Dimensionality Reduction

Peng Lingwei

August 21, 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 $S = (x_1, \ldots, x_m) \in \mathbb{R}^d$, 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

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

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

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

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

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

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 $D_{1,1} \geq D_{2,2} \geq \ldots \geq D_{d,d}$,

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

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

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

□

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

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

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

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*

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

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

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

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*

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

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

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

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

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

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

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

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

□

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

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

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. □

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

Then we can proof the theorem

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

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

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

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

$$\|h_{T_0^c}\|_1 \leq \|\vec{x}\|_1 - \|x_{T_0}\|_1 + \|x_{T_0^c}\|_1 + \|h_{T_0}\|_1 = 2\|x_{T_0^c}\|_1 + \|h_{T_0}\|_1$$

$$\|h_{T_{0,1}^c}\|_2 \leq s^{-1/2}\left(2\|x_{T_0^c}\|_1 + \|h_{T_0}\|\right) \leq \|h_{T_0}\|_2 + 2s^{-1/2}\|x-x_s\|$$

Then we prove claim2: For RIP condition,

$$(1-\epsilon)\|h_{T_{0,1}}\|_2^2 \leq \|Wh_{T_{0,1}}\|_2^2 = \|Wh - \sum_{j\geq 2}Wh_{T_j}\|_2^2 = \|\sum_{j\geq 2}Wh_{T_j}\|_2^2$$

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

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

□

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*

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

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 $(\epsilon, s) - RIP$

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).