Improved Guarantees for Correlated-PCA (PCA when Data and Noise are Correlated)

Namrata Vaswani and Han Guo
Dept of ECE, Iowa State University, Ames IA 50010, USA

Problem Setting. We study Principal Components Analysis (PCA) in the setting where some of the corrupting "noise" or interference is correlated with the true data. Such corruption is often also called "data-dependent" noise. We are given n-length data vectors,

$$\boldsymbol{y}_t := \boldsymbol{\ell}_t + \boldsymbol{w}_t + \boldsymbol{v}_t, t = 1, 2, \dots, \text{ where } \boldsymbol{\ell}_t = \boldsymbol{P}\boldsymbol{a}_t, \ \boldsymbol{w}_t = \boldsymbol{M}_t\boldsymbol{\ell}_t,$$

$\boldsymbol{P}$ 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; and $\boldsymbol{v}_t$ is the uncorrelated noise component with $\mathbb{E}[\boldsymbol{\ell}_t\boldsymbol{v}_t{}'] = 0$. The matrices $\boldsymbol{M}_t$ are *unknown*. We need to estimate $\text{range}(\boldsymbol{P})$. We assume the following about $\boldsymbol{\ell}_t$, $\boldsymbol{M}_t$.

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

Define $\lambda^- := \lambda_{\min}(\boldsymbol{\Lambda})$, $\lambda^+ := \lambda_{\max}(\boldsymbol{\Lambda})$ and $f := \frac{\lambda^+}{\lambda^-}$. Since the $\boldsymbol{a}_t$'s are bounded, we can also define a finite constant $\eta := \max_{j=1,2,\dots r} \max_t \frac{(\boldsymbol{a}_t)_j^2}{\lambda_j}$. Thus, $(\boldsymbol{a}_t)_j^2 \le \eta\lambda_j$.

Asssumption 2. *The data-dependency matrices $\boldsymbol{M}_t$ can be expressed 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,*

$$\|\boldsymbol{M}_{1,t}\boldsymbol{P}\|_2 \le q < 1, \ \|\boldsymbol{M}_{2,t}\|_2 \le 1, \tag{1}$$

and, for any $\alpha \ge \alpha_0$, and any α-length sequence of positive semi-definite Hermitian matrices, $\boldsymbol{A}_t$,

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

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

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, clearly, (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].

Main Result. With the above assumptions, we study the most commonly used PCA solution, *simple eigenvalue decomposition (EVD)* on the empirical covariance matrix of the observed data. Let $\hat{\boldsymbol{P}}$ be the matrix of top r eigenvectors of $\frac{1}{\alpha}\sum_{t=1}^{\alpha}\boldsymbol{y}_t\boldsymbol{y}_t{}'$. We bound the subspace recovery error, $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) := \|(\boldsymbol{I} - \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}')\boldsymbol{P}\|_2$.

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

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

For an $\alpha \ge \alpha_0$, let $\hat{\boldsymbol{P}}$ be as defined above. Assume that Assumptions 1 and 2 hold with α_0 defined above. If $3.3\sqrt{b_0}qf \le \varepsilon_{\text{SE}}/4$ and $3.3\frac{\lambda_{\boldsymbol{v}}^+}{\lambda^-} < \varepsilon_{\text{SE}}/4$, then, with probability at least $1 - 10n^{-10}$,

$$\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P}) \le 6.6(\sqrt{b_0}qf + \frac{\lambda_{\boldsymbol{v}}^+}{\lambda^-}) \le \varepsilon_{\text{SE}}$$

Proof: http://www.ece.iastate.edu/~namrata/ImprovedCorPCA.pdf

Consider the large n regime so that $d = 1$. To compare the effects of correlated and uncorrelated noise, suppose that we equate the time-averaged correlated noise power bound and the uncorrelated noise power bound, and we 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,

1) we need $\alpha_0 = 2500 \cdot 32 \cdot 11\eta^2(\log n)r^2q^2f^2$; and
2) the bound $3.3\sqrt{b_0}qf \le \varepsilon_{\text{SE}}/4$ implies $3.3\frac{\lambda_{\boldsymbol{v}}^+}{\lambda^-} < \varepsilon_{\text{SE}}/4$,

i.e., the bounds due to the correlated noise, $\boldsymbol{w}_t$, dominate. The reason is that the bound on $\text{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$ is governed by the ratio between the norm of the perturbation matrix, $\boldsymbol{D} := \frac{1}{\alpha}\sum_t \boldsymbol{y}_t\boldsymbol{y}_t{}' - \frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{\ell}_t{}'$, and the minimum eigenvalue along the principal subspace, λ^- [2]. The dominant terms in $\boldsymbol{D}$ are $\frac{1}{\alpha}\sum_t \boldsymbol{\ell}_t\boldsymbol{w}_t{}'$ and its transpose.

Discussion. To our best knowledge, most existing results that study the simple EVD solution to PCA assume that the true data and the corrupting noise are uncorrelated. This is valid in practice often, but not always. There is, of course, a large amount of work on robust PCA (PCA in the presence of additive sparse outliers) that assumes nothing about the dependence between the outlier magnitudes and the true data, e.g., [3], [4], [5], [6], [7]. In particular, these allow the outlier magnitudes to be dependent on (correlated with) the true data. However, these works focus on large magnitude sparse outliers and hence need much more expensive solutions than simple EVD. Moreover, these also need the columns of $\boldsymbol{P}$ to be dense (not sparse). We compare simple EVD with two popular robust PCA solutions for a problem involving small magnitude sparse outliers in Table I.

In recent work [1], we studied the correlated-PCA problem described above. Our new result given in Theorem 3 above addresses two important limitations of [1]. First, we generalize the observed data model to also include an uncorrelated noise term. This is a more practically valid noise model. Second, and most importantly, we provide a significantly improved sample complexity bound. Theorem 3 shows that, in the large n regime, the sample complexity, α, is lower bounded by $Cr^2(\log n)\frac{q^2f^2}{\varepsilon_{\text{SE}}^2}$. This is much better than our earlier bound of $Cr^2(\log n)\frac{f^2}{\varepsilon_{\text{SE}}^2}$ [1]. For example, to get the subspace error to below $q/4$, the current bound gives a sample complexity of $\alpha \ge 16Cr^2(\log n)f^2$ samples instead of $\alpha \ge 16Cr^2(\log n)\frac{f^2}{q^2}$.

	Mean Subspace Error (SE)			Execution Time (seconds)		
	EVD	PCP	A-M-RPCA	EVD	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 EVD with two robust PCA solutions - PCP (Principal Components' Pursuit [3]) and A-M-RPCA (Alt-Min-RPCA [6]).* **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 and Alt-Min-RPCA fail. Both have average* $\mathrm{SE}(\hat{\boldsymbol{P}}, \boldsymbol{P})$ *close to one whereas the average SE of c-EVD and EVD is 0.0908 and 0.0911 respectively. Moreover, both of these are much slower than EVD 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 EVD.* **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.*

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] 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.

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

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

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

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

[7] 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.