Supplementary Material:
Memory and Computation Efficient PCA via Very Sparse Random Projections

Farhad Pourkamali-Anaraki FARHAD.POURKAMALI@COLORADO.EDU
Shannon M. Hughes SHANNON.HUGHES@COLORADO.EDU
Department of Electrical, Computer, and Energy Engineering, University of Colorado at Boulder, Boulder, CO 80309

1. Introduction

In this supplementary material, we will present the proofs of the main theorems given in Section 5 of our paper. Following the Proof of Theorem 3, we also include some discussion of the 2-norm of the error term $\mathbf{E}$ that appears in it. In particular, we give bounds on the norm of $\mathbf{E}$ for various types of PCs (smooth, sparse, etc.).

We start with stating the following two Lemmas. Throughout, we will follow the notation given in Section 3 of our paper.

Lemma 1. *Let $\mathbf{R} \in \mathbb{R}^{p\times m}$ be a random matrix consisting of i.i.d. entries with zero mean $\mu_1 = 0$, finite second and fourth moments μ_2 and μ_4, and kurtosis $\kappa \triangleq \frac{\mu_4}{\mu_2^2} - 3$. We define $\mathbf{t}_k$ as the k^{th} column of the matrix $\mathbf{R}\mathbf{R}^T$. Let $\{\mathbf{e}_i\}_{i=1}^p \in \mathbb{R}^p$ denote the standard basis. If we define the matrix $\Lambda_{k,l} \triangleq \mathbb{E}[\mathbf{t}_k\mathbf{t}_l^T]$, $1 \le k, l \le p$, then*

$$\begin{aligned}\Lambda_{k,k} &= m\mu_2^2\mathbf{I}_{p\times p} + m\mu_2^2\left(\kappa + m + 1\right)\mathbf{e}_k\mathbf{e}_k^T \\ \Lambda_{k,l} &= m^2\mu_2^2\mathbf{e}_k\mathbf{e}_l^T + m\mu_2^2\mathbf{e}_l\mathbf{e}_k^T, \quad k \ne l. \end{aligned} \tag{1.1}$$

Moreover, we have that

$$\mathbb{E}\left[\mathbf{R}\mathbf{R}^T\mathbf{R}\mathbf{R}^T\right] = m\mu_2^2\left(\kappa + m + p + 1\right)\mathbf{I}_{p\times p}. \tag{1.2}$$

Proof. Let us consider the matrix $\mathbf{R} \in \mathbb{R}^{p\times m}$, where the entry in the i^{th} row and j^{th} column is denoted by r_{ij}. It is easy to verify that the k^{th} column of the matrix $\mathbf{R}\mathbf{R}^T$ has the following form

$$\mathbf{t}_k = \left[\begin{array}{ccccccc} \sum_{j=1}^m r_{1j}r_{kj} & \sum_{j=1}^m r_{2j}r_{kj} & \cdots & \sum_{j=1}^m r_{kj}^2 & \cdots & \sum_{j=1}^m r_{pj}r_{kj} \end{array}\right]^T.$$

We calculate the entries of the matrix $\Lambda_{k,k} = \mathbb{E}[\mathbf{t}_k\mathbf{t}_k^T]$ using the assumption that $\mu_1 = 0$.

(1) The k^{th} diagonal entry of $\Lambda_{k,k}$:

$$\mathbb{E}\left[\left(\sum_{j=1}^m r_{kj}^2\right)^2\right] = \mathbb{E}\left[\sum_{j=1}^m r_{kj}^4 + \sum_{i\ne j} r_{ki}^2 r_{kj}^2\right] = m\mu_4 + \left(m^2 - m\right)\mu_2^2.$$

(2) The other diagonal entries of $\Lambda_{k,k}$ (i.e. the i^{th} diagonal entry for $i \ne k$):

$$\mathbb{E}\left[\left(\sum_{j=1}^m r_{ij}r_{kj}\right)^2\right] = \mathbb{E}\left[\sum_{j=1}^m r_{ij}^2 r_{kj}^2\right] = m\mu_2^2.$$

(3) The off-diagonal entries of $\Lambda_{k,k}$ (i.e. the entry in the i^{th} row and the l^{th} column when $i \ne l$):

$$\mathbb{E}\left[\left(\sum_{j=1}^m r_{ij}r_{kj}\right)\left(\sum_{j=1}^m r_{lj}r_{kj}\right)\right] = 0.$$

Observe that these entries have at least one term with degree 1, even for the case that either $i = k$ or $l = k$ ($i \ne l$). Thus, they all equal zero.

In summary, we get:

$$\begin{aligned}\Lambda_{k,k} &= m\mu_2^2\mathbf{I}_{p\times p} + \left(m\mu_4 + \left(m^2 - 2m\right)\mu_2^2\right)\mathbf{e}_k\mathbf{e}_k^T \\ &= m\mu_2^2\mathbf{I}_{p\times p} + m\mu_2^2\left(\frac{\mu_4}{\mu_2^2} - 3 + 1 + m\right)\mathbf{e}_k\mathbf{e}_k^T \\ &= m\mu_2^2\mathbf{I}_{p\times p} + m\mu_2^2\left(\kappa + m + 1\right)\mathbf{e}_k\mathbf{e}_k^T. \end{aligned} \tag{1.3}$$

Similarly, we can calculate the entries of the matrix $\Lambda_{k,l} = \mathbb{E}[\mathbf{t}_k \mathbf{t}_l^T]$. In this case, each entry has at least one term with degree 1, so it will be zero, except two terms:

(1) The entry in the k^{th} row and l^{th} column of $\Lambda_{k,l}$:

$$\mathbb{E}\left[\left(\sum_{j=1}^{m} r_{kj}^2\right)\left(\sum_{j=1}^{m} r_{lj}^2\right)\right] = \mathbb{E}\left[\sum_{j=1}^{m} r_{kj}^2\right]\mathbb{E}\left[\sum_{j=1}^{m} r_{lj}^2\right] = m^2\mu_2^2.$$

(2) The entry in the l^{th} row and k^{th} column of $\Lambda_{k,l}$:

$$\mathbb{E}\left[\left(\sum_{j=1}^{m} r_{lj} r_{kj}\right)^2\right] = \mathbb{E}\left[\sum_{j=1}^{m} r_{lj}^2 r_{kj}^2\right] = m\mu_2^2.$$

Hence, we get:

$$\Lambda_{k,l} = m^2\mu_2^2 \mathbf{e}_k \mathbf{e}_l^T + m\mu_2^2 \mathbf{e}_l \mathbf{e}_k^T. \tag{1.4}$$

Moreover, note that $\mathbf{R}\mathbf{R}^T \in \mathbb{R}^{p\times p}$ is a symmetric matrix, and we have

$$\mathbf{R}\mathbf{R}^T\mathbf{R}\mathbf{R}^T = \begin{bmatrix} - & \mathbf{t}_1^T & - \\ & \vdots & \\ - & \mathbf{t}_p^T & - \end{bmatrix}\begin{bmatrix} | & & | \\ \mathbf{t}_1 & \dots & \mathbf{t}_p \\ | & & | \end{bmatrix} = \begin{bmatrix} \langle \mathbf{t}_1, \mathbf{t}_1\rangle & \cdots & \langle \mathbf{t}_1, \mathbf{t}_p\rangle \\ \vdots & \ddots & \\ \langle \mathbf{t}_p, \mathbf{t}_1\rangle & & \langle \mathbf{t}_p, \mathbf{t}_p\rangle \end{bmatrix}. \tag{1.5}$$

We further note that we can obtain the expectation of each term using the trace property:

$$\begin{aligned} \mathbb{E}\left[\langle \mathbf{t}_k, \mathbf{t}_k\rangle\right] &= \mathbb{E}\left[tr\left(\mathbf{t}_k^T\mathbf{t}_k\right)\right] \\ &= tr\left(\mathbb{E}\left[\mathbf{t}_k\mathbf{t}_k^T\right]\right) = tr\left(\Lambda_{k,k}\right) = m\mu_2^2\left(\kappa + m + p + 1\right) \end{aligned}$$

and

$$\mathbb{E}\left[\langle \mathbf{t}_k, \mathbf{t}_l\rangle\right] = \mathbb{E}\left[tr\left(\mathbf{t}_k^T\mathbf{t}_l\right)\right] = tr\left(\Lambda_{l,k}\right) = 0, \quad k \neq l.$$

Hence,

$$\mathbb{E}\left[\mathbf{R}\mathbf{R}^T\mathbf{R}\mathbf{R}^T\right] = m\mu_2^2\left(\kappa + m + p + 1\right)\mathbf{I}_{p\times p}. \tag{1.6}$$

□

Lemma 2. *Assume that $\mathbf{R} \in \mathbb{R}^{p\times m}$ is a random matrix consisting of i.i.d. entries drawn from a distribution with zero mean and bounded moments μ_2, μ_4, μ_6, and μ_8. Also, let $\mathbf{x} \in \mathbb{R}^p$ be a fixed vector. Then*

$$\frac{1}{m^2\left(m+1\right)^2\mu_2^4}\mathbb{E}\left[\left\|\mathbf{R}\mathbf{R}^T\mathbf{x}\right\|_2^4\right] = \xi_1\sum_{i=1}^{p} x_i^4 + \xi_2\sum_{i\neq j} x_i^2 x_j^2$$

where

$$\xi_1 \leq \frac{\mu_8/\mu_2^4}{m^3} + \frac{\mu_6/\mu_2^3}{m^2}\left(4 + \frac{2}{m/p}\right) + \frac{\left(\mu_4/\mu_2^2\right)^2}{m^2}\left(3 + \frac{1}{m/p}\right) + \frac{\mu_4/\mu_2^2}{m}\left(6 + \frac{6}{m/p} + \frac{1}{\left(m/p\right)^2}\right) + \left(1 + \frac{1}{m/p}\right)^2 + \frac{3}{p\left(m/p\right)^2}$$

and

$$\xi_2 \leq \frac{\mu_6/\mu_2^3}{m^2}\left(\frac{6}{p^{m/p}}\right)+\frac{\left(\mu_4/\mu_2^2\right)^2}{m^2}\left(1+\frac{5}{p^{m/p}}\right)+\frac{\mu_4/\mu_2^2}{m}\left(2+\frac{26}{p^{m/p}}+\frac{2}{m/p}+\frac{13}{p\left(m/p\right)^2}\right)+\left(1+\frac{1}{m/p}\right)^2$$
$$+\frac{10}{p^{m/p}}+\frac{7}{p^2\left(m/p\right)^2}+\frac{13}{p\left(m/p\right)^2}+\frac{2}{p\left(m/p\right)^3}.$$

Proof. This is a straightforward, albeit tedious, computation. To start, let $\mathbf{x}_j$ denote the j^{th} entry of $\mathbf{x}$ so that $\mathbf{x} = [x_1, \ldots, x_p]^T$, let r_{ij} denote the entry in the i^{th} row and j^{th} column of $\mathbf{R}$, and let $\mathbf{t}_k$ denote the k^{th} column of the matrix $\mathbf{R}\mathbf{R}^T$. Then:

$$\begin{aligned}
\left\|\mathbf{R}\mathbf{R}^T\mathbf{x}\right\|_2^4 &= \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p} x_i x_j x_k x_l \langle \mathbf{t}_i, \mathbf{t}_j\rangle\langle \mathbf{t}_k, \mathbf{t}_l\rangle \\
&= \sum_{i=1}^{p} x_i^4 \langle \mathbf{t}_i, \mathbf{t}_i\rangle^2 + 4\sum_{i=1}^{p}\sum_{j\neq i} x_i^3 x_j \langle \mathbf{t}_i, \mathbf{t}_i\rangle\langle \mathbf{t}_i, \mathbf{t}_j\rangle + \sum_{i=1}^{p}\sum_{j\neq i} x_i^2 x_j^2 \langle \mathbf{t}_i, \mathbf{t}_i\rangle\langle \mathbf{t}_j, \mathbf{t}_j\rangle + 2\sum_{i=1}^{p}\sum_{j\neq i} x_i^2 x_j^2 \langle \mathbf{t}_i, \mathbf{t}_j\rangle^2 \\
&\quad +2\sum_{i=1}^{p}\sum_{j\neq i}\sum_{k\neq i,j} x_i^2 x_j x_k \langle \mathbf{t}_i, \mathbf{t}_i\rangle\langle \mathbf{t}_j, \mathbf{t}_k\rangle + 4\sum_{i=1}^{p}\sum_{j\neq i}\sum_{k\neq i,j} x_i^2 x_j x_k \langle \mathbf{t}_i, \mathbf{t}_j\rangle\langle \mathbf{t}_i, \mathbf{t}_k\rangle \\
&\quad +\sum_{i=1}^{p}\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k} x_i x_j x_k x_l \langle \mathbf{t}_i, \mathbf{t}_j\rangle\langle \mathbf{t}_k, \mathbf{t}_l\rangle. \qquad (1.7)
\end{aligned}$$

Using the assumption that $\mu_1 = 0$, we see that the following terms are zero:

$$\begin{aligned}
\mathbb{E}\left[\langle \mathbf{t}_i, \mathbf{t}_i\rangle\langle \mathbf{t}_i, \mathbf{t}_j\rangle\right] &= 0, \quad i \neq j \\
\mathbb{E}\left[\langle \mathbf{t}_i, \mathbf{t}_i\rangle\langle \mathbf{t}_j, \mathbf{t}_k\rangle\right] &= 0, \quad i \neq j \neq k \\
\mathbb{E}\left[\langle \mathbf{t}_i, \mathbf{t}_j\rangle\langle \mathbf{t}_i, \mathbf{t}_k\rangle\right] &= 0, \quad i \neq j \neq k \\
\mathbb{E}\left[\langle \mathbf{t}_i, \mathbf{t}_j\rangle\langle \mathbf{t}_k, \mathbf{t}_l\rangle\right] &= 0, \quad i \neq j \neq k \neq l. \qquad (1.8)
\end{aligned}$$

We show this only for one of the above terms to save space. Let us consider the first expression $\mathbb{E}[\langle \mathbf{t}_i, \mathbf{t}_i\rangle\langle \mathbf{t}_i, \mathbf{t}_j\rangle]$. For simplicity and without loss of generality, we assume that $i = 1$ and $j = 2$. However, it is obvious that the proof holds for any fixed $i \neq j$.

$$\begin{aligned}
\mathbb{E}\left[\langle \mathbf{t}_1, \mathbf{t}_1\rangle\langle \mathbf{t}_1, \mathbf{t}_2\rangle\right] &= \mathbb{E}\left[\sum_{l=1}^{p}\left(\mathbf{t}_1\right)_l^2\sum_{k=1}^{p}\left(\mathbf{t}_1\right)_k\left(\mathbf{t}_2\right)_k\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{p}\sum_{l=1}^{p}\left(\sum_{j=1}^{m} r_{lj}r_{1j}\right)^2\left(\sum_{j=1}^{m} r_{kj}r_{2j}\right)\left(\sum_{j=1}^{m} r_{kj}r_{1j}\right)\right] \\
&= \sum_{k=1}^{p}\mathbb{E}\left[\left(\sum_{j=1}^{m} r_{kj}r_{2j}\right)\left(\sum_{j=1}^{m} r_{kj}r_{1j}\right)^3\right] \\
&\quad +\sum_{k=1}^{p}\sum_{l\neq k}\mathbb{E}\left[\left(\sum_{j=1}^{m} r_{kj}r_{2j}\right)\left(\sum_{j=1}^{m} r_{kj}r_{1j}\right)\left(\sum_{j=1}^{m} r_{lj}r_{1j}\right)^2\right]. \qquad (1.9)
\end{aligned}$$

It is easy to verify that all the terms in Eq. 1.9 contain at least one odd degree of the random variables forming the random matrix $\mathbf{R}$. Let's examine this for the first summation in Eq. 1.9:

$$\begin{aligned}
&\sum_{k=1}^{p}\mathbb{E}\left[\left(\sum_{j=1}^{m}r_{kj}r_{2j}\right)\left(\sum_{j=1}^{m}r_{kj}r_{1j}\right)^3\right] \\
&=\mathbb{E}\left[\left(\sum_{j=1}^{m}r_{1j}r_{2j}\right)\left(\sum_{j=1}^{m}r_{1j}^2\right)^3\right]+\mathbb{E}\left[\left(\sum_{j=1}^{m}r_{2j}^2\right)\left(\sum_{j=1}^{m}r_{2j}r_{1j}\right)^3\right] \\
&+\sum_{k\geq 3}\mathbb{E}\left[\left(\sum_{j=1}^{m}r_{kj}r_{2j}\right)\left(\sum_{j=1}^{m}r_{kj}r_{1j}\right)^3\right].
\end{aligned}$$

For example, the third term in the above equation will have r_{2j} in all of its expansions, hence the expectation of all those terms equal zero. Using the similar argument for other terms in Eq. 1.9, we see that $\mathbb{E}[\langle\mathbf{t}_i,\mathbf{t}_i\rangle\langle\mathbf{t}_i,\mathbf{t}_j\rangle]=0$ for $i\neq j$.

Next, we see that:

$$\begin{aligned}
\mathbb{E}\left[\langle\mathbf{t}_i,\mathbf{t}_i\rangle^2\right] &= m\mu_8+\mu_6\mu_2\left\{4m\left(m-1\right)+2m\left(p-1\right)\right\}+\mu_4^2\left\{3m\left(m-1\right)+m\left(p-1\right)\right\} \\
&+\mu_4\mu_2^2\left\{6m\left(m-1\right)\left(m-2\right)+6m\left(m-1\right)\left(p-1\right)+m\left(p-1\right)\left(p-2\right)\right\} \\
&+\mu_2^4\left\{m\left(m-1\right)\left(m-2\right)\left(m-3\right)+3m\left(m-1\right)\left(p-1\right)\right. \\
&\left.+2m\left(m-1\right)\left(m-2\right)\left(p-1\right)+m\left(m-1\right)\left(p-1\right)\left(p-2\right)\right\}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\langle\mathbf{t}_i,\mathbf{t}_i\rangle\langle\mathbf{t}_j,\mathbf{t}_j\rangle\right] &= 2m\mu_6\mu_2+\mu_4^2\left\{m^2+m\right\}+\mu_4\mu_2^2\left\{6m\left(m-1\right)+2m^2\left(m-1\right)+3m\left(p-2\right)+2m^2\left(p-2\right)\right\} \\
&+\mu_2^4\left\{2m\left(m-1\right)\left(m-2\right)+3m\left(m-1\right)+m^2\left(m-1\right)^2+3m\left(m-1\right)\left(p-2\right)\right. \\
&\left.+2m^2\left(m-1\right)\left(p-2\right)+m^2\left(p-2\right)\left(p-3\right)\right\}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\langle\mathbf{t}_i,\mathbf{t}_j\rangle^2\right] &= 2m\mu_6\mu_2+2m\mu_4^2+\mu_4\mu_2^2\left\{10m\left(m-1\right)+5m\left(p-2\right)\right\} \\
&+\mu_2^4\left\{4m\left(m-1\right)\left(m-2\right)+5m\left(m-1\right)\left(p-2\right)+m\left(p-2\right)\left(p-3\right)+2m\left(m-1\right)\right\}. \quad (1.10)
\end{aligned}$$

Again, we show this only for one of the above terms. In the following, we compute $\mathbb{E}[\langle\mathbf{t}_i,\mathbf{t}_i\rangle^2]$ and for simplicity and without loss of generality, we fix $i=1$.

$$
\begin{aligned}
\mathbb{E}\left[\langle \mathbf{t}_1, \mathbf{t}_1\rangle^2\right] &= \mathbb{E}\left[tr\left(\mathbf{t}_1^T\mathbf{t}_1\mathbf{t}_1^T\mathbf{t}_1\right)\right] = \mathbb{E}\left[tr\left(\mathbf{t}_1\mathbf{t}_1^T\mathbf{t}_1\mathbf{t}_1^T\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{j=1}^m r_{1j}^2\right)^4\right] + \sum_{k=2}^p \mathbb{E}\left[\left(\sum_{j=1}^m r_{kj}r_{1j}\right)^4\right] + 2\sum_{k=2}^p \mathbb{E}\left[\left(\sum_{j=1}^m r_{1j}^2\right)^2\left(\sum_{j=1}^m r_{kj}r_{1j}\right)^2\right] \\
&\quad + \sum_{k\neq l>1} \mathbb{E}\left[\left(\sum_{j=1}^m r_{kj}r_{1j}\right)^2\left(\sum_{j=1}^m r_{lj}r_{1j}\right)^2\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\left(\sum_{j=1}^m r_{1j}^2\right)^4\right] + (p-1)\,\mathbb{E}\left[\left(\sum_{j=1}^m r_{2j}r_{1j}\right)^4\right] + 2\,(p-1)\,\mathbb{E}\left[\left(\sum_{j=1}^m r_{1j}^2\right)^2\left(\sum_{j=1}^m r_{2j}r_{1j}\right)^2\right] \\
&\quad + (p-1)\,(p-2)\,\mathbb{E}\left[\left(\sum_{j=1}^m r_{2j}r_{1j}\right)^2\left(\sum_{j=1}^m r_{3j}r_{1j}\right)^2\right]
\end{aligned}
$$

where (b) follows from the fact that

$$
tr\left(\mathbf{t}_1\mathbf{t}_1^T\mathbf{t}_1\mathbf{t}_1^T\right) = \left\|\mathbf{t}_1\mathbf{t}_1^T\right\|_F^2
$$

and

$$
\left(\mathbf{t}_1\mathbf{t}_1^T\right)_{k,l} = \left(\sum_{j=1}^m r_{kj}r_{1j}\right)\left(\sum_{j=1}^m r_{lj}r_{1j}\right).
$$

Using (1.7), (1.8), and (1.10), we get

$$
\frac{1}{m^2\,(m+1)^2\,\mu_2^4}\mathbb{E}\left[\left\|\mathbf{R}\mathbf{R}^T\mathbf{x}\right\|_2^4\right] = \xi_1\sum_{i=1}^p x_i^4 + \xi_2\sum_{i\neq j} x_i^2x_j^2
$$

where

$$
\begin{aligned}
\xi_1 &= \frac{1}{m^2\,(m+1)^2\,\mu_2^4}\mathbb{E}\left[\langle \mathbf{t}_i, \mathbf{t}_i\rangle^2\right] \\
&= \frac{1}{m\,(m+1)^2\,\mu_2^4}\left\{\mu_8 + \mu_6\mu_2\,(4\,(m-1) + 2\,(p-1)) + \mu_4^2\,(3\,(m-1) + (p-1))\right. \\
&\quad +\mu_4\mu_2^2\,(6\,(m-1)\,(m-2) + 6\,(m-1)\,(p-1) + (p-1)\,(p-2)) \\
&\quad \left.+\mu_2^4\,((m-1)\,(m-2)\,(m-3) + 3\,(m-1)\,(p-1) + 2\,(m-1)\,(m-2)\,(p-1) + (m-1)\,(p-1)\,(p-2))\right\} \\
&\leq \frac{\mu_8/\mu_2^4}{m^3} + \frac{\mu_6/\mu_2^3}{m^2}\left(4 + \frac{2}{m/p}\right) + \frac{\left(\mu_4/\mu_2^2\right)^2}{m^2}\left(3 + \frac{1}{m/p}\right) + \frac{\mu_4/\mu_2^2}{m}\left(6 + \frac{6}{m/p} + \frac{1}{(m/p)^2}\right) + \left(1 + \frac{1}{m/p}\right)^2 + \frac{3}{p\,(m/p)^2}
\end{aligned}
$$

and

$$\begin{aligned}\xi_2 &= \frac{1}{m^2\left(m+1\right)^2\mu_2^4}\left\{\mathbb{E}\left[\langle\mathbf{t}_i,\mathbf{t}_i\rangle\langle\mathbf{t}_j,\mathbf{t}_j\rangle\right]+2\mathbb{E}\left[\langle\mathbf{t}_i,\mathbf{t}_j\rangle^2\right]\right\}\\ &\leq \frac{\mu_6/\mu_2^3}{m^2}\left(\frac{6}{p^{m}/_{p}}\right)+\frac{\left(\mu_4/\mu_2^2\right)^2}{m^2}\left(1+\frac{5}{p^{m}/_{p}}\right)+\frac{\mu_4/\mu_2^2}{m}\left(2+\frac{26}{p^{m}/_{p}}+\frac{2}{{}^{m}/_{p}}+\frac{13}{p\left({}^{m}/_{p}\right)^2}\right)+\left(1+\frac{1}{{}^{m}/_{p}}\right)^2\\ &\quad+\frac{10}{p^{m}/_{p}}+\frac{7}{p^2\left({}^{m}/_{p}\right)^2}+\frac{13}{p\left({}^{m}/_{p}\right)^2}+\frac{2}{p\left({}^{m}/_{p}\right)^3}.\end{aligned}$$

This completes the proof. □

2. Proof of Theorem 1

Since w_j, $\mathbf{z}$, and $\mathbf{R}$ are independent with $\mathbb{E}[w_j]=0$ and $\mathbb{E}[\mathbf{z}]=\mathbf{0}$, we have $\mathbb{E}[\mathbf{R}\mathbf{R}^T\mathbf{x}]=\mathbb{E}[\mathbf{R}\mathbf{R}^T]\overline{\mathbf{x}}$. The diagonal entries of the matrix $\mathbf{R}\mathbf{R}^T$ have the form $\sum_{j=1}^{m} r_{kj}^2$, $1\leq k\leq p$, where $r_{i,j}$ denotes the i,j^{th} entry of the matrix $\mathbf{R}$, and the off-diagonal entries have the form $\sum_{j=1}^{m} r_{kj}r_{lj}$, $k\neq l$. The entries of $\mathbf{R}$ are i.i.d. with zero mean and variance μ_2, thus $\mathbb{E}[\mathbf{R}\mathbf{R}^T]=m\mu_2\mathbf{I}_{p\times p}$, and we get $\frac{1}{m\mu_2}\mathbb{E}\left[\mathbf{R}\mathbf{R}^T\mathbf{x}\right]=\overline{\mathbf{x}}$. Theorem 1 then follows from linearity and the strong law of large numbers. □

3. Proof of Theorem 2

Consider our proposed center estimator $\widehat{\overline{\mathbf{x}}}_n=\frac{1}{m\mu_2}\frac{1}{n}\sum_{i=1}^{n}\mathbf{R}_i\mathbf{R}_i^T\mathbf{x}_i$. We showed in Theorem 1 that this is an unbiased estimator for the true center, i.e. $\mathbb{E}[\widehat{\overline{\mathbf{x}}}_n]=\overline{\mathbf{x}}$. Now, we find the variance:

$$\begin{aligned}\mathrm{Var}\left(\widehat{\overline{\mathbf{x}}}_n\right) &= \mathbb{E}\left[\left\|\widehat{\overline{\mathbf{x}}}_n-\overline{\mathbf{x}}\right\|_2^2\right]\\ &= \frac{1}{n^2}tr\left(\mathbb{E}\left[\sum_{i=1}^{n}\left(\frac{1}{m\mu_2}\mathbf{R}_i\mathbf{R}_i^T\mathbf{x}_i-\overline{\mathbf{x}}\right)\sum_{i=1}^{n}\left(\frac{1}{m\mu_2}\mathbf{R}_i\mathbf{R}_i^T\mathbf{x}_i-\overline{\mathbf{x}}\right)^T\right]\right)\\ &\overset{(a)}{=} \frac{1}{n^2}tr\left(\mathbb{E}\left[\sum_{i=1}^{n}\left(\frac{1}{m\mu_2}\mathbf{R}_i\mathbf{R}_i^T\mathbf{x}_i-\overline{\mathbf{x}}\right)\left(\frac{1}{m\mu_2}\mathbf{R}_i\mathbf{R}_i^T\mathbf{x}_i-\overline{\mathbf{x}}\right)^T\right]\right)\\ &= \frac{1}{n}tr\left(\mathbb{E}\left[\left(\frac{1}{m\mu_2}\mathbf{R}\mathbf{R}^T\mathbf{x}-\overline{\mathbf{x}}\right)\left(\frac{1}{m\mu_2}\mathbf{R}\mathbf{R}^T\mathbf{x}-\overline{\mathbf{x}}\right)^T\right]\right)\\ &= \frac{1}{n}\frac{1}{m^2\mu_2^2}tr\left(\mathbb{E}\left[\mathbf{R}\mathbf{R}^T\left(\mathbf{x}-\overline{\mathbf{x}}\right)\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{R}\mathbf{R}^T\right]\right)\\ &\quad+\frac{2}{n}tr\left(\mathbb{E}\left[\frac{1}{m\mu_2}\mathbf{R}\mathbf{R}^T\left(\mathbf{x}-\overline{\mathbf{x}}\right)\left(\frac{1}{m\mu_2}\mathbf{R}\mathbf{R}^T\overline{\mathbf{x}}-\overline{\mathbf{x}}\right)^T\right]\right)\\ &\quad+\frac{1}{n}tr\left(\mathbb{E}\left[\left(\frac{1}{m\mu_2}\mathbf{R}\mathbf{R}^T\overline{\mathbf{x}}-\overline{\mathbf{x}}\right)\left(\frac{1}{m\mu_2}\mathbf{R}\mathbf{R}^T\overline{\mathbf{x}}-\overline{\mathbf{x}}\right)^T\right]\right)\end{aligned}\tag{3.1}$$

where (a) follows from the fact that the vectors $(\frac{1}{m\mu_2}\mathbf{R}_i\mathbf{R}_i^T\mathbf{x}_i-\overline{\mathbf{x}})$ are independent and identically distributed around $\mathbf{0}$ for each i. Also, note that we have dropped the i subscript in the last two steps, since all $\mathbf{R}_i$ and $\mathbf{x}_i$ are i.i.d. and the dependence on i no longer matters.

Now, the second term in Eq. 3.1 is zero because $\mathbb{E}[\mathbf{x}-\overline{\mathbf{x}}]=\mathbf{0}$. Thus, the variance consists of two terms. The first term in Eq. 3.1 is computed using the result of Theorem 3:

$$
\begin{aligned}
& tr\left(\mathbb{E}\left[\mathbf{R}\mathbf{R}^T\left(\mathbf{x}-\overline{\mathbf{x}}\right)\left(\mathbf{x}-\overline{\mathbf{x}}\right)^T\mathbf{R}\mathbf{R}^T\right]\right) \\
&= \left(m^2+m\right)\mu_2^2\left(tr\left(\mathbf{C}_{true}+\alpha\mathbf{I}\right)+\frac{\kappa}{m+1}\sum_{i=1}^{p}\sum_{j=1}^{d}\sigma_j^2 v_{j,i}^2\right) \\
&\overset{(b)}{=} \left(m^2+m\right)\mu_2^2\left(h+\alpha p+\frac{\kappa}{m+1}h\right) \\
&= m\left(m+1\right)\mu_2^2\left(h+\frac{ph}{m+1}+\frac{\left(\kappa+m+p+1\right)}{m+1}\epsilon^2+\frac{\kappa}{m+1}h\right) \\
&= m\mu_2^2\left(\kappa+m+p+1\right)\left(h+\epsilon^2\right) \\
&= m\mu_2^2 h\left(\kappa+m+p+1\right)\left(1+\frac{1}{\text{SNR}}\right)
\end{aligned} \tag{3.2}
$$

where (b) follows from $\sum_{i=1}^{p}\sum_{j=1}^{d}\sigma_j^2 v_{j,i}^2 = \sum_{j=1}^{d}\sigma_j^2\sum_{i=1}^{p}v_{j,i}^2 = h$, since each PC has unit norm. Next, we find the third term in Eq. 3.1 as follows:

$$
\begin{aligned}
& tr\left(\mathbb{E}\left[\left(\frac{1}{m\mu_2}\mathbf{R}\mathbf{R}^T\overline{\mathbf{x}}-\overline{\mathbf{x}}\right)\left(\frac{1}{m\mu_2}\mathbf{R}\mathbf{R}^T\overline{\mathbf{x}}-\overline{\mathbf{x}}\right)^T\right]\right) \\
&= \mathbb{E}\left[\left\|\frac{1}{m\mu_2}\mathbf{R}\mathbf{R}^T\overline{\mathbf{x}}-\overline{\mathbf{x}}\right\|_2^2\right] \\
&\overset{(c)}{=} \frac{1}{m^2\mu_2^2}\overline{\mathbf{x}}^T\mathbb{E}\left[\mathbf{R}\mathbf{R}^T\mathbf{R}\mathbf{R}^T\right]\overline{\mathbf{x}}-\|\overline{\mathbf{x}}\|_2^2 \\
&\overset{(d)}{=} \frac{1}{m}\left(\kappa+p+1\right)\|\overline{\mathbf{x}}\|_2^2
\end{aligned} \tag{3.3}
$$

where in (c) we have used $\mathbb{E}[\mathbf{R}\mathbf{R}^T] = m\mu_2\mathbf{I}_{p\times p}$ and (d) follows from Lemma 1. By substituting Eq. 3.2 and 3.3 in Eq. 3.1, we get the variance of our center estimator. □

4. Proof of Theorem 3

Consider the model, $\mathbf{x} = \sum_{j=1}^{d} w_j\sigma_j\mathbf{v}_j + \mathbf{z}$, given in Section 3 of our paper. Since w_j, $\mathbf{z}$, and $\mathbf{R}$ are independent with $\mathbb{E}[w_j] = 0$, $\mathbb{E}[w_j^2] = 1$, and $\mathbb{E}[\mathbf{z}] = \mathbf{0}$, it is easy to check that

$$
\mathbb{E}\left[\mathbf{R}\mathbf{R}^T\mathbf{x}\mathbf{x}^T\mathbf{R}\mathbf{R}^T\right] = \sum_{j=1}^{d}\sigma_j^2\mathbf{C}_{\mathbf{v}}^{(j)} + \mathbf{C}_{\epsilon} \tag{4.1}
$$

where we have defined $\mathbf{C}_{\mathbf{v}}^{(j)} \triangleq \mathbb{E}[\mathbf{R}\mathbf{R}^T\mathbf{v}_j\mathbf{v}_j^T\mathbf{R}\mathbf{R}^T]$ for $j = 1,\ldots,d$, and $\mathbf{C}_{\epsilon} \triangleq \mathbb{E}[\mathbf{R}\mathbf{R}^T\mathbf{z}\mathbf{z}^T\mathbf{R}\mathbf{R}^T]$. We first compute the covariance matrix $\mathbf{C}_{\mathbf{v}}^{(j)}$. For simplicity and without loss of generality, we omit the dependence on j. Consider a principal component $\mathbf{v} = [\nu_1,\ldots,\nu_p]^T$ with unit norm. Then,

$$
\begin{aligned}
\mathbf{C}_{\mathbf{v}} &= \mathbb{E}\left[\mathbf{R}\mathbf{R}^T\mathbf{v}\mathbf{v}^T\mathbf{R}\mathbf{R}^T\right] \overset{(a)}{=} \mathbb{E}\left[\left(\sum_{k=1}^{p}\nu_k\mathbf{t}_k\right)\left(\sum_{l=1}^{p}\nu_l\mathbf{t}_l\right)^T\right] \\
&= \sum_{k=1}^{p}\nu_k^2\Lambda_{k,k}+\sum_{k\neq l}\nu_k\nu_l\Lambda_{k,l} \\
&\overset{(b)}{=} \left(m^2+m\right)\mu_2^2\mathbf{v}\mathbf{v}^T+m\mu_2^2\mathbf{I}_{p\times p}+m\mu_2^2\kappa\text{diag}\left(\nu_1^2,\ldots,\nu_p^2\right)
\end{aligned}
$$

where in (a) we have again defined $\mathbf{t}_k$ as the k^{th} column of $\mathbf{R}\mathbf{R}^T$ (see pages 1-2 of these supplementary materials), and in (b) we have used Lemma 1. Hence,

$$\mathbf{C}_{\mathbf{v}}^{(j)} = \left(m^2+m\right)\mu_2^2\mathbf{v}_j\mathbf{v}_j^T + m\mu_2^2\mathbf{I}_{p\times p} + m\mu_2^2\kappa\text{diag}\left(v_{j,1}^2,\ldots,v_{j,p}^2\right) \tag{4.2}$$

where $v_{j,k}$ denotes the k^{th} element of the principal component $\mathbf{v}_j \in \mathbb{R}^p$. Next, we need to find the covariance matrix $\mathbf{C}_\epsilon$ induced by the noise:

$$\begin{aligned}\mathbf{C}_\epsilon &= \mathbb{E}\left[\mathbf{R}\mathbf{R}^T\mathbf{z}\mathbf{z}^T\mathbf{R}\mathbf{R}^T\right] = \frac{\epsilon^2}{p}\mathbb{E}\left[\mathbf{R}\mathbf{R}^T\mathbf{R}\mathbf{R}^T\right] \\ &= \frac{m\mu_2^2\epsilon^2}{p}\left(\kappa+m+p+1\right)\mathbf{I}_{p\times p}.\end{aligned} \tag{4.3}$$

We substitute the covariance matrices calculated in Eq. 4.2 and Eq. 4.3 into Eq. 4.1 to find the covariance matrix of the random projections of data:

$$\frac{1}{\left(m^2+m\right)\mu_2^2}\mathbb{E}\left[\mathbf{R}\mathbf{R}^T\mathbf{x}\mathbf{x}^T\mathbf{R}\mathbf{R}^T\right] = \widehat{\mathbf{C}}_{true} + \mathbf{E} \tag{4.4}$$

where $\widehat{\mathbf{C}}_{true} \triangleq \sum_{j=1}^d \sigma_j^2\mathbf{v}_j\mathbf{v}_j^T + \alpha\mathbf{I}_{p\times p}$, $\alpha \triangleq \frac{h}{m+1} + (\frac{\kappa}{p(m+1)} + \frac{(m+p+1)}{p(m+1)})\epsilon^2$, and $\mathbf{E} \triangleq \frac{\kappa}{m+1}\text{diag}(\sum_{j=1}^d \sigma_j^2 v_{j,1}^2,\ldots,\sum_{j=1}^d \sigma_j^2 v_{j,p}^2) = \frac{\kappa}{m+1}\sum_{j=1}^d \sigma_j^2\text{diag}(\mathbf{v}_j\mathbf{v}_j^T)$.

The rest of Theorem 3 then follows from linearity and the strong law of large numbers. □

4.1. Notes on the error matrix E for various types of PCs

In high dimensions, many real-world datasets, e.g. images and videos, tend to have PCs that are "smooth" in the sense that their maximum magnitude entry is of size $O(\frac{1}{\sqrt{p}})$ (See (Ailon and Chazelle, 2009)). In other words, the energy of these signals is scattered in the time/space domain, making them distributed rather than spiky. For example, predefined bases such as the DCT have been widely used for signal compression (smooth signals in the time domain are sparse in the frequency domain).

At times, however, we may be interested in cases when the PCs are sparse. Sparse PCs for example may facilitate *interpretation* of the meaning of the principal components. On the other hand, there are already many existing algorithms specifically tailored for recovering sparse PCs efficiently, so we consider this case to be of less interest for our algorithm than that of non-sparse PCs, for which there are fewer algorihms, Nevertheless, we still wish to examine the case of sparse PCs as well for completeness.

Hence, in this section, we will focus on the consequences of Theorem 3 for smooth PCs, but will also examine the implications of Theorem 3 for various other subtypes of PCs. We will find that our method is particularly effective for smooth PCs, but that many different types of PCs can also be recovered by our algorithm with great savings in memory/computation.

This may seem surprising compared with the literature on the Johnson–Lindenstrauss (JL) theorem, where sparse-Bernoulli random projections are frequently used. In this literature, (Ailon and Chazelle, 2009) for example, observe that sparse data samples (which they call "bad inputs"), can cause high variance in the JL transform when using sparse-Bernoulli random projections. Intuitively, this is due to the fact that sparse projections can capture information about the structure of the original data only when the nonzero entries of the random projections are aligned with nonzero coordinates of the data. Thus, when the data is very sparse, the probability of the nonzero entries of the random projections failing to overlap with the few nonzero coordinates of the data, and hence receiving zero information from a projection, is high. (Ailon and Chazelle, 2009) suggest preconditioning the data using a randomized Hadamard matrix in the JL transform as a fix for this problem.

However, our approach for recovering principal components is fundamentally different from the JL transform in the sense that multiple and distinct sparse projection matrices are used. There are many such sparse projections for different data samples, and only a few such projections must succeed to recover the principal component. In effect, there is *redundancy* introduced by needing only to recover a small number of PCs from a much greater number of data samples. This results

in a dramatically higher success rate for PCA with sparse random projections as opposed to JL embeddings using these projections.

In the following, we examine our results in detail for the case of several different types of PCs and provide some simulation results as well. We see that our method is effective for many different types of PCs, although the greatest savings in memory/computation are achieved for smooth PCs. Let us consider the result of Theorem 3 which states that

$$\mathbf{C}_\infty = \sum_{j=1}^{d} \sigma_j^2 \mathbf{v}_j \mathbf{v}_j^T + \alpha \mathbf{I}_{p\times p} + \frac{\kappa}{m+1} \sum_{j=1}^{d} \sigma_j^2 \text{diag}\left(\mathbf{v}_j \mathbf{v}_j^T\right).$$

As we have explained in the paper, the term $\alpha \mathbf{I}_{p\times p}$ does not change the eigenvectors and it only perturbs the eigenvalues, no matter what eigenvectors we have. We have further noticed that the amount of perturbation α is very small in high dimensions. Hence, for convenience of discussion, let's first look at only the term for a single j value (the single PC case) and omit both $\alpha \mathbf{I}_{p\times p}$ and the dependence on j, so that we are just working with:

$$\mathbf{C}'_\infty = \sigma^2 \left(\mathbf{v}\mathbf{v}^T + \frac{\kappa}{m+1} \text{diag}\left(\mathbf{v}\mathbf{v}^T\right) \right)$$

instead. We will examine this term for several cases of the PC $\mathbf{v}$ to gain intuition, then return later to examine $\mathbf{C}_\infty$ as a whole. We also remind the reader that the factor $\frac{\kappa}{m+1} = \frac{s-3}{m+1} \approx \frac{s}{m} = \frac{1}{\gamma}$ represents the sparsity of the random projections. Before we begin, we define the mutual coherence μ_{max} between the PCs and the canonical basis $\{\mathbf{e}_j\}_{j=1}^p$ as $\mu_{max} \triangleq \max_{i=1,\ldots,d, j=1,\ldots,p} |\langle \mathbf{v}_i, \mathbf{e}_j \rangle|$ (Elad, 2007) and will also define μ_{min} as $\mu_{min} \triangleq \min_{i=1,\ldots,d} \max_{j=1,\ldots,p} |\langle \mathbf{v}_i, \mathbf{e}_j \rangle|$. Now, let's begin wih the first case for $\mathbf{v}$:

(1) Low-coherence with the canonical basis, i.e. $\mu_{max}^2 = O(\frac{1}{p})$:

Note that

$$\mathbf{C}'_\infty = \sigma^2 \left(\mathbf{v}\mathbf{v}^T + \frac{\kappa}{m+1} \text{diag}\left(\mathbf{v}\mathbf{v}^T\right) \right)$$

where

$$\left\| \text{diag}\left(\mathbf{v}\mathbf{v}^T\right) \right\|_2 = \mu_{max}^2.$$

Thus, we may express $\mathbf{C}'_\infty$ as $\mathbf{C}'_\infty = \mathbf{C}_{true} + \mathbf{E}$, where $\|\mathbf{E}\|_2 \leq \frac{\kappa}{m+1} \mu_{max}^2 \sigma^2$. Hence, as mentioned in Section 5.3 of our paper, we should have $\gamma \geq C(\theta_0) \mu_{max}^2 = O(\frac{1}{p})$ to maintain a certain fixed error θ_0 for PC estimation.

(2) PC is a canonical basis element, i.e. $\mu_{max} = \mu_{min} = 1$:

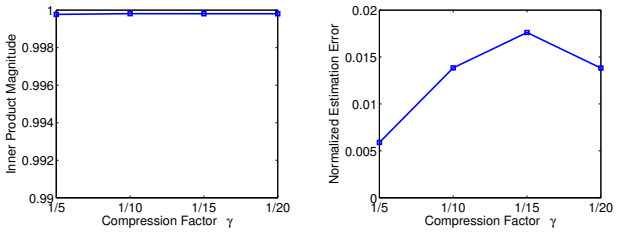
In this case, it is easy to verify that $\mathbf{v}\mathbf{v}^T = \text{diag}(\mathbf{v}\mathbf{v}^T)$. This results in the following

$$\mathbf{C}'_\infty = \sigma^2 \left(1 + \frac{\kappa}{m+1} \right) \mathbf{v}\mathbf{v}^T$$

which implies that the original principal component can be recovered accurately without any perturbation. However, the corresponding eigenvalue will be increased heavily by the **known** factor $1 + \frac{\kappa}{m+1}$. So we see that, unlike in previous Johnson-Lindenstrauss results, sparse PCs do not cause any serious issue here, apart from possible scaling of the eigenvalue, which could potentially result in reordering of the PCs.

We further provide some simulation results to validate the theoretical analysis. We synthetically generate 25,000 data samples in $\mathbb{R}^{500}$ distributed along one PC which is the first element of the canonical basis, i.e. $\mathbf{v} = [1, 0, \ldots, 0]^T$. We fix the measurement ratio $m/p = 0.3$ and estimate the principal component and its corresponding eigenvalue for various compression factors γ. Fig. 4.1(a) shows the accuracy for the estimated PC and Fig. 4.1(b) shows the estimation accuracy for the eigenvalue when scaled by the given factor. We see that our approach works well in this situation, which is consistent with the theoretical analysis.

(3) High-coherence with the canonical basis, i.e. μ_{min} is close to 1, or the principal component has a few nonzero entries:



(a) (b)

Figure 4.1. PC estimation results for synthetic data with the single PC $\mathbf{v} = [1, 0, \ldots, 0]^T$. Plot of (a) magnitude of the inner product between the estimated and true PCs, and (b) error between the estimated singular value (scaled by $1/\sqrt{1+\frac{\kappa}{m+1}}$) and the true singular value σ, normalized by the latter. Notice that both the PC and its associated singular value are well-estimated.

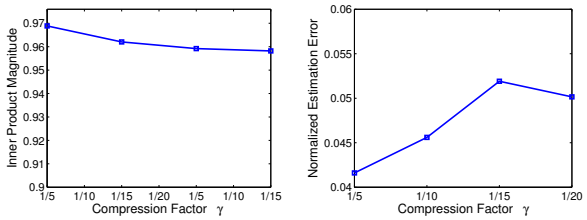
Based on the previous case, it is straightforward to re-write $\mathbf{C}'_\infty$ as

$$\begin{aligned}
\mathbf{C}'_\infty &= \sigma^2\left(\mathbf{v}\mathbf{v}^T + \frac{\kappa}{m+1}\text{diag}\left(\mathbf{v}\mathbf{v}^T\right)\right) \\
&= \sigma^2\left(\left(1+\frac{\kappa}{m+1}\right)\mathbf{v}\mathbf{v}^T + \frac{\kappa}{m+1}\left(\text{diag}\left(\mathbf{v}\mathbf{v}^T\right) - \mathbf{v}\mathbf{v}^T\right)\right) \\
&= \sigma^2\left(1+\frac{\kappa}{m+1}\right)\left(\mathbf{v}\mathbf{v}^T + \frac{\frac{\kappa}{m+1}}{1+\frac{\kappa}{m+1}}\left(\text{diag}\left(\mathbf{v}\mathbf{v}^T\right) - \mathbf{v}\mathbf{v}^T\right)\right)
\end{aligned} \tag{4.5}$$

where we note that the factor $\frac{\frac{\kappa}{m+1}}{1+\frac{\kappa}{m+1}} < 1$. We may this regard this as a new term $(1+\frac{\kappa}{m+1})\mathbf{C}_{true}$ plus a new error perturbation term $\mathbf{E}_0 = \sigma^2\frac{\kappa}{m+1}(\text{diag}(\mathbf{v}\mathbf{v}^T) - \mathbf{v}\mathbf{v}^T)$. Moreover, we provide an upper bound for the amount of perturbation error, by bounding $\left\|\mathbf{v}\mathbf{v}^T - \text{diag}\left(\mathbf{v}\mathbf{v}^T\right)\right\|_2$ as follows:

$$\begin{aligned}
\left\|\mathbf{v}\mathbf{v}^T - \text{diag}\left(\mathbf{v}\mathbf{v}^T\right)\right\|_2 &\leq \left\|\mathbf{v}\mathbf{v}^T - \text{diag}\left(\mathbf{v}\mathbf{v}^T\right)\right\|_F \\
&= \left(\left\|\mathbf{v}\mathbf{v}^T - \text{diag}\left(\mathbf{v}\mathbf{v}^T\right)\right\|_F^2\right)^{1/2} \\
&= \left(\sum_{k\neq j} v_k^2 v_j^2\right)^{1/2} \\
&= \left(\sum_{k,j} v_k^2 v_j^2 - \sum_{k=1}^p v_k^4\right)^{1/2} \\
&= \left(\left\|\mathbf{v}\mathbf{v}^T\right\|_F^2 - \sum_{k=1}^p v_k^4\right)^{1/2} \\
&\overset{(a)}{=} \left(1 - \sum_{k=1}^p v_k^4\right)^{1/2} \\
&\overset{(b)}{=} \left(1 - \frac{\sum_{k=1}^p v_k^4}{\left(\sum_{k=1}^p v_k^2\right)^2}\right)^{1/2}
\end{aligned} \tag{4.6}$$

where (a) and (b) both follow from the fact that $\|\mathbf{v}\|_2 = 1$.



(a) (b)

Figure 4.2. PC estimation results for synthetic data with the single PC $\mathbf{v} = [100, 30, 10, 0, \ldots, 0]^T$, normalized to have unit norm. Plot of (a) magnitude of the inner product between the estimated and true PCs, and (b) error between the estimated singular value (scaled by $1/\sqrt{1 + \frac{\kappa}{m+1}}$) and the true singular value σ, normalized by the latter. Notice that both the PC and its associated singular value are well-estimated.

In fact, the upper bound in Eq. 4.6 gives us an important information about the amount of perturbation. The term $\frac{\sum_{k=1}^{p} v_k^4}{\left(\sum_{k=1}^{p} v_k^2\right)^2}$ is close to 1 (i.e. the amount of perturbation is very small) when a few entries of the principal component are nonzero. Moreover, this term gets closer to 1 when the nonzero entries have a fast rate of decay. For example, let us consider the extreme case where $\mu_{max} = 1$. In this case, it is easy to see that the term $\frac{\sum_{k=1}^{p} v_k^4}{\left(\sum_{k=1}^{p} v_k^2\right)^2}$ is exactly one and thus the amount of perturbation is zero. This is consistent with our prior observation.

Furthermore, a useful upper bound for the 2-norm of the perturbation term $\mathbf{E}_0$ is obtained by considering μ_{min}. First, we see that

$$\begin{aligned} \left\|\mathbf{v}\mathbf{v}^T - \text{diag}\left(\mathbf{v}\mathbf{v}^T\right)\right\|_2 &\leq \left(1 - \sum_{k=1}^{p} v_k^4\right)^{1/2} \\ &\leq \left(1 - \mu_{min}^4\right)^{1/2} \end{aligned}$$

and this results in $\|\mathbf{E}_0\|_2 \leq \frac{\kappa}{m+1}\sqrt{1 - \mu_{min}^4}\sigma^2$. Hence, as in the previous cases, we can recover the principal component (provided that μ_{min} is sufficiently large). However, as in case (2), the corresponding eigenvalue is heavily increased by the known factor.

To validate this theoretical analysis, we consider the same synthetic simulation described in the previous case. However, we now consider the principal component $\mathbf{v} = [100, 30, 10, 0, \ldots, 0]^T$, normalized to have unit ℓ_2-norm. Fig. 4.2(a) shows the accuracy for the estimated PC and Fig. 4.2(b) shows the estimation accuracy for the singular value when scaled by the given factor $1/\sqrt{1 + \frac{\kappa}{m+1}}$. We see that our approach is less accurate than it was for the canonical basis vector, but still works quite well. We note that μ_{min} is about 0.95 here.

(4) Medium-coherence with the canonical basis:

In this case, we may follow the model of the low-coherence case, and still obtain the result that γ may scale with μ_{max}^2 to achieve fixed accuracy. However, in this case, μ_{max}^2 is not as small as in the low-coherence case, and the memory and computation savings will thus not be as aggressive, although they may still be substantial.

We have observed that our approach can recover the principal component accurately in several cases. However, when the principal component is close to a canonical basis element, the corresponding eigenvalue is increased by the known factor $1 + \frac{\kappa}{m+1} \approx 1 + \frac{1}{\gamma}$.

It is straightforward to generalize this for the situation of multiple principal components. Let us define $\mathcal{I}_{low}$ and $\mathcal{I}_{high}$ as the set of indices of principal components with low/medium and high coherence respectively, then we see that

$$\begin{aligned}
\mathbf{C}_\infty &= \sum_{j=1}^{d} \sigma_j^2 \left(\mathbf{v}_j \mathbf{v}_j^T + \frac{\kappa}{m+1} \text{diag}\left(\mathbf{v}_j \mathbf{v}_j^T \right) \right) + \alpha \mathbf{I}_{p\times p} \\
&= \sum_{j \in \mathcal{I}_{low}} \sigma_j^2 \left(\mathbf{v}_j \mathbf{v}_j^T + \frac{\kappa}{m+1} \text{diag}\left(\mathbf{v}_j \mathbf{v}_j^T \right) \right) \\
&\quad + \sum_{j \in \mathcal{I}_{high}} \sigma_j^2 \left(1 + \frac{\kappa}{m+1} \right) \left(\mathbf{v}_j \mathbf{v}_j^T + \frac{\frac{\kappa}{m+1}}{1+\frac{\kappa}{m+1}} \left(\text{diag}\left(\mathbf{v}_j \mathbf{v}_j^T \right) - \mathbf{v}_j \mathbf{v}_j^T \right) \right) + \alpha \mathbf{I}_{p\times p} \qquad (4.7) \\
&= \mathbf{C}_{true} + \frac{\kappa}{m+1} \sum_{j \in \mathcal{I}_{high}} \sigma_j^2 \mathbf{v}_j \mathbf{v}_j^T + \alpha \mathbf{I}_{p\times p} \\
&\quad \underbrace{+ \frac{\kappa}{m+1} \left(\sum_{j \in \mathcal{I}_{low}} \sigma_j^2 \text{diag}\left(\mathbf{v}_j \mathbf{v}_j^T \right) + \sum_{j \in \mathcal{I}_{high}} \sigma_j^2 \left(\text{diag}\left(\mathbf{v}_j \mathbf{v}_j^T \right) - \mathbf{v}_j \mathbf{v}_j^T \right) \right)}_{\text{2 error terms}} \qquad (4.8)
\end{aligned}$$

In this case, if all the PCs belong to either $\mathcal{I}_{low}$ or $\mathcal{I}_{high}$, then performance is not majorly impacted, and the PCs are still recovered in order. If there is a mix however, the two error terms remain small, but for large γ, scaling of the eigenvalues in $\mathcal{I}_{high}$ may reorder the original eigenvalues, and it may become possible only to recover the d-dimensional subspace containing all the PCs, not the original ordering.

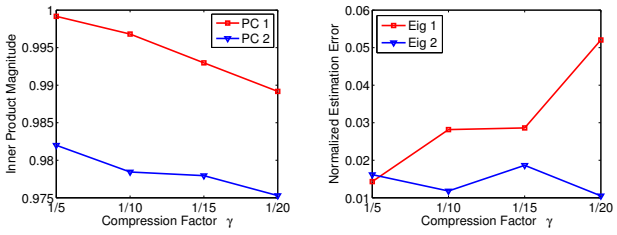
We also provide one last simulation result for the case when we have two principal components, each belonging to a different case of high/low coherence with the canonical basis elements. We synthetically generate 25,000 data samples in $\mathbb{R}^{500}$ distributed along two principal components, where the first one is drawn from the Gaussian distribution, i.e. a low-coherence PC, and the second one is $\mathbf{v} = [100, 20, 10, 0 \ldots, 0]^T$, normalized to have unit norm, i.e. a high coherence PC. We also choose $\sigma_1 = 50$ and $\sigma_2 = 10$. We fix the measurement ratio $m/p = 0.3$ and estimate the principal components and corresponding eigenvalues for various compression factors γ. Fig. 4.3(a) shows the accuracy for the estimated PCs and Fig. 4.3(b) shows the estimation accuracy for the singular values. In fact, this simple example gives us useful intuition about the performance of our method for the case of a mixture of low and high coherence PCs. We observe that for preserving the order of the PCs, the compression factor γ also needs to respect the gap between the eigenvalues. Note that in this example, even for $\gamma = \frac{1}{20}$, we have $\sqrt{1+\frac{\kappa}{m+1}} \approx \sqrt{1+\frac{1}{\gamma}} < \frac{\sigma_1}{\sigma_2}$, and the PCs do not reorder themselves. Fortunately, the rate of decay of eigenvalues for many real-world datasets such as images and videos is typically very fast. Indeed, because of this property, principal components have been widely used for signal compression, and as a sparsifying basis for compressive sensing. Hence, in these cases of fast decay, we may decrease the compression factor γ aggressively for substantial savings of memory and computation with little loss in accuracy in the case of mixed-type PCs. Furthermore, even if the PCs reorder themselves, we can still recover the d-dimensional subspace in which they all lie.

5. Proof of Theorem 4

Let us consider $\widehat{\mathbf{C}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{C}_i$, where $\mathbf{C}_i = \frac{1}{m(m+1)\mu_2^2}(\mathbf{R}_i\mathbf{y}_i)(\mathbf{R}_i\mathbf{y}_i)^T$. In the proof of Theorem 3, we showed that $\mathbb{E}[\mathbf{C}_i] = \mathbf{C}_\infty = \mathbf{C}_{true} + \alpha\mathbf{I} + \mathbf{E}$. Hence, we have $\mathbb{E}[\widehat{\mathbf{C}}_n] = \mathbf{C}_\infty$. In the following, we calculate an upper bound for the deviation of the empirical covariance matrix from its mean value:

$$\mathbb{E}\left[\left\| \widehat{\mathbf{C}}_n - \mathbf{C}_\infty \right\|_F^2 \right] = \frac{1}{n^2} tr\left(\mathbb{E}\left[\left(\sum_{i=1}^{n} (\mathbf{C}_i - \mathbf{C}_\infty) \right)^2 \right] \right) = \frac{1}{n} \left(\mathbb{E}\left[tr\left(\mathbf{C}_i^2 \right) \right] - \|\mathbf{C}_\infty\|_F^2 \right). \qquad (5.1)$$

First, we find an upper bound for $\mathbb{E}[tr(\mathbf{C}_i^2)]$. Without loss of generality, we omit the dependence on i:



(a) (b)

Figure 4.3. PC estimation results for synthetic data for the case of two principal components, one with high coherence with the canonical basis and one with low coherence with the canonical basis. Plot of (a) magnitude of the inner product between the estimated and true PCs, and (b) error between the estimated singular values (scaled by $1/\sqrt{1+\frac{\kappa}{m+1}}$ for the estimated PC with high coherence with the canonical basis) and the true singular values σ_1 and σ_2, normalized by the original values. Notice that both high coherence and low coherence PCs and their associated singular values are well-estimated.

$$\begin{aligned}\mathbb{E}\left[tr\left(\mathbf{C}^2\right)\right] &= \frac{1}{m^2\left(m+1\right)^2\mu_2^4}\mathbb{E}\left[\left\|\mathbf{R}\mathbf{R}^T\mathbf{x}\right\|_2^4\right] \overset{(a)}{=} \xi_1\mathbb{E}\left[\sum_{i=1}^{p}x_i^4\right]+\xi_2\mathbb{E}\left[\sum_{i\neq j}x_i^2x_j^2\right] \\ &\leq \xi\mathbb{E}\left[\left\|\mathbf{x}\right\|_2^4\right]=\tau_1h^2 \end{aligned} \tag{5.2}$$

where (a) follows from Lemma 2, $\xi = \max\left(\xi_1,\xi_2\right)$, $\tau_1 \triangleq \xi\left\{\left(1+\frac{1}{\text{SNR}}\right)^2+2\left(\frac{\widetilde{h}}{h^2}+\frac{2}{p}\frac{1}{\text{SNR}}+\frac{1}{p}\left(\frac{1}{\text{SNR}}\right)^2\right)\right\}$, and $\widetilde{h} \triangleq \sum_{j=1}^{d}\sigma_j^4$. The fact that

$$\mathbb{E}\left[\left\|\mathbf{x}\right\|_2^4\right]=h^2\left\{\left(1+\frac{1}{\text{SNR}}\right)^2+2\left(\frac{\widetilde{h}}{h^2}+\frac{2}{p}\frac{1}{\text{SNR}}+\frac{1}{p}\left(\frac{1}{\text{SNR}}\right)^2\right)\right\} \tag{5.3}$$

for our model was previously shown in (Qi, 2013).

Next, we find a lower bound for the second term in Eq. 5.1:

$$\begin{aligned}\left\|\mathbf{C}_\infty\right\|_F^2 &= tr\left(\left(\mathbf{C}_{true}+\alpha\mathbf{I}+\mathbf{E}\right)^2\right)=tr\left(\mathbf{C}_{true}^2\right)+2\alpha tr\left(\mathbf{C}_{true}\right)+\alpha^2p+2\alpha tr\left(\mathbf{E}\right)+2tr\left(\mathbf{C}_{true}\mathbf{E}\right)+tr\left(\mathbf{E}^2\right) \\ &\overset{(b)}{=} \widetilde{h}+2\alpha h+\alpha^2p+2\alpha\frac{\kappa}{m+1}h+\left(2\frac{\kappa}{m+1}+\left(\frac{\kappa}{m+1}\right)^2\right)\left(\sum_{i=1}^{p}\left(\sum_{j=1}^{d}\sigma_j^2v_{j,i}^2\right)^2\right)\end{aligned} \tag{5.4}$$

where (b) follows from the fact that the i^{th} diagonal entry of $\mathbf{C}_{true}=\sum_{j=1}^{d}\sigma_j^2\mathbf{v}_j\mathbf{v}_j^T$ has the form $\sum_{j=1}^{d}\sigma_j^2v_{j,i}^2$, hence

$$tr\left(\mathbf{C}_{true}\mathbf{E}\right)=\frac{\kappa}{m+1}\sum_{i=1}^{p}\left(\sum_{j=1}^{d}\sigma_j^2v_{j,i}^2\right)^2.$$

We also need to find a lower bound for the following:

$$\begin{aligned}&\sum_{i=1}^{p}\left(\sum_{j=1}^{d}\sigma_j^2v_{j,i}^2\right)^2 \overset{(c)}{\geq} \sum_{i=1}^{p}\sum_{j=1}^{d}\sigma_j^4v_{j,i}^4=\sum_{j=1}^{d}\sigma_j^4\sum_{i=1}^{p}v_{j,i}^4 \\ &\overset{(d)}{\geq} \frac{1}{p}\widetilde{h}\end{aligned} \tag{5.5}$$

where (c) follows from the inequality $(\sum_i a_i)^2 \geq \sum_i a_i^2$ for non-negative $\{a_i\}$, and (d) is due to the inequality for p-dimensional vectors $\|\mathbf{v}\|_{r_2} \leq p^{(1/r_2 - 1/r_1)} \|\mathbf{v}\|_{r_1}$, when $r_1 > r_2 > 0$. Therefore, we get the following lower bound using Eq. 5.4 and 5.5:

$$\|\mathbf{C}_\infty\|_F^2 \geq \tau_2 h^2 \tag{5.6}$$

where $\tau_2 \triangleq \frac{\widetilde{h}}{h^2}\left(\frac{p-1}{p} + \frac{1}{p}\left(1 + \frac{\kappa}{m+1}\right)^2\right) + \beta\left(2 + \beta p + 2\frac{\kappa}{m+1}\right)$ and $\beta \triangleq \frac{\alpha}{h}$. □

References

N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009. 4.1

M. Elad. Optimized projections for compressed sensing. *IEEE Transactions on Signal Processing,*, 55(12):5695–5702, 2007. 4.1

H. Qi. Low-dimensional signal models in compressive sensing. *Ph.D. Thesis, Dept. of Electrical, Computer, and Energy Eng., University of Colorado at Boulder*, 2013. 5