Principal Component Analysis (PCA) with Data-Dependent Noise: Understanding usefulness of matrix Bernstein and Vershynin’s sub-Gaussian matrices result

Praneeth Narayanamurthy and Namrata Vaswani
Iowa State University

1 Introduction

This note contains a simplified version of one result from our recent ISIT 2018 paper ”PCA in Sparse Data-Dependent Noise” and its meaning as well as its proof. The point of this note is to compare the use of two key random matrix theory results – matrix Bernstein and Vershynin’s independent sub-Gaussian rows result.

If you use this set of notes, please cite above paper.

2 Preliminaries

The key elements that will be used in this document are as follows. Here we prove Theorem 4.4. The overall proof relies on a systematic application of the standard Davis Kahan sin θ theorem [1] summarized here.

**Theorem 2.1** (Davis-Kahan sin θ theorem). Let $D_0$ be a Hermitian matrix whose span of top $r$ eigenvectors equals $\text{span}(P_1)$. Let $D$ be the Hermitian matrix with top $r$ eigenvectors $P_2$. Then,

$$\text{SE}(P_2, P_1) \leq \frac{\| (D - D_0) P_1 \|_2}{\lambda_r(D_0) - \lambda_{r+1}(D)} \leq \frac{\| (D - D_0) P_1 \|_2}{\lambda_r(D_0) - \lambda_{r+1}(D_0) - \lambda_{\text{max}}(D - D_0)} \leq \frac{\| D - D_0 \|_2}{\lambda_r(D_0) - \lambda_{r+1}(D_0) - \| D - D_0 \|}$$

as long as the denominator is positive. The second inequality follows from the first using Weyl’s inequality.

The following theorem is adapted from [2, Theorem 1.6].

**Theorem 2.2** (Matrix Bernstein Concentration). Given an $d$-length sequence of $n_1 \times n_2$ dimensional random matrices. Assume the following holds. (i) the matrices $Z_i$ are mutually independent, (ii) $P(\|Z_i\| \leq R) = 1$, and (iii) max $\{\| \frac{1}{d} \sum \hat{t} E[Z_i'Z_i] \|, \| \frac{1}{d} \sum \hat{t} E[Z_iZ_i'] \| \} \leq \sigma^2$. Then, for an $\epsilon > 0$

$$P\left( \left\| \frac{1}{d} \sum Z_i - \frac{1}{d} \sum \hat{t} E[Z_i] \right\| \leq \epsilon \right) \geq 1 - (n_1 + n_2) \exp \left( \frac{-d \epsilon^2}{2(\sigma^2 + R\epsilon)} \right).$$

The following theorem is adapted from [3, Theorem 5.39].

**Theorem 2.3** (Sub-Gaussian Rows). Given an $d$-length sequence of sub-Gaussian random vectors $w_i$ in $\mathbb{R}^{n_1}$. Assume the following holds. (i) $w_i$ are independent; (ii) the sub-Gaussian norm of $w_i$ is bounded by $K$ for all $i$. Then for an $\epsilon$ satisfy $0 < \epsilon/K^2 < 1$,

$$P\left( \left\| \frac{1}{d} \sum w_i w_i' - \frac{1}{d} \sum \hat{t} E[Z_iZ_i'] \right\| \leq \epsilon \right) \geq 1 - 2 \exp \left( n_1 \log 9 - \frac{c\epsilon^2 d}{4K^2} \right).$$

This theorem can be re-stated by replacing $\epsilon$ by $\tilde{\epsilon} \| \frac{1}{d} \sum \hat{t} E[Z_iZ_i'] \|$ or by $\tilde{\epsilon} \max_i \|E[w_i w_i']\|$. If the $w_i$’s are “nice” sub-Gaussians, then $K^2$ will be a constant times $\max_i \|E[w_i w_i']\|$ and thus, in the latter re-statement, the exponent in the probability expression will simplify to $(n_1 \log 9 - c\tilde{\epsilon}^2)$. 
3 Problem Setup

The goal of this document is to dissect the usage of matrix Bernstein and the matrix Sub-Gaussian Result as applied to a practical problem of Principal Components Analysis (PCA) with missing data/PCA with data-dependent noise. Mathematically, the problem is defined as follows. At each time $t = 1, 2, \cdots, d$, we observe an $n$-dimensional real-valued vector, $y_t$ that satisfies

$$y_t = \ell_t + w_t, \quad \text{with} \quad \ell_t = Pa_t, \quad w_t = M_t \ell_t$$

where $P$ is a “basis matrix” that defines the true, underlying subspace that we want to learn, $a_t$’s denote the principal subspace coefficients, and $w_t$’s denote the data-dependent noise. Furthermore, the matrices $M_t \in \mathbb{R}^{n \times n}$ are deterministic but unknown. We assume that $a_t$’s are either (i) bounded random vectors, or (ii) element-wise bounded random vectors.

**Model on $M_t$.** We assume that $M_t = M_{2,t}M_{1,t}$ and define $q = \max_1 \|M_{1,t}P\|$ and $b = \frac{1}{\delta}\|\sum_1 M_{2,1}M_{2,1}'\|$ with $q, b < 1$. We will provide bounds for these quantities in the main result.

As an application, observe that if $M_t = -I_{T_t}I_{T_t}'$, this translates to PCA with missing data problem and $T_t$ denotes the support of the missing entries at time $t$. Thus, we access the following observations,

$$y_t = Pa_t - I_{T_t}I_{T_t}'Pa_t := P_{\Omega_t}(\ell_t) \quad (5)$$

In the Machine Learning literature, this is also referred to as the Matrix Completion problem, and various approaches to this problem have been analyzed in the past 10+ years [4, 5].

**Statistical Assumptions.** Let $E[a_t] = 0, E[a_t a_t'] = \Lambda$ ($\Lambda$ is diagonal) with $\lambda_{\max}(\Lambda) = \lambda^+, \lambda_{\min}(\Lambda) = \lambda^-$, and $f = \lambda^+ / \lambda^-$. Additionally, assume that $E[e_t e_t'] \neq 0$.

We provide results for two cases. (1) **Bounded:** $a_t$’s are bounded s.t. $\|a_t\|^2_2 \leq \eta_1\lambda^+$ for some numerical constant $\eta_1$; (2) **Element-Wise Bounded:** $a_t$’s are element-wise bounded s.t. $\|a_t\|^2_2 \leq \eta_2\lambda_1(\Lambda)$ for some numerical constant, $\eta_2$. This assumption makes $a_t$’s nice sub-Gaussians as the sub-Gaussian norm is proportional to $\sqrt{\eta_1\lambda^2}$ and thus do not depend on $r$ or $n$.

4 Main Result and Proof

**Theorem 4.4.** Assume that the data satisfies the model described above. Define $H(d) = C_1\sqrt{\eta f}\sqrt{\frac{\log n}{d}}, G_{\text{elem}}(d) = C_2\sqrt{\eta f}\sqrt{\frac{\log n}{d}} + G_{\text{bound}}(d) = C_3\sqrt{\eta f}\sqrt{\frac{\log n}{d}}$. Furthermore, for the element-wise bounded case assume that the data-dependency matrices $M_{1,t}$’s satisfy the assumption with constants $b, q$ which satisfy

$$6\sqrt{bf} + H(d) + G_{\text{elem}}(d) < 1$$

and for the bounded case assume that the $b, q$ satisfy

$$6\sqrt{bf} + H(d) + G_{\text{bound}}(d) < 1$$

Then, with probability at least $1 - 10n^{-10}$, the matrix $P$ of top-$r$ eigenvectors of the sample covariance matrix, $\frac{1}{n} \sum y_t y_t'$ satisfy the following.

1. **Element-Wise bounded:**

$$\text{SE}(P, P) \leq \frac{2\sqrt{bf} + H(d)}{1 - 6\sqrt{bf} - H(d) - G_{\text{elem}}(d)}$$

2. **Bounded:**

$$\text{SE}(P, P) \leq \frac{2\sqrt{bf} + H(d)}{1 - 6\sqrt{bf} - H(d) - G_{\text{bound}}(d)}$$

**Proof of Theorem 4.4.** We will first define matrices in accordance with Theorem 2.1. For this example, we define $D_0 = \frac{1}{n} \sum_t \ell_t \ell_t'$. Notice that this is a Hermitian matrix $P$ as the top $r$ eigenvectors. Next, let $D = \frac{1}{n} \sum y_t y_t'$ and let $P$ denote the matrix of $D$’s top $r$ eigenvectors.
Observe

\[
D - D_0 = \frac{1}{d} \sum_i (y_i y_i' - \ell_i \ell_i') = \frac{1}{d} \sum_i \ell_i w_i' + \frac{1}{d} \sum_i w_i \ell_i' + \frac{1}{d} \sum_i w_i w_i'
\]

:= cross + cross' + noise

Also notice that \( \lambda_{r+1}(D_0) = 0, \lambda_r(D) = \lambda_{\min} \left( \frac{1}{d} \sum_i a_i a_i' \right) \). Now, applying Theorem 2.1,

\[
\text{SE}(\hat{P}, P) \leq \frac{2 ||\text{cross}|| + ||\text{noise}||}{\lambda_{\min} \left( \frac{1}{d} \sum_i a_i a_i' \right) - \text{numerator}}
\]

Now, we can bound \( ||\text{cross}|| \leq ||E[\text{cross}]|| + ||\text{cross} - E[\text{cross}]|| \) and similarly for the noise term. We use the Cauchy-Schwarz inequality for bounding the expected values of cross, noise as follows.

Recall that \( M_t = M_{2,t} M_{1,t} \) with \( b := \| \frac{1}{d} \sum_t M_{2,t} M_{2,t}' \| \) and \( q := \max_t \| M_{1,t} P \| \leq q < 1 \). Thus,

\[
||E[\text{noise}]||^2 = \left\| \frac{1}{d} \sum_t M_t P A P' M_{1,t}' M_{2,t}' \right\|_2^2 \leq \left\| \frac{1}{d} \sum_t M_t P A P' M_{1,t}' M_{1,t} P A P' M_{1,t}' M_{2,t}' \right\|_2 \left\| \frac{1}{d} \sum_t M_{2,t} M_{2,t}' \right\|_2 \\
\leq \max_t ||M_t P A P' M_{1,t}'||_2 b \leq (q^2 \lambda^+)^2 b.
\]

Similarly,

\[
||E[\text{cross}]||^2 = \left\| \frac{1}{d} \sum_t M_{2,t} M_{1,t} P A P' \right\|_2^2 \leq \left\| \frac{1}{d} \sum_t P A P' M_{1,t}' M_{1,t} P A P' \right\|_2 \left\| \frac{1}{d} \sum_t M_{2,t} M_{2,t}' \right\|_2 \\
\leq \max_t ||M_{1,t} P A_t P'||_2^2 b \leq (q \lambda^+)^2 b.
\]

We use the similar idea to lower bound \( \lambda_{\min} \left( \frac{1}{d} \sum_i a_i a_i' \right) \) as

\[
\lambda_{\min} \left( \frac{1}{d} \sum_i a_i a_i' \right) = \lambda_{\min} \left( \Lambda - \left( \frac{1}{d} \sum_i a_i a_i' - \Lambda \right) \right) \\
\geq \lambda_{\min} (\Lambda) - \lambda_{\max} \left( \frac{1}{d} \sum_i a_i a_i' - \Lambda \right) \\
\geq \lambda - \left\| \frac{1}{d} \sum_i a_i a_i' - \Lambda \right\|
\]

and thus we have

\[
\text{SE}(\hat{P}, P) \leq \frac{3 \sqrt{q} f + 2 ||\text{cross} - E[\text{cross}]|| + ||\text{noise} - E[\text{noise}]||}{\lambda^* - ||\frac{1}{d} \sum_i a_i a_i' - \Lambda|| - \text{numerator}}
\]

Bounding the “Statistical Errors”. We use concentration bounds from the Lemma 4.5. Notice that

\[
||\text{noise} - E[\text{noise}]|| + 2 ||\text{cross} - E[\text{cross}]|| = \left\| \frac{1}{d} \sum_i (w_i w_i' - E[w_i w_i']) \right\| + 2 \left\| \frac{1}{d} \sum_i (\ell_i w_i' - E[\ell_i w_i']) \right\| \\
\leq c \sqrt{q} f \sqrt{\frac{r \log n}{d}} \lambda^* + c \sqrt{q} f \sqrt{\frac{r \log n}{d}} \lambda^* \leq C \sqrt{q} f \sqrt{\frac{r \log n}{d}} \lambda^* := H(d) \lambda^*
\]

where the last line follows from using \( q \leq 1 \). The bound on \( ||\frac{1}{d} \sum_i a_i a_i' - \Lambda||_2 \) follows directly from the first item of Lemma 4.5. This completes the proof. \( \square \)
Lemma 4.5. With probability at least $1 - 10n^{-10}$, if $d > r \log n$, then,

$$\left\| \frac{1}{d} \sum_t a_t a_t' - \Lambda \right\| \leq cnf \sqrt{\frac{r + \log n}{d}} \lambda^- := G_{elem}(d) \lambda^-, \quad \text{(element-wise bounded r.v.'s)}$$

$$\left\| \frac{1}{d} \sum_t a_t a_t' - \Lambda \right\| \leq cnf \sqrt{\frac{r \log n}{d}} \lambda^- := G_{bound}(d) \lambda^-, \quad \text{(bounded r.v.'s)}$$

$$\left\| \frac{1}{d} \sum_t \ell_t w_t' - \frac{1}{d} E \left[ \sum_t \ell_t w_t' \right] \right\|_2 \leq c \sqrt{\eta q} \sqrt{\frac{r \log n}{d}} \lambda^- := H(d) \lambda^-, \quad \text{and w.p. at most}$$

$$\left\| \frac{1}{d} \sum_t w_t w_t' - \frac{1}{d} E \left[ \sum_t w_t w_t' \right] \right\|_2 \leq c \sqrt{\eta q^2} \sqrt{\frac{r \log n}{d}} \lambda^- := H(d) q \lambda^-$$

Proof of Lemma 4.5.

1. $a_t a_t'$ term with element-wise bounded-ness

Using sub-Gaussian rows result result applied to $\frac{1}{d} \sum_t a_t a_t'$, and using the fact that the $a_t$'s are $r$-length independent sub-Gaussian vectors with sub-Gaussian norm bounded by $K = \sqrt{\eta \lambda^+}$, we get the following:

$$\text{with probability at least} \ 1 - 2 \exp \left( r \log 9 - d \frac{c^2 \lambda^2}{(4 \eta \lambda^+)^2} \right) = 1 - 2 \exp \left( r \log 9 - d \frac{c^2}{16 \eta q^2} \right),$$

$$\left\| \frac{1}{d} \sum_t a_t a_t' - \Lambda \right\|_2 \leq \epsilon_1 \lambda^-$$

Set $\epsilon_1 = cnf \sqrt{\frac{r + 11 \log n}{d}}$. Then, the above event holds w.p. at least $1 - 2n^{-10}$

2. $a_t a_t'$ term for bounded r.v.'s. This and all other items use Matrix Bernstein for rectangular matrices. Let $\bar{Z}_t := a_t a_t'$ and apply the above result to $\bar{Z}_t = Z_t - E[Z_t]$. with $s = ed$. Now it is easy to see that

$$\|Z_t\| \leq 2 \|a_t a_t'\| \leq 2 \|a_t\|_2^2 \leq 2 \eta r \lambda^+ := R$$

and similarly,

$$\left\| \frac{1}{d} \sum_t E[Z_t^2] \right\| = \frac{1}{d} \sum_t \|E[a_t^2 a_t a_t']\| \leq \max_{a_t} \|a_t\|_2^4 \cdot \max_{a_t} E[a_t a_t'] \leq \eta r (\lambda^+)^2 := \sigma^2$$

and thus, w.p. at most $2r \exp \left( - c \min \left( \frac{c^2 d}{(4 \eta \lambda^+)^2}, \frac{c^2 d}{(4 \eta \lambda^+)^2} \right) \right)$. Now we set $\epsilon = \epsilon_5 \lambda^-$ with $\epsilon_5 = cnf \sqrt{\frac{r \log n}{d}}$ so that

$$\left\| \frac{1}{d} \sum_t (a_t a_t' - E[a_t a_t']) \right\| \geq cnf \sqrt{\frac{r \log n}{d}} \lambda^-$$

3. $\ell_t w_t'$ term.

Let $Z_t := \ell_t w_t'$. We apply this result to $Z_t := Z_t - E[Z_t]$. To get the values of $R$ and $\sigma^2$ in a simple fashion, we use the facts that (i) if $\|Z_t\|_2 \leq R_1$, then $\|Z_t\| \leq 2 R_1$; and (ii) $\sum_t E[Z_t Z_t'] \leq \sum_t E[Z_t Z_t']$. Thus, we can set $R$ to two times the bound on $\|Z_t\|_2$ and similarly for $\sigma^2$

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

$$\left\| \frac{1}{d} \sum_t E[w_t \ell_t \ell_t' w_t'] \right\|_2 \leq \left( \max_{\ell_t} \|\ell_t\|^2 \right) \cdot \max_{\ell_t} \|E[w_t w_t']\| \leq \eta r \lambda^+ \cdot q^2 \lambda^+ = \eta r q^2 (\lambda^+)^2.$$ 

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

Thus, we conclude that,

$$\left\| \frac{1}{d} \sum_t \ell_t w_t' - E[\sum_t \ell_t w_t'] \right\|_2 \geq \epsilon$$

w.p. at most $2n \exp \left( - c \min \left( \frac{c^2 d}{(4 \eta r q^2 \lambda^+)^2}, \frac{c^2 d}{(4 \eta q^2 \lambda^+)^2} \right) \right)$. Set $\epsilon = \epsilon_0 \lambda^-$ with $\epsilon_0 = cnf \sqrt{\frac{r \log n}{d}}$ so that (8) hold w.p. at most $2n^{-10}$.
4. \( \mathbf{w}_t \mathbf{w}_t^t \) term. We again apply matrix Bernstein and proceed as above. In this case, \( R = 2\eta q^2 \lambda^+ + \sigma^2 = \eta q^2 (\lambda^+)^2 \). Set \( \epsilon = c_2 \lambda_{\text{avg}} \) with \( c_2 = c \sqrt{\eta} q^2 f \sqrt{\frac{1}{3} \log n} \). Then again, the probability of the bad event is bounded by \( 2n^{-10} \).

\[ \] 4.1 Vershynin versus Matrix Bern

In Lemma 4.5, if we wanted to use the Vershynin result for the \( \ell_t \mathbf{w}_t^t \) term or the \( \mathbf{w}_t \mathbf{w}_t^t \) term, we would need to first re-write them in terms of \( \sum_t \mathbf{a}_t \mathbf{a}_t^t \). This is possible to do in one of two ways: (i) use Cauchy-Schwarz with \( \mathbf{P} \mathbf{a}_t \mathbf{a}_t^t \) as the first matrix; or (ii) use a bound on \( \max_t \| \mathbf{P}^t \mathbf{M}_1^t \mathbf{M}_2^t \| \). In (i), we will end up squaring \( \mathbf{a}_t \mathbf{a}_t^t \) which means we will need to bound \( \max_t \| \mathbf{a}_t \| \) which is bounded by \( r \log n \). In case of (ii), we will not be able to exploit the assumption on \( \sum_t \mathbf{M}_2^t \mathbf{M}_2^t \). The latter is what helps us get a nice bound on \( \mathbb{E} [ \ell_t \mathbf{w}_t^t ] \).

For the \( \mathbf{a}_t \mathbf{a}_t^t \) term, we can use either result. Vershynin will give a better bound. But since the overall sample complexity is dictated by the other terms, it does not matter.

5 Discussion

5.1 Discussion of assumptions

Notice that \( \lambda^+ \) is the (maximum) signal power. The model on \( \mathbf{M}_1^t \) implies that the noise power \( \| \mathbb{E} [ \mathbf{w}_t \mathbf{w}_t^t ] \|_2 \leq q^2 \lambda^+ \) (thus \( q^2 \) is the noise-to-signal ratio) and the signal-noise correlation \( \| \mathbb{E} [ \ell_t \mathbf{w}_t^t ] \|_2 \leq q \lambda^+ \). Without the assumption on \( \mathbf{M}_2^t \), (i.e. if \( b = 1 \)), this implies that it is not possible to achieve subspace error that is anything smaller than a constant times \( qf \). The reason is that PCA error depends on the ratio between noise power plus signal-noise correlation and \( \lambda^- \) (minimum signal space eigenvalue). However, with the assumption on \( \mathbf{M}_2^t \), one can show that the time-averaged values of both the above quantities satisfy

\[ \frac{1}{d} \sum_{t=1}^d \mathbb{E} [ \mathbf{w}_t \mathbf{w}_t^t ] \|_2 \leq \sqrt{b} q^2 \lambda^+ \text{ and } \frac{1}{d} \sum_{t=1}^d \mathbb{E} [ \ell_t \mathbf{w}_t^t ] \|_2 \leq \sqrt{b} q \lambda^+ \].

Without the assumption on \( \mathbf{M}_2^t \), our result tells us that the noise support changes enough over time so that \( b \) is small; (iii) \( d \geq C f^2 \log n \). This sample complexity is near-optimal since \( r \) is the minimum number of samples needed to even define a subspace.

5.2 Discussion of Result

Observe that to obtain an error of \( \varepsilon \) it suffices to ensure that \( H(d) \leq 0.1 \varepsilon \) and this is satified as long as

\[ d \geq C \frac{q^2 f^2}{\varepsilon^2} r \log n \]

Furthermore, notice that in both models, we only require that \( G_{\text{elem}}(d) (G_{\text{bound}}(d)) \) is less than a constant, \( c = 0.01 \). Finally, to understand the difference in the result of Matrix Bernstein vs the sub-Gaussian row result, consider the following setting. If we only want to estimate the principal components of a large fraction of the noise level, i.e., if we only wanted \( \varepsilon = 2qf \), and not to make it arbitrarily small, the sample complexity is dominated by the \( G_{\text{elem}}(d) (G_{\text{bound}}(d)) \) term, and thus, in the element-wise bounded setting, it suffices to have \( d \geq C f^2 (r + \log n) \) samples, whereas, in the bounded setting, we would still need \( d \geq C f^2 r \log n \).

References


\[1\] follows with a careful application of Cauchy-Schwartz inequality.