: We used images of a low-rankified real video sequence (escalator sequence from http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html) as $\boldsymbol{\ell}_t$'s.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 small.

VI. Conclusions and Extensions

In this work, we studied the PCA problem when the noise and data are correlated (a part of the noise is data-dependent). We showed that, with as few as $\alpha = Cr(\log n)f^2$ samples, one can achieve subspace recovery error that is a constant fraction of q. Recall that q bounds the noise-to-signal ratio. If the condition number f is assumed to be a constant, then, up to constants, this is only $(\log n)$ times the minimum required which would be r.

Further improvements. The result given here assumes that the $\boldsymbol{\ell}_t$'s are bounded and mutually independent random variables. Both assumptions can be relaxed. Mutual independence can be replaced by an autoregressive model on the $\boldsymbol{\ell}_t$'s, then, as long as the autocorrelation parameter is not too large, it is be possible to get a result that is slightly weaker than the one above by using the matrix Azuma inequality [17] (the approach will be similar to that used to analyze the subspace update step of ReProCS in [18]; this step also involves a correlated-PCA problem). It will require $r^2 \log n$ samples instead of $r \log n$. We can also replace the boundedness assumption by a sub-Gaussianity assumption, as long as $\alpha \ge Cf^2n$. Thus, in the unbounded case one would need $O(n)$ samples; this is similar to the sample complexity of various other PCA results for uncorrelated or independent Gaussian noise, e.g., [5], [7].

	Mean Subspace Error (SE)			Execution Time (seconds)		
	SVD	PCP	A-M-RPCA	SVD	PCP	A-M-RPCA
Experiment 1 ($\boldsymbol{\ell}_t = \boldsymbol{P}\boldsymbol{a}_t$, $\boldsymbol{P}$ sparse)	0.0911	1.0000	1.0000	0.0255	0.2361	0.0810
Experiment 2 ($\boldsymbol{\ell}_t = \boldsymbol{P}\boldsymbol{a}_t$, $\boldsymbol{P}$ dense)	0.07233	0.00000015686	0.000011865	0.0237	0.6989	0.1504
Experiment 3 ($\boldsymbol{\ell}_t$'s from real video)	0.3821	0.4970	0.4846	0.0223	1.6784	5.5144

TABLE I: *Comparison of* $\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$ *and execution time (in seconds). We compare SVD (Algorithm 1) with two robust PCA solutions - PCP (Principal Components' Pursuit [2]) and A-M-RPCA (Alt-Min-RPCA [10]). Table taken from [1].*

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 $\boldsymbol{\Lambda}$ are clustered, i.e., there exists a partition of the index set $\{1, 2, \dots, r\}$ into subsets $\mathcal{G}_1, \mathcal{G}_2, \dots \mathcal{G}_K$ so that $\lambda^+_k := \max_{i\in\mathcal{G}_k}\lambda_i(\boldsymbol{\Lambda})$ and $\lambda^-_k := \min_{i\in\mathcal{G}_k}\lambda_i(\boldsymbol{\Lambda})$ satisfy the following: $\lambda^+_{k+1} < \lambda^-_k$, $\lambda^+_k/\lambda^-_k \le g < f$ and $\lambda^+_{k+1}/\lambda^-_k < \chi < 1$. In words, the clusters are arranged in decreasing order of eigenvalues; the condition number within a cluster is at most g, and the normalized gap between consecutive clusters' eigenvalues is at least $1 - \chi$. We say that the eigenvalues are well-clustered when $g \ll f$ and $\chi \ll 1$.

To understand the basic idea of the cluster-EVD algorithm, suppose that the clusters are known[2]. Thus $r_k := |\mathcal{G}_k|$ is also known. Let $\boldsymbol{G}_k := (\boldsymbol{P})_{\mathcal{G}_k}$ and let $\hat{\boldsymbol{G}}_k$ denote its estimate. Cluster-EVD computes $\hat{\boldsymbol{G}}_1$ as the top r_1 eigenvectors of $\sum_{t=1}^{\alpha}\boldsymbol{y}_t\boldsymbol{y}_t{}'$. For each $k > 1$, it computes $\hat{\boldsymbol{G}}_k$ as the top r_k eigenvectors of $\boldsymbol{\Phi}\sum_{t=(k-1)\alpha+1}^{k\alpha}\boldsymbol{y}_t\boldsymbol{y}_t{}'\boldsymbol{\Phi}$ where $\boldsymbol{\Phi} := \boldsymbol{I} - \hat{\boldsymbol{G}}_1\hat{\boldsymbol{G}}_1{}' - \hat{\boldsymbol{G}}_2\hat{\boldsymbol{G}}_2{}' - \dots\hat{\boldsymbol{G}}_{k-1}\hat{\boldsymbol{G}}_{k-1}{}'$. After K such steps, it sets $\hat{\boldsymbol{P}} = [\hat{\boldsymbol{G}}_1, \hat{\boldsymbol{G}}_2, \dots \hat{\boldsymbol{G}}_K]$.

By using the Vershynin result and matrix Bernstein to replace matrix Hoeffding at various places in the cluster-EVD proof of [1], we can significantly improve its sample complexity (as compared to the result given in [1]). It is possible to show that, to get $\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \le K\varepsilon$, cluster-EVD needs

$$\alpha \ge C\eta d\ r(\log n + \log K)\frac{q^2}{\varepsilon^2}\max(g, (\varepsilon f))^2$$

This will hold as long as $\sqrt{b_0}gq < 0.15\varepsilon$, $\chi < 0.4$, and $\varepsilon f < g$. By substituting $\varepsilon \le g/f$, we get $\alpha \ge C\eta dr(\log n +$

[2] As explained in [1], these can be estimated automatically also.

$\log K)q^2f^2$. The sample complexity is $K\alpha$. On the other hand, for such an ε, EVD needs $\alpha \geq C\eta dr(\log n)q^2f^2(f/g)^2$. Thus, when $\varepsilon < g/f$, and K is small, say $K = 2$, the cluster-EVD sample complexity is roughly $(g/f)^2$ times smaller than that of EVD.

REFERENCES

[1] N. Vaswani and H. Guo, "Correlated-pca: Principal components' analysis when data and noise are correlated," in *Adv. Neural Info. 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. Hogben, "Recursive robust pca or recursive sparse recovery in large but structured noise," *IEEE Trans. Info. Th.*, pp. 5007–5039, August 2014.

[4] Jussi Gillberg, Pekka Marttinen, Matti Pirinen, Antti J Kangas, Pasi Soininen, Mehreen Ali, Aki S Havulinna, Marjo-Riitta Marjo-Riitta Järvelin, Mika Ala-Korpela, and Samuel Kaski, "Multiple output regression with latent noise," *Journal of Machine Learning Research*, 2016.

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

[6] R. Vershynin, "How close is the sample covariance matrix to the actual covariance matrix?," *Journal of Theoretical Probability*, pp. 1–32, 2010.

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

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

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

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

[11] 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.

[12] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis, "Fast algorithms for robust pca via gradient descent," in *Neural Info. Proc. Sys. (NIPS)*, 2016.

[13] Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain, "Nearly-optimal robust matrix completion," *arXiv preprint arXiv:1606.07315*, 2016.

[14] 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.

[15] 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.

[16] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Compressed sensing*, pp. 210–268, 2012.

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

[18] J. Zhan, B. Lois, H. Guo, and N. Vaswani, "Online (and Offline) Robust PCA: Novel Algorithms and Performance Guarantees," in *Intnl. Conf. Artif. Intell. and Stat. (AISTATS)*, 2016.