PCA

We are given a set of vectors $(x_1, \dots, x_m)$, $x_i \in \mathbb{R}^d$.

Goal: "reduce the dimensionality" of these vectors. i.e., for some $n < d$, find a matrix $W \in \mathbb{R}^{n \times d}$ s.t. Wx is compressed representation. Let $U \in \mathbb{R}^{d \times n}$ be the reconstruction matrix s.t. UWx is reconstruction of x.

Formally, want to solve

$$\min_{\substack{U \in \mathbb{R}^{d \times n} \\ W \in \mathbb{R}^{n \times d}}} \sum_{i=1}^{m} \|x_i - UWx_i\|_2^2 \qquad (*)$$

Note: if a_i are columns of A, then $\sum \|a_i\|_2^2 = \|A\|_F^2$, so we are minimizing the Frobenius norm

$$\|X - UWX\|_2^2, \text{ where } X = [x_1, \dots, x_m] \in \mathbb{R}^{d \times m}.$$

The following claim simplifies the problem

Claim: Let (U, W) be a soln to $(*)$. Then the columns of U are orthonormal, & $U^T U = I_n$, & $W = U^T$.

pf

1. Fix any U, W.
2. Let $R := \{UWx : x \in \mathbb{R}^d\}$ be range of map. Note $R \subset \mathbb{R}^d$ but $\dim R = n$. Let $V \in \mathbb{R}^{d \times n}$ be matrix whose cols form orthonormal basis for R.
3. Let x be given & consider $\min_y \|x - Vy\|^2$. Using calculus, this is

$$V^T V x = \operatorname*{argmin}_{\tilde{x} \in R} \|x - \tilde{x}\|.$$

To see this, note

$$\|x - Vy\|_2^2 = \|x\|^2 + y^T \underbrace{V^T V}_{I_n} y - 2 y^T V^T x = \|x\|^2 + \|y\|^2 - 2 y^T V^T x.$$

This is strongly cvx. Set gradient in y $= 0$ to get $y^* = V^T x$.

Replacing U, W w/ V, V^T doesn't increase the objective.

$$\sum_i \| x_i - UWx_i \|^2 \geq \sum_i \| x_i - VV^T x_i \|^2 .$$

We had W, U fixed. But since this holds $\forall W, U$, we're done. ∎

So, an equivalent optimization problem is

$$\min_{U \in \mathbb{R}^{d \times n}} \sum_i \| x_i - UU^T x_i \|^2 \qquad (**)$$
$$U^T U = I_n$$

Trick: observe

$$\| x - U^T U x \|^2 = \| x \|^2 - 2 x^T U U^T x + x^T U \underbrace{U^T U}_{I_n} U^T x$$

$$= \| x \|^2 - x^T U U^T x$$

$$= \| x \|^2 - \mathrm{tr}\left(U^T x x^T U \right)$$

↑

Note: If $z \in \mathbb{R}^n$, then $z^T z = \mathrm{tr}(z z^T)$

$$\| \quad \sum z_i^2 \quad = \quad \sum z_i^2 \quad \|$$

Let $A = \sum_{i=1}^m x_i x_i^T$. A is symmetric w/ decomposition

$$A = V D V^T \quad \text{where } D = \mathrm{diag} \text{ & } V^T V = V V^T = I_d$$

Since A is PSD, diag of D is ≥ 0.

So problem is equivalent to

$$\underset{U \in \mathbb{R}^{d \times n}}{\operatorname{argmax}} \ \mathrm{tr}\left(U^T A U \right) \qquad (***)$$
$$U^T U = I_n$$

Thm Let $(x_1, \ldots, x_m) \subset \mathbb{R}^d$. Let $A = \sum_{i=1}^{m} x_i x_i^T$ & let $u_1, \ldots, u_n$ be n eig vectors corresponding to n largest eigvals. Then soln of ~~(**)~~ (*) is to set $u = (u_1, \ldots, u_n)$ & $w = u^T$

pf.

1. Suffices to look at (**)
2. Will show that $\operatorname{tr}(u^T A u) \leq \sum_{j=1}^{n} D_{j,j}$ ← n largest eigvals of A.

   But this value is attained by our selection. So if we show this, we're done.

3. Fix some arbitrary matrix $u \in \mathbb{R}^{d \times n}$ w/ orthonorm cols. Let $B = V^T u$ so $VB = VV^T u = u$. Have

$$u^T A u = \underbrace{B^T V^T}_{= u^T} \underbrace{V D V^T}_{= A} VB = B^T D B$$

Have

$$DB = \begin{pmatrix} D_1 (\text{row } 1 \text{ of } B) \\ \vdots \\ D_d \text{ row } d \text{ of } B \end{pmatrix} \qquad B \in \mathbb{R}^{d \times n}$$

$$\operatorname{tr}(BDB) = \operatorname{tr}\left(\begin{bmatrix} B_1^T & \cdots & B_d^T \end{bmatrix} \begin{bmatrix} D_1 B_1 \\ \vdots \\ D_d B_d \end{bmatrix} \right) \qquad B_1 \in \mathbb{R}^{1 \times n}$$

multiply out 1st block to get $D_1 \sum_{j=1}^{n} B_{1,j}$

$$\operatorname{tr}(\cdot) = D_1 \sum_{j=1}^{n} b_{1,j}^2 + \cdots + D_d \sum_{j=1}^{n} b_{d,j}^2 = \sum_{i=1}^{d} D_{i,i} \sum_{j=1}^{n} b_{i,j}^2$$

4. This is annoying, but have $\sum_{i=1}^{d} \sum_{j=1}^{n} b_{i,j}^2 = n$. by orthonormality, [with order of sum to find (col i) · (col j)]

$$\sum_{j=1}^{n} \underbrace{\sum_{i=1}^{d} b_{i,j}^2}_{(\text{col } j)\cdot(\text{col } j) = 1} = n$$

If we augment B to ~~not first~~ $\tilde{B} = [B, \text{other}]$ w/ $\tilde{B}^T \tilde{B} = I$, then

$$\sum_{j=1}^{d} \tilde{B}_{i,j}^2 = 1 \quad \forall i \implies \sum_{j=1}^{n} B_{i,j}^2 \le 1$$

Hence,

$$\text{tr}(U^T A U) \le \max_{\substack{B \in [0,1]^d \\ \|B\|_1 \le n}} \sum_{i=1}^{d} D_{i,i} B_i.$$

The max on RHS is attained by very B given all mass on n largest d. Hence

$$\text{tr}(U^T A U) = \sum_{i=1}^{n} D_{i,i}' \quad \text{n largest eigenvalues of } A$$

as desired

∎

Perspective 2: An SVD based perspective

Want to solve

$$\min_{\Sigma} \| \Sigma - \underbrace{UW\Sigma}_{\text{rank } n} \|_F^2$$

equivalently, solve

$$\min_{\text{rnk } B \le n,\ \dim B = \dim \Sigma} \| \Sigma - B \|_F^2$$

Thm (Eckart-Young-Mirsky) Let $A = U\Sigma V^T$ be the SVD of A. The best rank n approx to A in Frobenius norm is

$$A_n := \sum_{i=1}^{n} \sigma_i u_i v_i .$$

pf 1. As matter of notation, say $\sigma_1(B) \ge \sigma_2(B) \ge \cdots$ for a matrix.

Groundwork claim:

$$\sigma_{i+j-1}(\Sigma + T) \le \sigma_i(\Sigma) + \sigma_j(T)$$

To do: prove this. Weyl's thm for eigenvalues/singular values?

Would like to show that if rank $B_n = n$, $\| A - A_n \|_F \le \| A - B_n \|_F$

observe

$$\| A - A_n \|_F^2 = \Big\| \sum_{i=n+1}^{r} \sigma_i u_i v_i^T \Big\|^2 = \sum_{i=n+1}^{r} \sigma_i^2(A)$$

(rank of A)

$$\uparrow$$

Let $u = u_i$, $v = v_j$,

$$\| uv^T \|_F = \sum_k \sum_l (u_k v_l)^2 = \underbrace{\sum_k u_k^2}_{=1} \underbrace{\sum_l v_l^2}_{=1}$$

Note that $\sigma_{n+1}(B_n) = 0$. In above claim, let $\gamma = B_n$ & $A = A - B_n$, & $j = n+1$

Get,

$$\sigma_{i+n}(A) \le \sigma_i(A - B_n).$$ Hence,

$$\| A - B_n \|_F^2 \ge \sum_{i=1}^{r} \sigma_i^2(A - B) \ge \sum_{i=n+1}^{r} \sigma_i^2(A) = \| A - A_n \|^2$$

(Exact same reasons as before)

What's the connection between the SVD low rank approx perspective & the "compression" perspective?

First, note that PCA perspective gives us a low dimensional representation of X, namely $U^T X$. The SVD perspective just gives us a low rank matrix - not helpful for simplifying dimensionality issues.

Let's clarify what we have

$$X = U \Sigma V^T$$
$$d \times m \quad d \times d \; d \times m \; m \times m$$

$$U = [u_1, \ldots, u_d]$$
$$V = [v_1, \ldots v_d]$$

To disambiguate, call matrix U from PCA stuff W, so

$WW^T X$ is low-rank approx to X, $W = [u_1, \ldots, u_n]$.

$W: d \times n$, $W^T: n \times d$.

also cols of U in SVD

W compresses compressed cols of X as $W^T x_i$, $WW^T x_i$ brings them back.

So, it appears we have that $WW^T X$ & $\sum_{i=1}^{n} \sigma_i u_i v_i^T$ are both optimal low rank approximations. What gives? Are these the same? ... Yes!

To make this more obvious, call $W = [u_1, \ldots u_n] =: U_n$

Note
$$\sum_{i=1}^{n} \sigma_i u_i v_i^T = [u_1, \ldots, u_n] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}$$
$$1 \times n \qquad n \times n \qquad n \times m$$

But, from PCA,

$$WW^T X = U_n U_n^T X \overset{\text{SVD}}{=} U_n U_n^T U \Sigma V^T$$

$$U_n^T U = \begin{bmatrix} u_1^T \\ \vdots \\ u_n^T \end{bmatrix} [u_1 \ldots u_n \ldots u_d]$$
$$n \times d \quad d \times d$$

$$= [I_n \; 0_{n-n}]$$
$$n \times d$$

$$= U_n [I_n \; 0] \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_n & & \\ & & & \sigma_{n+1} & \\ & & & & \ddots \end{bmatrix} V^T$$

$$= U_n \Sigma_n V_n^T$$

w/ obvious modifications

A couple of quick follow up thoughts:

- There is a nice 2d Python demo of PCA in my notes folder
- Also, a demo using eigenfaces (pulled the demo from the internet)

  Thinking quickly now. I think eigenfaces are just $\sigma_i U_i V_i^T$ for $i = 1$ to k Singular Values/Vectors.