2 SVD

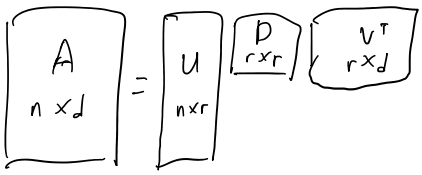
Friday, January 10, 2020 11:10 AM

Last time we left off with the Johnson-Lindenstrauss Lemma, which tells us that a projection via a Gaussian matrix gives a dimensionality reduction that preserves pairwise distances.

This time we will be covering dimensionality reductions that preserve variance of a dataset. We will also approach it from the other direction. Instead of specifying in advance how much variance to preserve and computing the number of dimensions needed, we simply maximize the amount of variance preserved for a given number of dimensions.

Define: The Singular Value Decomposition (SVD) of an $n \times d$ matrix A is a factorization of the form

$$A = UDV^T$$ where U is an orthogonal (unitary) $n \times r$ matrix
D is a diagonal $r \times r$ matrix with nonnegative real entries
V is an orthogonal $d \times r$ matrix.
(r is the rank of the matrix)



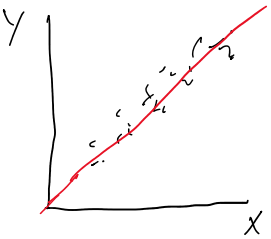
Note: this factorization always exists, unlike e.g. eigendecompositions.

cf. $M = QDQ^{-1}$
(M: $n \times n$ matrices; Q: square matrix of eigenvectors; D: diagonal)

But instead of just defining the SVD and proving properties, let's motivate this definition.

Suppose we have n points in a d-dimensional subspace.

Let's try to find the best-fitting 1-dim subspace (i.e. fit a line).

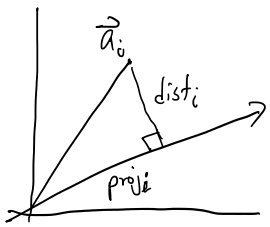


One idea is to minimize the squared distance from the line.

How should we measure squared distance?

- Linear regression: measure distance only in a privileged predicted var.
- SVD: measure ordinary Euclidean distance from the line.

Consider projecting a single point $\vec{a}_i$ to a line.



Consider projecting a single point $\vec{a}_i$ to a line.

$|\vec{a}_i|^2 = dist_i^2 + proj_i^2$.

But $|\vec{a}_i|^2$ is ind. of the line chosen, so is effectively constant for minimizing $dist_i^2$ is equivalent to maximizing $proj_i^2$.

This is true even for the sums of many distances or projection lengths, and also if you have a higher dimensional subspace.

More rigorously

Let $\vec{a}_1, \dots, \vec{a}_n \in \mathbb{R}^d$, and let $A = \begin{bmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_n \end{bmatrix} \in \mathbb{R}^{n \times d}$.

Let $\vec{v} \in \mathbb{R}^d$ be a unit vector, defining a line in $\mathbb{R}^d$.
Then the projection of a vector $\vec{a}_i$ has length $|\vec{a}_i \cdot v|$.
Then $\sum_{i=1}^{n} |\vec{a}_i \cdot v|^2 = |A\vec{v}|^2$.

Define: The first singular vector $\vec{v}_1$ of A is

$$v_1 = \arg\max_{|v|=1} |A\vec{v}|.$$

(We arbitrarily pick one if there are multiple, e.g. $-v_1$).
(and will always arbitrarily pick when there's a tie in today's lecture)

v_1 defines the best-fit line in terms of minimizing squared distance or maximizing squared projected lengths.

Define: The first (right) singular value $\sigma_1(A) = |A\vec{v}_1|$

Note that $\sigma_1^2 = \sum_{i=1}^{n} (\vec{a}_i \cdot \vec{v}_1)^2$.

(v_1 is the line capturing the most variance of $\vec{a}_1, \dots, \vec{a}_n$.)

This gives the best-fit 1D subspace. What about higher-dimensional subspaces?

Let's try the greedy approach.

Define: The second (right) singular vector v_2 of A is

$$\vec{v}_2 = \arg\max_{\substack{\vec{v} \perp \vec{v}_1 \\ |\vec{v}|=1}} |A\vec{v}|.$$

$\sigma_2(A) = |A\vec{v}_2|$ is the 2nd singular value.

v_2 $\operatorname{argmax}_{\vec{v}\perp\vec{v}_1,\ |\vec{v}|=1} |A\vec{v}|$ v_2 $|Av_2|$ is the 2nd singular value.

i.e. We look for a second direction to explain as much of the remaining variance.

Define: The kth right singular vector $\vec{v}_k$ of A is

$$\vec{v}_k = \operatorname{argmax}_{\vec{v}\perp\vec{v}_1,\dots,\vec{v}\perp\vec{v}_{k-1},\ |\vec{v}|=1} |A\vec{v}|$$

and $\sigma_k(A) = |A\vec{v}_k|$ is the kth singular value.

Does this work? Let r be the rank(A).

Thm 3.1 Let $A \in \mathbb{R}^{n\times d}$ with right singular vectors $\vec{v}_1, \dots, \vec{v}_r$. For $1 \le k \le r$, let V_k be the subspace spanned by $\vec{v}_1, \dots, \vec{v}_k$. For each k, V_k is the best-fit k-dimensional subspace for A.

proof. By induction.

Trivially true for base case $k=1$ by definition.

Assume V_{k-1} is a best-fit $(k-1)$-dimensional subspace.

Suppose W is a best-fit k-dimensional subspace.

Choose an orthonormal basis $\vec{w}_1, \dots, \vec{w}_k$ of W so that $\vec{w}_k$ is perpendicular to V_{k-1}.

(Possible because $\dim(W) = k$ and $\dim(V_{k-1}) = k-1$)

Then $|A\vec{w}_1|^2 + \cdots + |A\vec{w}_{k-1}|^2 \le |A\vec{v}_1|^2 + \cdots + |A\vec{v}_{k-1}|^2$

because V_{k-1} is an optimal $k-1$ dim subspace.

Furthermore, $|A\vec{w}_k|^2 \le |A\vec{v}_k|^2$ because $\vec{v}_k = \operatorname{arg\,max}_{v\perp V_{k-1},\ |v|=1} |Av|^2$.

Thus, $|A\vec{w}_1|^2 + \cdots + |A\vec{w}_k|^2 \le |A\vec{v}_1|^2 + \cdots + |A\vec{v}_k|^2$,

proving the V_k is at least as good as W for any W. ∎

Lemma: Let the Frobenius norm be defined by $\|A\|_F = \sqrt{\sum_{j,k} a_{jk}^2}$.

Then $\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2(A)$.

proof. Let $\vec{a}_1, \dots, \vec{a}_n \in \mathbb{R}^d$ be the rows of A.

Since $\vec{v}_1, \dots, \vec{v}_r$ is an orthonormal basis, $\vec{a}_j = (\vec{a}_j \cdot \vec{v}_1)\vec{v}_1 + \dots + (\vec{a}_j \cdot \vec{v}_r)\vec{v}_r$.

$$|\vec{a}_j|^2 = (\vec{a}_j \cdot \vec{v}_1)^2 + \dots + (\vec{a}_j \cdot \vec{v}_r)^2$$ (Pythagorean theorem)

$$\sum_{j=1}^{n} |\vec{a}_j|^2 = \sum_{j=1}^{n}\sum_{i=1}^{r} (\vec{a}_j \cdot \vec{v}_i)^2 = \sum_{i=1}^{r}\sum_{j=1}^{n} (\vec{a}_j \cdot \vec{v}_i)^2 = \sum_{i=1}^{r} |A\vec{v}_i|^2 = \sum_{i=1}^{r} \sigma_i^2(A).$$

The singular vectors we've been working with are the right singular vectors

Def. The left singular vectors are defined as $\vec{u}_1, \dots, \vec{u}_r$, where

$$\vec{u}_i = \frac{1}{\sigma_i(A)} \cdot A\vec{v}_i .$$

Note that $\vec{u}_i = \underset{\substack{\vec{u} \perp \vec{u}_1, \dots, \vec{u} \perp \vec{u}_{i-1} \\ |\vec{u}|=1}}{\arg\max} |u^T A|$, and are also orthogonal. (proof later)

Now we are ready to reintroduce the matrix decomposition.

Lemma 3.3 $A = B \iff A\vec{v} = B\vec{v} \quad \forall \vec{v}$. (trivial)

$\iff A\vec{e}_1 = B\vec{e}_1, \dots, A\vec{e}_r = B\vec{e}_r$ where $\vec{e}_1, \dots, \vec{e}_r$ is a basis of $\mathrm{Im}(A)$.

Theorem 3.4 Let $A \in \mathbb{R}^{n \times d}$ with right-singular vectors $\vec{v}_1, \dots, \vec{v}_r$, left-singular vectors $\vec{u}_1, \dots, \vec{u}_r$, and singular values $\sigma_1, \dots, \sigma_r$.

Let $V = [\vec{v}_1 \ \cdots \ \vec{v}_r] \in \mathbb{R}^{d \times r}$

$U = [\vec{u}_1 \ \cdots \ \vec{u}_r] \in \mathbb{R}^{n \times r}$

$D = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix} \in \mathbb{R}^{r \times r}$

$$D = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{bmatrix} \in \mathbb{R}^{r \times r}$$

Then $A = UDV^T = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T$.

proof. $\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T \vec{v}_j = \sum_{i=1}^{r} \sigma_i \vec{u}_i \underbrace{(\vec{v}_i \cdot \vec{v}_j)}_{\substack{=0 \text{ if } i \neq j \\ =1 \text{ if } i=j}} = \sigma_j \vec{u}_j = A\vec{v}_j$. ▨

This is also why we defined the left singular values by reference to the right singular values. Had we not done so, the lack of guaranteed uniqueness would've messed us up in the matrix decomposition. If singular values are equal, then subspace where any orthonormal basis can be chosen.

Theorem: Let A be a rank r matrix. The left singular vectors of A $\vec{u}_1, \ldots, \vec{u}_r$ are orthogonal.

proof. By induction, on r.

Clearly $\{\vec{u}_1\}$ is orthogonal. (base case) ($r=1$)

Consider the matrix $B = A - \sigma_1 \vec{u}_1 \vec{v}_1^T$. (removing the first component of the SVD)

$$B\vec{v}_1 = A\vec{v}_1 - \sigma_1 \vec{u}_1 \vec{v}_1^T \vec{v}_1 = 0.$$

Say $\vec{z}$ is the first right singular vector of B.

$\vec{z} \perp \vec{v}_1$ because if not $\left| B \frac{\vec{z} - (\vec{z} \cdot \vec{v}_1)\vec{v}_1}{|\vec{z} - (\vec{z} \cdot \vec{v}_1)\vec{v}_1|} \right| = \frac{|B\vec{z}|}{|\vec{z} - (\vec{z} \cdot \vec{v}_1)\vec{v}_1|} > |B\vec{z}|$, which contradicts the argmax def. of first sing. vector.

But for any $\vec{v} \perp \vec{v}_1$, $B\vec{v} = A\vec{v}$.

Thus, the first right singular vector of B is a second right singular vector of A.

Repeating this argument all the singular vectors of B are the sing. vectors of A except v_1.

Thus, by the induction hypothesis, $\vec{u}_2, \ldots, \vec{u}_r$ are orthogonal.

Need to prove $\vec{u}_1$ is orthogonal to all other $\vec{u}_i$.

Suppose not, that for some $i \geq 2$, $\vec{u}_1 \cdot \vec{u}_i \neq 0$.

WLOG, $\vec{u}_1 \cdot \vec{u}_i = \delta > 0$.

For $\varepsilon > 0$, $\left| A\left(\frac{\vec{v}_1 + \varepsilon \vec{v}_i}{|\vec{v}_1 + \varepsilon \vec{v}_i|} \right) \right| = \left| \frac{\sigma_1 \vec{u}_1 + \varepsilon \sigma_i \vec{u}_i}{\sqrt{1+\varepsilon^2}} \right|$

length of vector is at least length along one component

$$\geq \vec{u}_1 \cdot \left(\frac{\sigma_1 \vec{u}_1 + \varepsilon \sigma_i \vec{u}_i}{\sqrt{1+\varepsilon^2}} \right)$$

$$= \left(\sigma_1 + \varepsilon \sigma_i \vec{u}_1 \cdot \vec{u}_i \right) \left(1 - \frac{\varepsilon^2}{2} + O(\varepsilon^4) \right)$$

(Taylor series of $\frac{1}{\sqrt{1+x^2}}$)

$$> \left(\sigma_1 + \varepsilon \sigma_i \delta \right) \left(1 - \frac{\varepsilon^2}{2} \right)$$

$$> \sigma_1 - \frac{\varepsilon^2}{2}\sigma_1 + \varepsilon \sigma_i \delta - \frac{\varepsilon^3}{2}\sigma_i \delta$$

$$= \sigma_1 + \varepsilon \left(\sigma_i \delta - \frac{\varepsilon}{2}\sigma_1 - \frac{\varepsilon^2}{2}\sigma_i \delta \right) > \sigma_1$$ for sufficiently small ε.

($\varepsilon < \min(\delta, 1)$)

Contradiction, because σ_1 is the largest singular value.

Thus, all $\vec{u}_i$'s are orthogonal!

Low-rank approximations

We will prove that under both the Frobenius & spectral norms, the SVD provides the best rank-k approximation of a matrix.

Lemma 3.5 The rows of $A_k = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T$ are the projections of the rows of A onto the subspace V_k spanned by the first k singular vectors $\vec{v}_1, \ldots, \vec{v}_k$ of A.

proof Let $\vec{a}$ be an arbitrary row vector.

The proj. of $\vec{a}$ onto V_k is $\sum_{i=1}^{k} (\vec{a} \cdot \vec{v}_i) \vec{v}_i^T$ (because the v_i are orthonormal)

Let $A = \begin{bmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_n \end{bmatrix}$. Then

$$A_k = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T = \sum_{i=1}^{k} A \vec{v}_i \vec{v}_i^T$$

$$= \sum_{i=1}^{k} \begin{bmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_n \end{bmatrix} \vec{v}_i \vec{v}_i^T = \sum_{i=1}^{k} \begin{bmatrix} \vec{a}_1 \vec{v}_i \vec{v}_i^T \\ \vdots \\ \vec{a}_n \vec{v}_i \vec{v}_i^T \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{k} (\vec{a}_1 \cdot \vec{v}_i) \vec{v}_i^T \\ \vdots \\ \sum_{i=1}^{k} (\vec{a}_n \cdot \vec{v}_i) \vec{v}_i^T \end{bmatrix}$$

Theorem 3.6 For any matrix B of rank at most k,

$\| A - A_k \| \leq \| A - B \|$

Theorem 3.6 For any matrix B of rank at most k,

$$\|A - A_k\|_F \leq \|A - B\|_F,$$

where $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T = A_k$.

proof. Let $B = \underset{B' \mid \text{rank}(B') \leq k}{\text{argmin}} \|A - B'\|_F^2$.

Let V be the space spanned by the rows of B. $\dim(V) \leq k$.
Let $\vec{b}_i$ be a row of B.
Let $\vec{a}_i'$ be the projection of $\vec{a}_i$ onto V.
Suppose $\vec{b}_i \neq \vec{a}_i'$.
Then we can replace $\vec{b}_i$ with $\vec{a}_i'$ in B, forming B',
But $\vec{a}_i' \in V$, so the rows of B' are still contained in V, so $\text{rank}(B') \leq k$.
However $\|A - B'\|_F^2 \leq \|A - B\|_F^2$ because the proj. of $\vec{a}_i$ onto V is the closest any vector in V gets to $\vec{a}_i$.

Repeating this argument for all rows, and we can conclude WLOG that all rows of B are projections of rows of A onto a k-dim. subsp V.

But we proved earlier that A_k minimizes the sum of squared distances of rows of A to a k-dim subsp., so

$$\|A - A_k\|_F \leq \|A - B\|_F.$$

Define: The spectral (2-norm) of a matrix A is

$$\|A\|_2 = \max_{|\vec{x}| \leq 1} |A\vec{x}|.$$

Note: $\|A\|_2 = \sigma_1(A)$

Lemma 3.8: $\|A - A_k\|_2^2 = \sigma_{k+1}^2$

Let $A = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T$ be the SVD of A.
k

From our earlier proof that the left singular vectors are orthogonal, we showed that the SVD of $A - A_1$ is just all but but the first singular value/vectors

The same argument can be extended to show that the SVD of $A - A_r$

First k singular values/vectors

The same argument can be extended to show that the SVD of $A - A_k$ is all but the first k singular values/vectors

Thus, the SVD of $A - A_k$ is $\sum_{i=k+1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T$, so $\|A - A_k\|_2 = \sigma_1(A - A_k) = \sigma_{k+1}$. ▨

Thm 3.9. Let $A \in \mathbb{R}^{n \times d}$. For any matrix B of rank at most k,

$$\|A - A_k\|_2 \leq \|A - B\|_2.$$

Proof. If $\text{rank}(A) \leq k$, then $A - A_k = 0$, so $\|A - A_k\|_2 = 0$, making the thm trivially true.

Assume $\text{rank}(A) > k$. Then $\|A - A_k\|_2^2 = \sigma_{k+1}^2$.

$\dim(\text{Null}(B)) \geq d - k$.

Let $\vec{v}_1, \ldots, \vec{v}_{k+1}$ be the first $k+1$ sing. vec. of A. $(\vec{v}_i \in \mathbb{R}^d)$

Then $\exists\, \vec{z} \neq 0$ s.t. $\vec{z} \in \text{Null}(B) \cap \text{span}(\vec{v}_1, \ldots, \vec{v}_k)$. WLOG, say $|\vec{z}| = 1$.

Then $\|A - B\|_2^2 \geq |(A - B)\vec{z}|^2 = |A\vec{z}|^2$

$$= \left| A \sum_{i=1}^{n} (\vec{z} \cdot \vec{v}_i) \vec{v}_i \right|^2 = \left| \sum_{i=1}^{n} (\vec{z} \cdot \vec{v}_i) \sigma_i \vec{u}_i \right|^2 \quad \text{(by orthogonality + Pythagoras)}$$

$$= \sum_{i=1}^{n} \sigma_i^2 (\vec{z} \cdot \vec{v}_i)^2 = \sum_{i=1}^{k+r} \sigma_i^2 (\vec{z} \cdot \vec{v}_i)^2 \geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} (\vec{z} \cdot \vec{v}_i)^2 = \sigma_{k+1}^2$$

↑ because $|\vec{z}| = 1$.

$$\|A - B\|_2^2 \geq \sigma_{k+1}^2 = \|A - A_k\|_2^2.$$

SVD vs. eigendecomposition

Lemma 3.10 $A\vec{v}_i = \sigma_i \vec{u}_i$ and $A^T \vec{u}_i = \sigma_i \vec{v}_i$.

proof. By def. $\vec{u}_i = \frac{1}{\sigma_i} A \vec{v}_i$; proving the first claim.

By SVD, $A^T \vec{u}_i = \left(\sum_{j=1}^{r} \sigma_j \vec{u}_j \vec{v}_j^T \right)^T \vec{u}_i = \sum_{j=1}^{r} \sigma_j \vec{v}_j \vec{u}_j^T \vec{u}_i = \sum_{j=1}^{r} \sigma_j \vec{v}_j (\vec{u}_j \cdot \vec{u}_i) = \sigma_i \vec{v}_i$. ▨

Corollary $\vec{v}_1, \ldots, \vec{v}_r$ are eigenvectors of $A^T A$ with eigenvalues σ_i^2.

$\vec{u}_1, \ldots, \vec{u}_r$ are eigenvectors of AA^T with eigenvalues σ_i^2.

$A^T A$ is positive semi-definite. $\left(x^T A^T A x = (Ax)^T (Ax) = |Ax|^2 \right)$