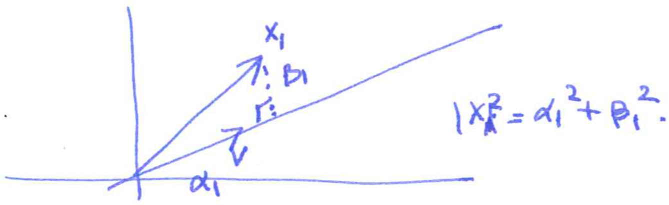
Singular Value Decomposition

We're given a set of points $X = \{x_1, x_2 \cdots x_n\}$ in $\mathbb{R}^D$. They're placed as rows of a matrix $A = \begin{pmatrix} -x_1- \\ -x_2- \\ \vdots \\ -x_n \end{pmatrix}$. We want to represent them using a small # of dimensions, but now not to preserve distances but preserve the "energy".

<u>The 1 dimensional problem</u>: find a 1-dimensional subspace such that the sum of projections (squared) is maximized — or the sum of distances squared is minimized. These are equivalent (by pythagoras).



So want v st. $\sum_i (x_i \cdot v)^2$ is maximized st $|v| = 1$. but $\sum (x_i \cdot v)^2 = \|Av\|^2$.

first (right) singular vector $\quad v_1 := \underset{v}{\arg\max} \dfrac{\|Av\|^2}{\|v\|^2} = \underset{\|v\|=1}{\arg\max} \|Av\|^2$.

first singular value $\quad \sigma_1(A) := \max_{|v|=1} \|Av\| = \|Av_1\|$.

$\sigma_1^2 =$ sum of squared projections of points on v_1.

Suppose want 2-d subspace in some greedy fashion :— let's find

$$v_2 = \underset{\|v\|=1,\ v \perp v_1}{\arg\max} \|Av\| . \quad \text{and} \quad \sigma_2(A) = \max_{\|v\|=1,\ v\perp v_1} \|Av\|.$$

similarly:

$$v_i = \underset{\|v\|=1,\ v \perp v_1, v_2 \cdots v_{i-1}}{\arg\max} \|Av\| \quad \text{and} \quad \sigma_i(A) = \max_{\|v\|=1,\ v\perp v_1 \cdots v_{i-1}} \|Av\|.$$

<u>Observe</u>: we've just defined things greedily. Find v_1, then v_2, then v_3 . . . . then v_k.

What if the best 2-dim subspace did some global optimization nongreedily ?

Thankfully, things are pretty good for us!

Claim: Suppose $A \in \mathbb{R}^{n \times d}$ with singular vectors $v_1, v_2 \dots v_r$. ← means $\max_{v \perp v_1 \dots v_r} \|Av\| = 0$.

Let $V_k = \text{span}(v_1 \dots v_k)$. $\forall k \leq r$.

then V_k is best-fit k-dimensional subspace for A.

(i.e. sps S is another k-dim subspace spanned by $w_1 \dots w_k$ (orthonormal basis)

$$\Rightarrow \sum_{i=1}^{k} \|Aw_i\|^2 \leq \sum_{i=1}^{k} \|Av_i\|^2 .)$$

Pf: . Base case $k=1$ (trivially true)

- $k=2$: v_1, v_2 are sing vectors. Say S is another subspace. Let $w_1 w_2$ span S.

  Choose $w_2 \perp v_1$ in the subspace, and then w_1 accordingly.

  . $\|Aw_2\|^2 \leq \|Av_2\|^2$. by choice of v_2. ✓.

  . $\|Aw_1\|^2 \leq \|Av_1\|^2$ by choice of v_1.

- k general. By induction V_{k-1} is best fit k-dim subspace.

  Choose $w_1 \dots w_k$ to span S such that $w_k \perp v_1 \dots v_{k-1}$.

  Again same argument.

∎

Fact: $\sum \sigma_i^2 = \|A\|_F = \sum_{ij} a_{ij}^2$.

Pf: $\sum_i \sigma_i^2 = \sum_i \|Av_i\|^2 = \sum_i \sum_j \langle a_j, v_i \rangle^2 = \sum_j \sum_i \langle a_j, v_i \rangle^2$

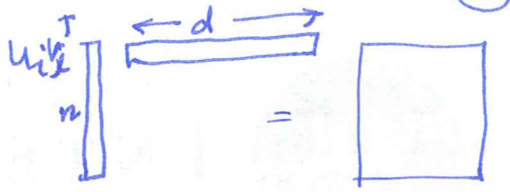
but v_i form an orthonormal basis for the row space of A.

$$= \sum_j \|a_j\|^2 = \sum_{jk} a_{jk}^2 = \|A\|_F .$$ ∎

Now define: $u_i = \dfrac{Av_i}{\|Av_i\|} = \dfrac{Av_i}{\sigma_i}$. clearly unit vectors. (also in fact orthonormal)

$u_i :=$ ith left singular vector of A.

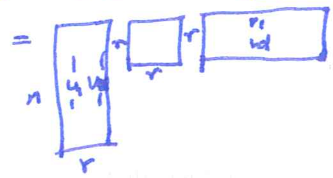
$$= \underset{\substack{u \perp u_1 \dots u_{i-1} \\ |u|=1}}{\arg\max} \|uA\|$$ (by the way) much like the right singular vector.



Claim: $A = \sum_i \sigma_i u_i v_i^T$

(BTW: if $U = \begin{pmatrix} | & | & & | \\ u_1 & u_2 & \dots & u_r \\ | & | & & | \end{pmatrix}$

$D = \begin{pmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \ddots \\ & & & \sigma_r \end{pmatrix}$ $\quad V = \begin{pmatrix} | & | & & | \\ v_1 & v_2 & \dots & v_r \\ | & | & & | \end{pmatrix}$ then $\sum_i \sigma_i u_i v_i^T = UDV^T$.



Pf: [**SubClaim:** A and B are same iff $Av = Bv$ $\forall v \in \mathbb{R}^n$.

Pf: $\Rightarrow$ trivial $\Leftarrow$ sps not then choose $v = e_i$. where differ in i^{th} column.

[**Subclaim:** A and B are same iff $Av = Bv$ $\forall v$ in basis. Pf write each v in this basis.

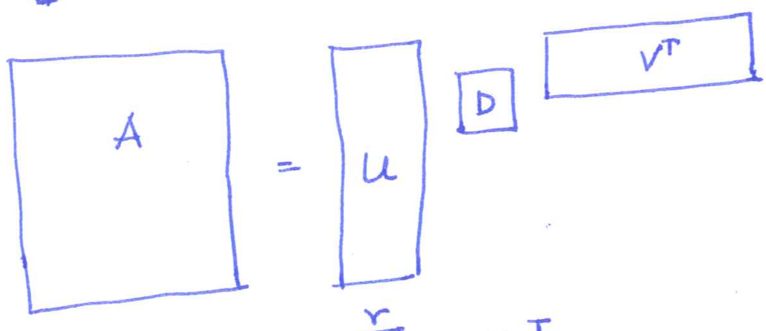
Now: consider the basis of $v_1, v_2 \dots v_r, \underbrace{v_{r+1} \dots v_d}_{\text{extend to orthonormal basis}}$

for any $j \in [r]$ $\left(\sum_i \sigma_i u_i v_i^T\right) . v_j = \sigma_j u_j = A v_j$ (by defn of u_j)

$\boxed{\Rightarrow A \text{ and } \sum \sigma_i u_i v_i^T \text{ are identical.}}$

for $j \notin [r]$ $A v_j = 0 = \left(\sum_i \sigma_i u_i v_i^T\right) v_j$

∎



$= \sum_{i=1}^{r} \sigma_i u_i v_i^T$

the columns of U & V are orthonormal (haven't proved it for columns of U yet $\rightarrow$ do it in HW)

Claim: Suppose $A = \check{U} \check{D} \check{V}^T$ for matrices $\check{U}, \check{D}, \check{V}$ ~~for~~ where $\check{D}$ is diagonal, $\check{U}, \check{V}$ have orthonormal columns. then $\check{U} = U$, $\check{D} = D$, $\check{V} = V$ (apart from degeneracies due to subspaces). (say all singular values are distinct then it is indeed unique.)

Thm: [Eckert Young]

Sps $\sum_{i=1}^{r} \sigma_i u_i v_i^T = A_{n\times d}$. For any $k \le r$, define $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$.

Then for any matrix B of rank at most k,

$$\boxed{\|A - B\|_F \geqslant \|A - A_k\|_F.}$$

Also: define the 2-norm or spectral norm of a matrix A to be $\|A\|_2 = \max_{\|v\|=1} \|Av\|$.
$= \max$ singular value of A.

$$\Rightarrow \boxed{\|A - B\|_2 \geqslant \|A - A_k\|_2}$$

See HW exercises.

~~The top k singular values give the best k dimensional subspace.~~

Lem: the rows of A_k are the projections of rows of A onto $V_k = \text{span}(v_1, \dots v_k)$.

Pf: if $\bar{a}$ is a row of A, a's projection is given by $\sum_{i=1}^{k} (a, v_i) . v_i^T$

$\Rightarrow$ All these projections are given by $\sum_{i=1}^{k} (Av_i) . v_i^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T = A_k$.

$\Rightarrow \|A - A_k\|_F$ is the remaining mass/energy

Fact: for a (square) symmetric matrix B, can find a basis of eigenvalues

x_i st $Bx_i = \lambda_i x_i \quad \forall i = 1..n$. $\quad x_i$ are orthonormal.

$\Rightarrow B = X \Lambda X^T \qquad X = \begin{pmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{pmatrix} \quad \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$.

$= \sum_i \lambda_i x_i x_i^T \qquad$ (again, unique upto rotation within subspaces).

Now: sps matrix is general $A_{n\times d}$. Then

$AA^T = (UDV^T)(UDV^T)^T = UDV^TVDU^T = UD^2U^T \Leftarrow$ eigenvalues are σ_i^2. eigenvectors are u_i.

$A^TA = VD^2V^T \leftarrow$ eigenvectors are now v_i

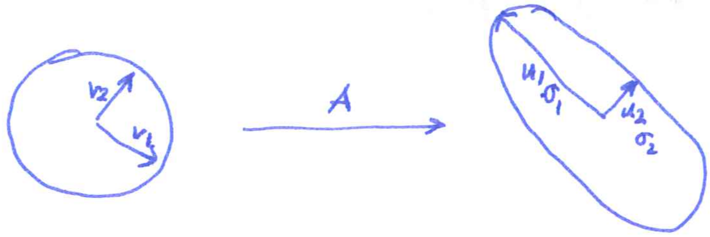
Moreover: having found (left) eigenvalues & eigenvectors (say) $AA^T = X\Lambda X^T$ know $U = X$. $D^2 = \Lambda$.

and set $V^T = D^{-1}U^TA$.

One view: v_i's are the directions along which A has most energy.

Another view: $A = UDV^T \Rightarrow AV = UD$.

cols of both U & V are orthonormal. So A takes these orthonormal cols of V into scaled orthonormal cols of U.



Example: hand writing recognition; face compression, etc.

~~Learning Mixtures of Gaussians~~

Topic Models (naive version)

A = documents $\times$ words = UDV^T

= docs ( ) ( ) ( ) topics

topics — weights of topics — ← words →

Learning Gaussian Mixtures: [Vempala Wang]

Given k gaussians, with weights w_i, want to (a) cluster the points into k clusters, and (b) find the mean, variance, weight etc of the gaussians.

(this is fine given the clustering).

Known: if inter-center distances are at least $\Omega(d^{1/4})$ then can cluster.

So: idea: find the SVD of space, with enough samples, the top k-dimensions ~~the~~ give us the subspace containing the k centers. Now: ~~this space~~ projecting down onto this space still gives us k Gaussians. (b/c of spherical symmetry) And hence a separation of $\Omega(k^{1/4})$ suffices

Subspace Embeddings:

Suppose $W \subseteq \mathbb{R}^n$ is a linear subspace: an ε-subspace embedding into m dimensions ~~ensures that~~ $\Pi : \mathbb{R}^n \to \mathbb{R}^m$ (Π is a $m \times n$ matrix) ensures that

$$1-\epsilon \leq \frac{\|\Pi x\|_2}{\|x\|_2} \leq 1+\epsilon \qquad \forall x \in W.$$

So it maintains all ~~distances~~ vectors in thin subspace.

[Fact: JL applied to a fine enough net of points in W will give us such a property.]

[Suppose W is given by column space of matrix $A \in \mathbb{R}^{n \times d}$. Then can use SVDs to find an embedding of W into $\mathbb{R}^d$. How?

write $A = \underset{n\times r}{U}\,\underset{r\times r}{D}\,\underset{r\times d}{V^T}$ $(r \leq d)$ hence $U^T A = D V^T$

Any vector $x \in W$ is a linear combination of columns of A. And U ~~has~~ has orthonormal columns so $U^T A$ rotates the columns of A into a r-dim subspace. $\|U^T x\| = \|x\| \;\; \forall x \in \text{span}(\text{col}(A))$

N.b. this is a 0-subspace embedding] (exact)

Least Squares Regression:

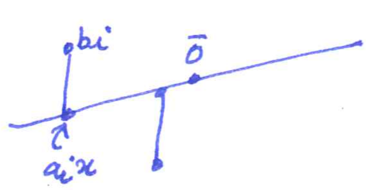
Given $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$ want an $x \in \mathbb{R}^d$ st $\min \|Ax - b\|$.

$$= \min \textstyle\sum_i (a_i x - b_i)^2.$$

($a_i x$ defines the subspace)

(Useful for solving overconstrained systems.)

Observe that Ax is in column span of $A = UDV^T$
so Ax^* is the projection of b onto this columnspan.



$\text{Proj}(b) = UU^T b$ (onto colspan(A)) because U is an orthonormal basis for colspace(A).

(alternatively, proj onto colspace(A) $= A(A^T A)^{-1} A^T$
$= UDV^T (VDU^T UDV^T)^{-1} (UDV^T)^T$
$= UU^T$)

hence $x^* = VD^{-1}U^T(UU^T b) = VD^{-1}U^T b$.

$\Rightarrow$ least squares regression in SVD(n,d) time.

BTW: If $A = UDV^T$

then $VD^{-1}U^T$ is called the pseudoinverse of A (D^{-1} is really D^+, zeroes stay 0)

denoted by A^+

$\boxed{\text{and it ~~maps~~ ensures that for all vectors in the columnspace of } A, A^+ \text{ acts like an inverse!}}$

Formally: Restricted to Colspace(A^T) and Colspace(A) respectively, the operators A and A^+ are inverses of each other.

Pf: sps $x \in$ Colspace(A^T) $\Rightarrow$ $x = A^T y$ for some y.

then
$$
\begin{aligned}
A^+Ax - x &= A^+A(A^Ty) - A^Ty \\
&= VD^+U^T \cdot UDV^T(VDU^Ty) - VDU^Ty \\
&= VD^+DDU^Ty - VDU^Ty \\
&= V\underbrace{(D^+D - I)D}_{\text{this multiplication gives zero.}}U^Ty = 0
\end{aligned}
$$
∎

(Another motivation, of course, from the least squares)

The pseudoinverses play an important role in Laplacians.

Recall: $L(G) = \text{diag(degree vector)} - A$

but $L(G).\mathbb{1} = 0$. since the sum of all rows gives 0.

For a connected graph, the eigen~~vectors~~ values of $L(G)$ are $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_n$

$\uparrow$ corresponding to eigenvector $\mathbb{1}$.

So can solve linear systems of the form $Lx = b$ as long as $b \perp \mathbb{1}$

i.e. $\sum_v b_v = 0$.

by $x = L^+b$. (gives potentials corresponding to electrical flow demands)

Fact: Also, can do calculations and show: $L^+L = LL^+ = I_n - \frac{1}{n}J$.

$\Rightarrow$ for any vector $b \perp \mathbb{1}$, $LL^+b = b$ since $bJ = 0$.