Dimensionality Reduction

Peng Lingwei

August 24, 2019

Contents

23 Dimensionality Reduction

In this chapter, we discuss linear transformation.

23.1 PRINCIPAL COMPONENT ANALYSIS (PCA)

Definition 1. ***(PCA target).*** *For a data $S = (x_1, \dots, x_m) \in \mathbb{R}^d$, finding a compression matrix W and a recovering matrix U, satisfy*

Lemma 1.

$$\arg \min_{W \in \mathbb{R}^{n,d}, U \in \mathbb{R}^{d,n}} \sum_{i=1}^{m} \|x_i - UWx_i\|_2^2. = \arg \min_{V \in \mathbb{R}^{d,n} : V^T V = I^n} \sum_{i=1}^{m} \|x_i - V^T V x_i\|_2^2$$

$$= \arg \max_{V \in \mathbb{R}^{d,n} : V^T V = I^n} trace \left(V^T \sum_{i=1}^{m} x_i x_i^T V \right)$$

And if V's column is the matrix $\sum_{i=1}^m x_i x_i^T$'s n leading eigenvectors, we reach the maximum.

Proof. Let $V \in \mathbb{R}^{d,n}$ be a matrix whose columns form an orthonormal basis of this subspace, then $\{UWx : x \in S\} \subset \{Vy : y \in \mathbb{R}^n\}$, then

$$\forall V \in \left\{ V^T V = I^n, \mathbb{R}^{d,n} \right\}, \quad \arg \min_{y_i} \|x_i - Vy_i\|^2 = V^T x_i$$

$$\min_{W \in \mathbb{R}^{n,d}, U \in \mathbb{R}^{d,n}} \sum_{i=1}^{m} \|x_i - UWx_i\|_2^2 \geq \min_{V : V^T V = I^n} \min_{y_1, \dots, y_m} \sum_{i=1}^{m} \|x_i - Vy_i\|^2$$

$$= \min_{V : V^T V = I^n} \sum_{i=1}^{m} \|x_i - VV^T x_i\| = \min_{V : V^T V = I^n} \sum_{i=1}^{m} \|x\|^2 - 2x^T VV^T x + x^T VV^T VV^T x$$

$$= \min_{V : V^T V = I^n} \sum_{i=1}^{m} \|x\|^2 - x^T VV^T x = \min_{V : V^T V = I^n} \sum_{i=1}^{m} \|x\|^2 - trace(V^T x x^T V)$$

$$= \max_{V \in \mathbb{R}^{d,n} : V^T V = I^n} trace \left(V^T \sum_{i=1}^{m} x_i x_i^T V \right)$$

Let $A = \sum_{i=1}^m x_i x_i^T$. The matrix A is symmetric and therefore it can be written using spectral decomposition as $A = UDU^T$, where D is diagonal and $U^T U = UU^T = I^d$.

$$\max_{V \in \mathbb{R}^{d,n} : V^T V = I^n} trace \left(V^T \sum_{i=1}^{m} x_i x_i^T V \right) = \max_{V \in \mathbb{R}^{d,n} : V^T V = I^n} trace \left(V^T U D U^T V \right)$$

$$= \max_{W \in \mathbb{R}^{d,n} : W^T W = I^n} trace \left(W^T D W \right) = \sum_{i=1}^{d} D_{i,i} \sum_{j=1}^{n} W_{i,j}^2$$

First we have $\sum_{i=1}^d \sum_{j=1}^n W_{i,j}^2 = n$.

Second, We expand W tobe $\tilde{W}$, whose first n columns are the columns of W, and $\tilde{W}^T\tilde{W} = I^d$.Then $\sum_{j=1}^{d} \tilde{W}_{i,j}^2 = 1 \Rightarrow \sum_{j=1}^{n} W_{i,j}^2 \leq 1$. $((\tilde{W}\tilde{W}^T - I^d)\tilde{W} = 0 \Rightarrow \tilde{W}\tilde{W}^T = I^d)$.Then,if $D_{1,1} \geq D_{2,2} \geq \ldots \geq D_{d,d}$,

$$\max_{W\in\mathbb{R}^{d,n}:W^TW=I^n} \sum_{i=1}^{d} D_{i,i} \sum_{j=1}^{n} W_{i,j}^2 \leq \max_{\beta\in[0,1]^d:\|\beta\|_1\leq n} \sum_{i=1}^{d} D_{i,i}\beta_i = \sum_{i=1}^{n} D_{i,i}$$

It's easy to varify that if V's column is U's first n columns, then

$$\max_{V\in\mathbb{R}^{d,n}:V^TV=I^n} trace\left(V^TUDU^TV\right) = \sum_{i=1}^{n} D_{i,i}$$

□

Because $\sum_{i=1}^{m} \|x_i\|^2 = trace(A) = \sum_{i=1}^{d} D_{i,i}$, so we obtain that

$$\min_{V:V^TV=I^n} \sum_{i=1}^{m} \|x\|^2 - trace(V^Txx^TV) = \sum_{i=n+1}^{d} D_{i,i}$$

23.1.1 A More Efficient Solution for the Case $d \gg m$

In previous section, constructing the matrix A need $O(md^2)$ and calculating eigenvalues of A need $O(d^3)$. If $d \gg m$, we can calculate the PCA solution more efficiently.

Instead of analysing $A = X^TX$, we consider $B = XX^T$. The B's eigenvector u satisfies $Bu = \lambda u \Rightarrow X^TXX^Tu = \lambda X^Tu \Rightarrow \frac{X^Tu}{\|X^Tu\|}$ is an eigenvector of A with eigenvalue of λ. Then the complexity is $O(m^3) + O(m^2d)$.

23.1.2 Kernel PCA

Any valid kernel $K(x, x')$ implies a mapping $\phi : \mathbb{R}^d \mapsto \mathbb{R}^n$, and $X = [\phi(\vec{x}_1), \ldots, \phi(\vec{x}_m)]$. Then

$$A = \sum_{i=1}^{m} \phi(x_i)\phi^T(x_i) = XX^T = UDU^T$$

We want use PCA without knowledge of ϕ. First, we can easily calculate the eigenvectors of $K(X, X) = X^TX = U_TD_TU_T^T$. We already have $U = XU_T = [Xu_1, \ldots, Xu_m]$.Let $V = [Xu_1, \ldots, Xu_k]$. By this kernel PAC, every sample $\vec{x}$ transforms into

$$V^T\phi(\vec{x}) = \begin{pmatrix} u_1^T \\ \vdots \\ u_k^T \end{pmatrix} X^T\phi(\vec{x}) = [u_1, \ldots, u_k]^T \begin{pmatrix} K(\vec{x}_1, \vec{x}) \\ \ldots \\ K(\vec{x}_m, \vec{x}) \end{pmatrix}$$

23.1.3 An Interpretation of PCA as Variance Maximization

Target:

$$\arg\max_{\vec{w}:\|\vec{w}\|=1} Var\left[\langle\vec{w}, X\rangle\right] = \arg\max_{\vec{w}:\|\vec{w}\|} \frac{1}{m} \sum_{i=1}^{m} \left(\langle\vec{w}, \vec{x}_i\rangle\right)^2 = \arg\max_{\vec{w}:\|\vec{w}\|} trace\left(\vec{w}^T \sum_{i=1}^{m} \vec{x}_i\vec{x}_i^T \vec{w}\right)$$

Then the first principal component $\vec{w}_1$ is the eigenvector of matrix $\sum_{i=1}^{m} \vec{x}_i \vec{x}_i^T$, coresponding to the larget eigenvalue.

Then, we want get second principal component $\vec{w}_2$ satisfying

$$\arg \max_{\vec{w}:\|\vec{w}\|=1, \mathbb{E}[\langle \vec{w}_1, \vec{x}\rangle \langle \vec{w}, \vec{x}\rangle]=0} Var\left[\langle \vec{w}, \vec{x}\rangle\right]$$

$$\vec{w}^* = \mathbb{E}[\langle \vec{w}_1, \vec{x}\rangle \langle \vec{w}, \vec{x}\rangle] = \vec{w}_1^T \mathbb{E}\left[\vec{x}\vec{x}^T\right] \vec{w} = \lambda_1 \vec{w}_1^T \vec{w} = 0$$

Then $\vec{w}^*$ is the second largest eigenvector $\vec{w}_2$.

23.2 RANDOM PROJECTIONS

For a random matrix W, we want $\frac{\|Wx_1 - Wx_2\|}{\|x_1 - x_2\|} \approx 1$.

Lemma 2. *Fix some $x \in \mathbb{R}^d$. Let $W \in \mathbb{R}^{n,d}$ be a random matrix such that each $W_{i,j}$ is an independent normal random variable. Then for every $\epsilon \in (0, 3)$ we have*

$$\mathbb{P}\left[\left|\frac{\|(1/\sqrt{n})Wx\|^2}{\|x\|^2} - 1\right| > \epsilon\right] \le 2e^{-\epsilon^2 n/6}$$

Proof. Wlog we can assume that $\|x\|^2 = 1$.Then we need to proof

$$\mathbb{P}\left[(1-\epsilon)n \le \|Wx\|^2 \le (1+\epsilon)n\right] \ge 1 - 2e^{-\epsilon^2 n/6}$$

Let w_i be the ith row of W. The random variable $\langle w_i, x\rangle$ is a combination of d independent normal random variables, which is still normal random variable. Then $\|Wx\|^2 = \sum_{i=1}^{n} \left(\langle w_i, x\rangle\right)^2 \sim \chi_n^2$

So we can use the measure concentration property of χ^2 random variables. □

Lemma 3. *Let $Z \sim \chi_k^2$. Then*

$$\forall \epsilon > 0, \quad \mathbb{P}\left[Z \le (1-\epsilon)k\right] \le e^{-\epsilon^2 k/6}$$

$$\forall \epsilon \in (0,3), \quad \mathbb{P}\left[Z \ge (1+\epsilon)k\right] \le e^{-\epsilon^2 k/6}$$

Proof. For normally distributed random variable, $\mathbb{E}[X] = 0, \mathbb{E}[X^2] = 1, \mathbb{E}[X^4] = 3$. Since $\forall a \ge 0, e^{-a} \le 1 - a + \frac{a^2}{2}$, then

$$\mathbb{E}\left[e^{-\lambda X^2}\right] \le 1 - \lambda \mathbb{E}\left[X^2\right] + \frac{\lambda^2}{2}\mathbb{E}\left[X^4\right] = 1 - \lambda + \frac{3}{2}\lambda^2 \le e^{-\lambda + \frac{3}{2}\lambda^2}$$

$$\begin{aligned}\mathbb{P}\left[-Z \ge -(1-\epsilon)k\right] =& \mathbb{P}\left[e^{-\lambda Z} \ge e^{-(1-\epsilon)k\lambda}\right] \le e^{(1-\epsilon)k\lambda}\mathbb{E}\left[e^{-\lambda Z}\right] \\ =& e^{(1-\epsilon)k\lambda} \prod_{i=1}^{k} \left(\mathbb{E}\left[e^{-\lambda X_i^2}\right]\right) \\ \le& e^{(1-\epsilon)k\lambda} e^{-\lambda k + \frac{3}{2}\lambda^2 k} = e^{-\epsilon k \lambda + \frac{3}{2}k\lambda^2} (= e^{-\epsilon^2 k/6} \ if \ \lambda = \epsilon/3)\end{aligned}$$

Here is a closed form expression for χ_k^2 distributed random variable:

$$\forall \lambda < \frac{1}{2}, \mathbb{E}\left[e^{\lambda Z^2}\right] = (1 - 2\lambda)^{-k/2}$$

$$
\begin{aligned}
&\mathbb{P}\left[Z \geq (1+\epsilon)k\right] = \mathbb{P}\left[e^{\lambda Z} \geq e^{(1+\epsilon)k\lambda}\right] \leq e^{-(1+\epsilon)k\lambda}\mathbb{E}\left[e^{\lambda Z}\right] \\
=&e^{-(1+\epsilon)k\lambda}(1-2\lambda)^{-k/2} \leq e^{-(1+\epsilon)k\lambda}e^{k\lambda} = e^{-\epsilon k\lambda}(= e^{-\epsilon^2 k/6},\ if\ \lambda = \epsilon/6)
\end{aligned}
$$

$\square$

Lemma 4. ***(Johnson-Lindenstrauss Lemma).*** *Let $x \in S$, then*

$$\mathbb{P}\left[\sup_{x\in S}\left|\frac{\|(1/\sqrt{n})Wx\|^2}{\|x\|^2} - 1\right| > \epsilon\right] \leq 2\,|S|\,e^{-\epsilon^2 n/6} \leq \delta \Rightarrow \epsilon \geq \sqrt{\frac{6\ln(2|S|/\delta)}{n}} \in (0,3)$$

The preceeding lemma does not depend on the original dimension of x.

23.3 COMPRESSED SENSING

1. Prior assumption: the original vector is sparse in some basis;
2. Denote: $\|\vec{x}\|_0 = |\{i : x_i \neq 0\}|$;
3. If $\|x\|_0 \leq s$, we can represent it using s (index, value) pairs;
4. Further assume: $\vec{x} = U\vec{\alpha}$, where $\|\vec{\alpha}\|_0 \leq s$, and U is a fixed orthonormal matrix;
5. Compressed sensing: get $\vec{x}$, compress $\vec{x}$ into $\vec{\alpha} = U^T x$ and represent $\vec{\alpha}$ by its s (index, value) pairs.

The key result:

1. It is possible to reconstruct any sparse signal fully if it wars compressed by $x \mapsto Wx$, where W is a matrix which satisfies a condition called the Restricted Isoperimentric Property.
2. The reconstruction can be calculated in polynomial timee by solving a linear program.
3. A random $n \times d$ matrix is likely to satisfy the RIP condition provided that n is greater than an order of $s\log(d)$

Definition 2. ***(Restricted Isoperimentric Property).*** *A matrix $W \in \mathbb{R}^{n,d}$ is $(\epsilon, s) - RIP$ if $x \neq 0$ s.t. $\|x\|_0 \leq s$*

$$\forall\vec{x} \in \left\{\|\vec{x}\|_0 \leq s \wedge \vec{x} \in \mathbb{R}^d\right\}, \quad \left|\frac{\|W\vec{x}\|_2^2}{\|\vec{x}\|_2^2} - 1\right| \leq \epsilon.$$

Theorem 1. *Let $\epsilon < 1$ and W be a $(\epsilon, 2s)-RIP$ matrix. Let $\vec{x} \in \left\{\|\vec{x}\|_0 \leq s \wedge \vec{x} \in \mathbb{R}^d\right\}$ and $\vec{y} = W\vec{x}$. Then,*

$$\vec{x} = \vec{z} \in \arg\max_{\vec{z}:W\vec{z}=\vec{y}} \|\vec{z}\|_0$$

Proof. If $\vec{x} \neq \vec{z}$, we can get $\|\vec{z}\|_0 \leq \|\vec{x}\|_0 \leq s$, so $\|\vec{x}-\vec{z}\| \leq 2s. \left|\frac{\|W(\vec{x}-\vec{z})\|_2^2}{\|\vec{x}-\vec{z}\|_2^2} - 1\right| \leq \epsilon$ which leads to a contradiction. $\square$

Theorem 2. *Further assume that $\epsilon < \frac{1}{1+\sqrt{2}}$, then*

$$\vec{x} = \arg\min_{\vec{v}:W\vec{v}=\vec{y}} \|\vec{v}\|_0 = \arg\min_{\vec{v}:W\vec{v}=\vec{y}} \|\vec{v}\|_1.$$

A stronger theorem follows

Theorem 3. *Let $\epsilon < \frac{1}{1+\sqrt{2}}$ and let $W \in \mathbb{R}^{n,d}$ be a $(\epsilon, 2s) - RIP$ matrix. Let $\vec{x} \in \mathbb{R}^d$ and denote*

$$\vec{x}_s \in \arg \min_{\vec{v}:\|\vec{v}\|_0 \leq s} \|\vec{x} - \vec{v}\|_1.$$

note that $\vec{x}_s$ is the vector which equals $\vec{x}$ on the s leargest elements of $\vec{x}$ and equals 0 elsewhere. Let $\vec{y} = W\vec{x}$ be the compression of $\vec{x}$ and let

$$\vec{x}^* \in \arg \min_{\vec{v}:W\vec{v}=\vec{y}} \|\vec{v}\|_1$$

Then,

$$\|\vec{x}^* - \vec{x}\|_2 \leq 2\frac{1+\rho}{1-\rho} s^{-1/2} \|\vec{x} - \vec{x}_s\|_1$$

where $\rho = \sqrt{2}\epsilon/(1-\epsilon)$.

Proof. Let $\vec{h} = \vec{x}^* - \vec{x}$. Given a vector $\vec{v}$ and a set of indices I we denote by $\vec{v}_I$ the vector whose ith element is v_i if $i \in I$ and 0 otherwise.

Then we partition the set of indices $[d] = \{1, \ldots, d\}$ into disjoint sets of size s, $[d] = T_0 \cup T_1 \cup T_2 \ldots T_{d/s-1}$.We assume d/s is an integer, then $|T_i| = s$.

T_0 has the s indices corresponding to the s largest elements in absolute values of $\vec{x}$. Let $T_0^c = [d] \backslash T_0$. Next, T_1 will be the s indices corresponding to the s largest elements in absolute value of $h_{T_0^c}$. Let $T_{0,1} = T_0 \cup T_1$ and $T_{0,1}^c = [d] \backslash T_{0,1}$. Next, T_2 will correspond to the s largest elements in absolute value of $h_{T_{0,1}^c}$. And soon on.

Lemma 5. *If W is an $(\epsilon, 2s) - RIP$ matrix. Then, for any two disjoint sets I,J, both of size at most s, and for any vector $\vec{u}$ we have that $\langle Wu_I, Wu_J \rangle \leq \epsilon \|u_I\|_2 \|u_J\|$*

Proof.

$$\left| \frac{\|W(\vec{u}_I + \vec{u}_J)\|_2^2}{\|\vec{u}_I + \vec{u}_J\|_2^2} - 1 \right| \leq \epsilon$$

$$\begin{aligned}
\langle W\vec{u}_I, W\vec{u}_J \rangle =& \frac{1}{4} \left(\|W\vec{u}_I + W\vec{u}_J\|_2^2 - \|W\vec{u}_I - W\vec{u}_J\|_2^2 \right) \\
\leq& \frac{1}{4} \left((1+\epsilon)\|\vec{u}_I + \vec{u}_J\|_2^2 + (\epsilon - 1)\|\vec{u}_I - \vec{u}_J\|_2^2 \right) \\
=& \frac{\epsilon}{2} \left(\|\vec{u}_I\|_2^2 + \|\vec{u}_J\|_2^2 \right)
\end{aligned}$$

W.l.o.g we assume $\|\vec{u}_I\| = k\|\vec{u}_J\|$, then

$$\begin{aligned}
\langle W\vec{u_I}, kW\vec{u_J} \rangle \leq& \frac{\epsilon}{2} \left(\|\vec{u}_I\|_2^2 + k^2\|\vec{u}_J\|_2^2 \right) = k\epsilon \|\vec{u}_I\| \|\vec{u}_J\| \\
\langle W\vec{u}_I, W\vec{u}_J \rangle \leq& \epsilon \|\vec{u}_I\| \|\vec{u}_J\|
\end{aligned}$$

□

Clearly, $\|h\|_2 = \|h_{T_{0,1}} + h_{T_{0,1}^c}\|_2 \leq \|h_{T_{0,1}}\|_2 + \|h_{T_{0,1}^c}\|_2$.
If we have following two claims:

1. $\|h_{T_{0,1}^c}\|_2 \leq \|h_{T_0}\|_2 + 2s^{-1/2}\|\vec{x} - \vec{x}_s\|_1$;

2. $\|h_{T_{0,1}}\|_2 \le \frac{2\rho}{1-\rho} s^{-1/2}\|\vec{x}-\vec{x}_s\|_1$.

Then we can proof the theorem

$$\begin{aligned}\|h\|_2 \le &\|h_{T_{0,1}}\|_2 + \|h_{T_{0,1}^c}\|_2 \le 2\|h_{T_{0,1}}\|_2 + 2s^{-1/2}\|\vec{x}-\vec{x}_s\|_1 \\ \le &2\left(\frac{2\rho}{1-\rho}+1\right)s^{-1/2}\|\vec{x}-\vec{x}_s\|_1 = 2\frac{1+\rho}{1-\rho}s^{-1/2}\|\vec{x}-\vec{x}_s\|_1\end{aligned}$$

Now we prove claims1: $\forall i \in T_j, i' \in T_{j-1}$, we have $|h_i| \le |h_i'|$. Therfore,

$$\begin{aligned}&\|h_{T_j}\|_\infty \le \|h_{T_{j-1}}\|_1/s \\ \Rightarrow &\|h_{T_j}\|_2 \le s^{1/2}\|h_{T_j}\|_\infty \le s^{-1/2}\|h_{T_{j-1}}\|_1 \\ \Rightarrow &\|h_{T_{0,1}^c}\| \le \sum_{j\ge 2}\|h_{T_j}\|_2 \le s^{-1/2}\|h_{T_0^c}\|_1\end{aligned}$$

$$\|\vec{x}\|_1 \ge \|\vec{x}+\vec{h}\|_1 = \sum_{i\in T_0}|x_i+h_i| + \sum_{i\in T_0^c}|x_i+h_i| \ge \|x_{T_0}\|_1 - \|h_{T_0}\|_1 + \|h_{T_0^c}\|_1 - \|x_{T_0^c}\|_1$$

$$\begin{aligned}&\|h_{T_0^c}\|_1 \le \|\vec{x}\|_1 - \|x_{T_0}\|_1 + \|x_{T_0^c}\|_1 + \|h_{T_0}\|_1 = 2\|x_{T_0^c}\|_1 + \|h_{T_0}\|_1 \\ &\|h_{T_{0,1}^c}\|_2 \le s^{-1/2}\left(2\|x_{T_0^c}\|_1 + \|h_{T_0}\|_1\right) \le \|h_{T_0}\|_2 + 2s^{-1/2}\|x-x_s\|\end{aligned}$$

Then we prove claim2: For RIP condition,

$$\begin{aligned}&(1-\epsilon)\|h_{T_{0,1}}\|_2^2 \le \|Wh_{T_{0,1}}\|_2^2 = \|Wh - \sum_{j\ge 2}Wh_{T_j}\|_2^2 = \|\sum_{j\ge 2}Wh_{T_j}\|_2^2 \\ =&\sum_{j\ge 2}\langle Wh_{T_0}+Wh_{T_1}, Wh_{T_j}\rangle \le \epsilon(\|h_{T_0}\|_2 + \|h_{T_1}\|_2)\sum_{j\ge 2}\|h_{T_j}\|_2 \\ \le&\sqrt{2}\epsilon\|h_{T_{0,1}}\|_2\|h_{T_{0,1}^c}\|_2 \le \sqrt{2}\epsilon\|h_{T_{0,1}}\|_2 s^{-1/2}\|h_{T_0^c}\|_1\end{aligned}$$

$$\begin{aligned}&\|h_{T_{0,1}}\|_2 \le \frac{\sqrt{2}\epsilon}{1-\epsilon}s^{-1/2}\|h_{T_0^c}\|_1 \le \frac{\sqrt{2}\epsilon}{1-\epsilon}s^{-1/2}\left(\|h_{T_0}\|_1 + 2\|x_{T_0^c}\|_1\right) \\ \le&\frac{\sqrt{2}\epsilon}{1-\epsilon}\left(\|h_{T_{0,1}}\|_1 + 2s^{-1/2}\|x_{T_0^c}\|_1\right) \le \frac{2\rho}{1-\rho}s^{-1/2}\|x_{T_0^c}\|_1, \quad \rho = \frac{\sqrt{2}\epsilon}{1-\epsilon}, \epsilon \le \frac{1}{\sqrt{2}+1}\end{aligned}$$

□

Theorem 4. *Let U be an arbitrary fixed $d\times d$ orthonormal matrix, let ϵ, δ be scalars in $(0,1)$, let s be an integer in $[d]$, and let n be an integer that satisfies*

$$n \ge 100\frac{s\log(40d/(\delta\epsilon))}{\epsilon^2}$$

Let $W \in \mathbb{R}^{n,d}$ be a matrix s.t. each element of W is distributed normally with zero mena and variance of 1/n. Then, with probability of at least $1-\delta$ over the choice of W, the matrix WU is $(\epsilon, s) - RIP$

Proof.

Lemma 6. *Let $\epsilon \in (0,1)$. There exists a finite set $Q \subset \mathbb{R}^d$ of size $|Q| \le (3/\epsilon)^d$ such that*

$$\sup_{\vec{x}:\|\vec{x}\|\le 1} \min_{\vec{v}\in Q} \|\vec{x}-\vec{v}\| \le \epsilon$$

Proof. Let k be an integer and let

$$Q' = \left\{ \vec{x} \in \mathbb{R}^d : \forall j \in [d], \exists i \in \{-k, -k+1, \dots, k\} \ s.t. \ x_j = \frac{i}{k} \right\}$$

Clearly, $\|Q'\| = (2k+1)^d$. We shall set $Q = Q' \cap B_2(1)$, where $B_2(1)$ is the unit l_2 ball of $\mathbb{R}^d$. The volumn of $B_2(1)$ is $\frac{\pi^{d/2}}{\Gamma(1+d/2)}$. If d is even therefore $\Gamma(1+d/2) = (d/2)! \ge (\frac{d/2}{\epsilon})^{d/2}$. Then

$$|Q| \le (2k+1)^d (\pi/e)^{d/2} (d/2)^{-d/2} 2^{-d}$$

$\forall \vec{x} \in B_2(1)$ let $\vec{v} \in Q$ that $v_i = sign(x_i) \lfloor |x_i| k \rfloor / k$. We can gurantee $|x_i - v_i| \le 1/k$ and thus

$$\|\vec{x}-\vec{v}\|_2 \le \frac{\sqrt{d}}{k} \le \epsilon \Rightarrow k = \left\lceil \sqrt{d}/\epsilon \right\rceil.$$

$$|Q| \le (3\sqrt{d}/(2\epsilon))^d (\pi/e)^{d/2} (d/2)^{-d/2} = \left(\frac{3}{\epsilon}\sqrt{\frac{\pi}{2e}}\right)^d \le \left(\frac{3}{\epsilon}\right)^d.$$

□

Lemma 7. *Let U be an orthonormal $d \times d$ matrix and let $I \subset [d]$ be a set of indices of size $|I| = s$. Let S be the span of $\{U_i : i \in I\}$, where U_i is the ith column of U. Let $\delta \in (0,1)$, $\epsilon \in (0,1)$, and $n \in \mathbb{N}$ such that*

$$n \ge 24 \frac{\log(2/\delta) + s\log(12/\epsilon)}{\epsilon^2}$$

$\forall W \in \mathbb{R}^{n,d}$ such that $W_{ij} \sim N(0, 1/n)$, we have

$$\mathbb{P}\left\{ \sup_{\vec{x}\in S} \left| \frac{\|W\vec{x}\|}{\|\vec{x}\|} - 1 \right| < \epsilon \right\} \ge 1 - \delta$$

Proof. It suffices to prove the lemma for all $\vec{x} \in S$ with $\|\vec{x}\|_2 = 1$. We can write $\vec{x} = U_I\vec{\alpha}$ where $\vec{\alpha} \in \mathbb{R}^s$, $\|\vec{\alpha}\|_2 = 1$ Then $\exists Q$ of size $|Q| \le (12/\epsilon)^s$ such that

$$\sup_{\vec{\alpha}:\|\vec{\alpha}\|_2=1} \min_{\vec{v}\in Q} \|\vec{\alpha}-\vec{v}\| \le \epsilon/4 \Rightarrow \sup_{\vec{\alpha}:\|\vec{\alpha}\|_2=1} \min_{\vec{v}\in Q} \|\vec{U_I\alpha} - U_I\vec{v}\| \le \epsilon/4.$$

If $n \ge 24\frac{\log(2/\delta)+s\log(12/\epsilon)}{\epsilon^2}$ then

$$\mathbb{P}\left\{ \sup_{\vec{v}\in Q} \left| \frac{\|WU_I\vec{v}\|^2}{\|U_I\vec{v}\|^2} - 1 \right| \le \epsilon/2 \right\} \ge 1 - \delta$$

$$\mathbb{P}\left\{ \sup_{\vec{v}\in Q} \left| \frac{\|WU_I\vec{v}\|}{\|U_I\vec{v}\|} - 1 \right| \le \sup_{\vec{v}\in Q} \left| \frac{\|WU_I\vec{v}\|^2}{\|U_I\vec{v}\|^2} - 1 \right| \le \epsilon/2 \right\} \ge 1 - \delta$$

We denote $\forall \vec{x} \in S$, $\frac{\|W\vec{x}\|}{\|\vec{x}\|} \leq 1 + a$, where a is the smallest number satisfying the previous inequation. Then

$$\|W\vec{x}\| \leq \|WU_I\vec{v}\| + \|W(\vec{x} - U_I\vec{v})\| \leq 1 + \epsilon/2 + (1+a)\epsilon/4$$

$$\forall \vec{x} \in S, \frac{\|W\vec{x}\|}{\|\vec{x}\|} \leq 1 + \epsilon/2 + (1+a)\epsilon/4$$

By the definition of a, we have $a \leq \epsilon/2 + (1+a)\epsilon/4 \Rightarrow \leq \frac{3\epsilon}{4-\epsilon} \leq \epsilon$. Similarily, we difine b as minimum number satifies $\forall \vec{x} \in S$, $\frac{\|W\vec{x}\|}{\|\vec{x}\|} \geq 1 - b$.

$$\|W\vec{x}\| \geq \|WU_I\vec{v}\| - \|W(\vec{x} - U_I\vec{v})\| \geq 1 - \epsilon/2 - (1-b)\epsilon/4.$$

$$b \leq \epsilon/2 + (1-b)\epsilon/4 \Rightarrow b \leq \frac{4}{4+\epsilon}\epsilon \leq \epsilon$$

□

The preceding lemma tells us that $\forall \vec{x} \in S$ of unit norm we have

$$(1-\epsilon) \leq \|W\vec{x}\| \leq (1+\epsilon) \Rightarrow (1-2\epsilon) \leq \|W\vec{x}\|^2 \leq (1+3\epsilon)$$

The total number of indices of I is $\mathbb{C}_d^s \leq (ed/s)^s$, by union bound, we need

$$n \geq 24\frac{\log\left(\frac{2}{\delta}\cdot\left(\frac{ed}{s}\right)^s\right) + s\log(36/\epsilon)}{(\epsilon/3)^2} = 216\frac{\log(2/\delta) + s\log\left(\frac{36ed}{s\epsilon}\right)}{\epsilon^2}$$

□

23.4 PAC OR COMPRESSED SENSING

1. PCA assumes that the set of examples is contained in an n dimensional subspace of $\mathbb{R}^d$;

2. Compressed sensing assumes the set of examples is sparse (in some basis).