The Singular Value Decomposition

Rob McCulloch

1. Singular Value Decomposition
2. Column space, Row space, and rank
3. Linear is just a bunch of linear
4. Reduced Form
5. SVD and Least Squares
6. SVD and Spectral
7. Condition Number of a Matrix
8. Moore Penrose Generalized Inverse
9. Matrix Approximation

1. Singular Value Decomposition

This is a key decomposition that applies to *any* matrix A, $m \times n$.

SVD:

Let A be $m \times n$.
Then there are

- orthogonal U, $m \times m$
- orthogonal V, $n \times n$
- diagonal Σ

such that

$$\underset{m\times n}{A} = \underset{m\times m}{U} \ \underset{m\times n}{\Sigma} \ \underset{n\times n}{V^T}$$

For integer r,

$\sigma_{11} \geq \sigma_{22} \ldots \geq \sigma_{rr} > 0$,
and $\sigma_{jj} = 0, j > r$, $\sigma_{ij} = 0, i \neq j$.

$$\Sigma = \begin{bmatrix} \begin{matrix} \sigma_{11} & 0 & \cdots & 0 \\ & \sigma_{22} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{rr} \end{matrix} & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{bmatrix}$$

2. Column space, Row space, and rank

We will see that the first r columns of U are an orthonormal basis for the column space of A.

We will see that the first r columns of V are an orthonormal basis for the row space of A.

Hence, the column rank = the row rank, which is then the rank.

So, r is the rank of the matrix.

Note:

A is $m \times n$, $A = [a_1, a_2, \ldots, a_n]$, $a_i \in R^m$.

The column space is the span of the a_i which is the set $\{Ab, b \in R^n\}$.

Suppose B is $n \times n$ invertible.

Then

$$\{Ab, b \in R^n\} \ = \ \{ABb, b \in R^n\}$$

so that the column space of A is the same as the column space of AB.

Similar result for premultiplying be an invertible matrix for the row space.

So, since V' is invertible, the column space of A is the column space of $U\Sigma$.

$$U\Sigma = [u_1, u_2, \ldots u_m] \begin{bmatrix} \sigma_{11} & 0 & \cdots & & & & \\ 0 & \sigma_{22} & & & & 0 & \\ \vdots & & \ddots & & & & \\ & & & \sigma_{rr} & 0 & & \\ & 0 & & & \ddots & & \\ & & & & & 0 & \\ 0 & & & & & & \end{bmatrix}$$

$$= \{u_1\sigma_{11}, u_2\sigma_{22}, \ldots u_r\sigma_{rr}, 0, 0, \ldots 0\}$$

Hence $[u_1, u_2, \ldots, u_r]$ is an orthonormal basis for the column space of A.

The column rank of A is r.

$[u_{r+1}, \ldots, u_m]$ is an orthonormal basis for the subspace perpendicular to the column space.

Similarly, the first r columns of V are an orthonormal basis for the row space of A.

So, the row rank $=$ the column rank $=$ the rank, all which are equal to r in our notation.

The $i = r + 1, \ldots, n$ columns of V form a basis for the subspace of R^n orthogonal to the row space.

3. Linear is just a bunch of linear

$$A_{m\times n} = U_{m\times m}\,\Sigma_{m\times n}\,V^T_{n\times n}$$

$U = [u_1, u_2, \dots u_m] \quad V = [v_1, v_2, \dots v_n]$

$$A v_j = U \Sigma V^T v_j = U \Sigma \begin{bmatrix} \langle v_1, v_j\rangle \\ \langle v_2, v_j\rangle \\ \vdots \\ \langle v_n, v_j\rangle \end{bmatrix}$$

$$= U \Sigma e_j$$

$j > r \Rightarrow \Sigma e_j = 0$

$\Rightarrow A v_j = 0$

$1 \le j \le r \qquad \Sigma e_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \sigma_{jj} \\ \vdots \\ 0 \end{bmatrix} - j\text{th}$

$A v_j$

$= U \Sigma e_j$

$= \sigma_{jj} u_j$

A truly remarkable result !!!

$$Av_j = \sigma_{jj}\, u_j,\ 1 \le j \le r,$$

$$Av_j = 0,\ (r+1) \le j \le n.$$

$$A : R^n \to R^m.$$

- $N(A) = \{x \in R^n,\ s.t.\ Ax = 0\}$, a subspace of dim $n - r$ with orthonormal basis $\{v_{r+1}, \ldots, v_n\}$.
- $R(A) = \{Ax, x \in R^n\}$, a subspace of dim r with orthonormal basis $\{u_1, u_2, \ldots, u_r\}$.

$A : \mathbb{R}^n \to \mathbb{R}^m \qquad A$ is $m \times n$

$x \in \mathbb{R}^n \qquad x = \sum_{j=1}^{n} \tilde{x}_j v_j \qquad (\tilde{x}_j = \langle v_j, x \rangle)$

$$A(x) = \sum_{j=1}^{n} \tilde{x}_j A v_j = \sum_{j=1}^{r} \tilde{x}_j \sigma_{jj} u_j$$

$y = Ax = \sum_{j=1}^{m} \tilde{y}_j u_j \qquad (\tilde{y}_j = \langle u_j, y \rangle)$

In terms of ~~the~~ orthonormal bases

$\{u_j\}_{j=1}^{m} \quad \{v_j\}_{j=1}^{n}$

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{bmatrix} \xrightarrow{A} \begin{bmatrix} \sigma_{11}\tilde{x}_1 \\ \sigma_{22}\tilde{x}_2 \\ \vdots \\ \sigma_{rr}\tilde{x}_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_m \end{bmatrix}$$

So, for A linear $R \Rightarrow R$ we have the simple form:

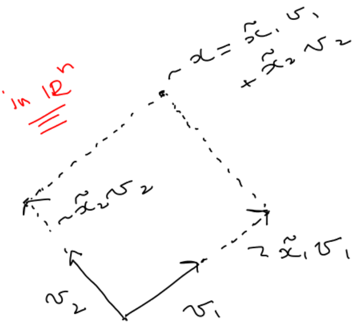
$$y = ax$$

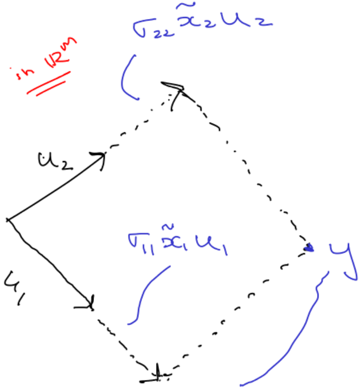
where $A = [a]$, 1×1.

In general, after you rotate to certain orthogonal bases, a rank r linear transformation $R^n \Rightarrow R^m$ is just the simple one r times.

$$\tilde{y}_i = \sigma_{ii}\,\tilde{x}_i, \quad i = 1, 2, \ldots, r.$$

$r = 2.$





$A : \mathbb{R}^n \to \mathbb{R}^m$

$\dim(R(A)) = r$

$\dim(N(A)) = n - r$

$$y = \sigma_{11} \tilde{x}_1 u_1 + \sigma_{22} \tilde{x}_2 u_2$$
$$= \tilde{y}_1 u_1 + \tilde{y}_2 u_2$$

Note:

U orthogonal.

$1 = |I| = |U'U| = |U'||U| = |U|^2 \Rightarrow |U| = \pm 1.$

Note:

$A, m \times m$ square of full rank so $r = m$.

Obviously, $\tilde{x}_i \to \sigma_{ii}\tilde{x}_i$ changes the volume by $\Pi_{i=1}^m \sigma_{ii}$.

$|A| = |U||\Sigma||V'| = (\pm 1)|\Sigma| = (\pm 1)\Pi_{i=1}^m \sigma_{ii}$.

Note:

Inverse of $\tilde{x}_i \to \tilde{y}_i = \sigma_{ii}\tilde{x}_i$ is

$$\tilde{y}_i \to \tilde{x}_i = \frac{1}{\sigma_{ii}}\tilde{y}_i$$

which is exactly $V\Sigma^{-1}U'$.

4. Reduced Form

You can simplify the construction to the “reduced form” by getting rid of the some zeros in Σ and corresponding columns in U and/or V.

Consider the case where $m > n$ and the rank is n so that the columns of A, $m \times n$ are linearly independent.

Full SVD

$$\underbrace{\begin{bmatrix} A \end{bmatrix}}_{m\times n} = \begin{bmatrix} \underbrace{U_1}_{m\times n} & \vdots & \underbrace{U_2}_{m\times(m-n)} \end{bmatrix} \underbrace{\begin{bmatrix} \overbrace{\tilde{\Sigma}}^{n\times n} \\ \underbrace{0}_{(m-n)\times n} \end{bmatrix}}_{m\times n} \underbrace{\begin{bmatrix} V^T \end{bmatrix}}_{n\times n}$$

$$= \underbrace{\begin{bmatrix} U_1 \end{bmatrix}}_{m\times n} \underbrace{\begin{bmatrix} \tilde{\Sigma} \end{bmatrix}}_{n\times n} \underbrace{\begin{bmatrix} V^T \end{bmatrix}}_{n\times n}$$

Reduced SVD

$$\boxed{\begin{aligned} AV &= U_1 \tilde{\Sigma} \\ Av_j &= \sigma_{jj} u_j \\ j &= 1, 2, \dots n \end{aligned}}$$

U_2 is just an orthonormal basic for $R(A)^{\perp}$, you don't need it.

In general we have:

$$\underset{m\times n}{A} = [\underbrace{U_1}_{m\times r}, \underbrace{U_2}_{m\times(m-r)}] \begin{bmatrix} \tilde{\Sigma} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \begin{array}{l} \} r\times n \\ \\ \end{array}$$

$$= U_1 \tilde{\Sigma} V_1^T$$

Columns of U_1 are an orthonormal basis for the column space of A.

Columns of V_1 are an orthonormal basis for the row space of A.

5. SVD and Least Squares

Let's see how the SVD decomposition can be used to compute the least squares solution.

Let's assume that X, $n \times p$ is of full rank p, where of course,

$$y = X\beta + \epsilon$$

is our model.

We simplify the SVD by using the reduced form.

$$X = \underset{n \times p}{U_1} \underset{p \times p}{\tilde{\Sigma}} \underset{p \times p}{V^T}$$

$$\tilde{\Sigma} = \mathrm{diag}(\sigma_{ii}) \quad i = 1, 2, \ldots p$$

$$V \text{ orthogonal}, \quad U_1^T U_1 = I_p$$

$$(X^T X) = V \tilde{\Sigma} U_1^T U_1 \tilde{\Sigma} V^T = V \tilde{\Sigma}^2 V^T$$

$$(X^T X)^{-1} = V \tilde{\Sigma}^{-2} V^T$$

$$X^T y = V \tilde{\Sigma} U_1^T y$$

$$(X^T X)^{-1} X^T y = \left[V \tilde{\Sigma}^{-2} V^T \right] \left\{ V \tilde{\Sigma} U_1^T y \right\}$$

$$= V \tilde{\Sigma}^{-1} U_1^T y = \sum_{i=1}^{p} v_i \frac{\langle u_i, y \rangle}{\sigma_{ii}}$$

This just says:

Want to solve $y \approx Xb$

First replace y with $\hat{y} = \sum_{i=1}^{p} \langle y, u_i \rangle u_i$

$$= \sum_{i=1}^{p} \tilde{y}_i u_i$$

$$b = \sum_{i=1}^{p} \tilde{x}_i v_i$$

So we have to solve: $\tilde{y}_i = \tilde{x}_i \sigma_{ii}$
for $\tilde{x}_i$ given $\tilde{y}_i = \langle y, u_i \rangle$

$$\Rightarrow \quad \tilde{x}_i = \frac{\tilde{y}_i}{\sigma_{ii}} \Rightarrow x = \sum_{i=1}^{p} \frac{\langle y, u_i \rangle}{\sigma_{ii}} v_i$$

6. SVD and Spectral

A, $m \times n$. $A = U\Sigma V'$.

$A'A = [V\Sigma' U'][U\Sigma V'] = V\Sigma'\Sigma V'$

$$\Sigma^T \Sigma = \underset{n\times m}{\begin{bmatrix} \tilde{\Sigma} & 0 \\ 0 & 0 \end{bmatrix}} \underset{m\times n}{\begin{bmatrix} \tilde{\Sigma} & 0 \\ 0 & 0 \end{bmatrix}}$$

$$= \underset{n\times n}{\begin{bmatrix} \tilde{\Sigma}^2 & 0 \\ 0 & 0 \end{bmatrix}} \equiv \Sigma_n^2$$

So, $A'A = V\Sigma_n^2 V'$.

Similarly, $AA' = U\Sigma_m^2 U'$.

:

In our our svd, the rank of A, $m \times n$ is the number of non-zero diagonals of Σ which is r in our basic notation.

For a symmetric matrix S, the rank is the number of non-zero eigen values which is the number of non-zero elements if the diagonal matrix D in $S = PDP'$.

So, from the previous slide we have that the rank of A is the same as the rank of $A'A$ and AA'.

Of course, the rank of $X'X$ is relevant.

7. Condition Number of a Matrix

If the columns (or rows) of a matrix A are linearly dependent, then it can cause a problem, depending on what you want to do.

In linear regression, if X is the design matrix, the if the columns are linearly dependent you cannot invert $X'X$.

More generally, if the colums are *close* to being linearly dependent then computation will become numerically unstable. That is, if some of the σ_{jj} are close to 0 for $j \in 1, 2, \ldots, r$ this can cause trouble.

We saw that computing the coefficients for the projection on the the column space involved $1/\sigma_{jj}$ so you can see if these are very small, we have trouble.

Suppose $X, n \times p$ is of full column rank so that $p = r$.

Then,

$$X = U_1 \tilde{\Sigma} V'$$

as we discussed above when we looked at the reduced form.

Here, U_1 is $n \times p$, $\tilde{\Sigma}$ is $p \times p$, and V is $p \times p$.

The diagonals of $\tilde{\Sigma}$ are $\sigma_{jj}, j = 1, 2, \ldots, p$, $\sigma_j > \sigma_{j+1,j+1} > 0$.

The degree to which ill-conditioning prevents a matrix from being inverted accurately depends on the ratio of its largest to smallest singular value, a quantity known as the condition number: which is

$$\text{Condition number } = \frac{\sigma_{11}}{\sigma_{pp}}$$

8. Moore Penrose Generalized Inverse

In solving the least squares problem, we have generally assumed that the design matrix X, $n \times p$ is of full rank p.

If X is not of full rank then there a many solutions to

$$\min_b ||y - Xb||^2$$

The Moore Penrose inverse chooses a solution for us.

Suppose we want to solve

$$y = Xb \quad \text{for } b$$
given y and X.

If X is $n \times n$ full rank

$$b^* = X^{-1} y \quad \text{is an exact solution.}$$

If X is $n \times p$ full rank (p) then

$$\hat{b} = (X^T X)^{-1} X^T y$$

is the closest we can get in that
$X\hat{b} \approx y$.

If X is not of full rank then

$\exists\ \alpha_i,\ i=1,2,\ldots p$, not all 0

such that $\sum_{i=1}^{p} x_i \alpha_i = 0$.

or $X\alpha = 0$

thus $X(b + c\alpha) = Xb + cX\alpha$

$= Xb$

so if b^* is a solution to

$\min_b \|y - Xb\|^2$ so is $b^* + c\alpha$

— there is not a unique solution.

Let $X = U_1 \tilde{\Sigma} V_1^T$

– the reduced form SVD of X

Let $X^+ \equiv V_1 \tilde{\Sigma}^{-1} U_1^T$

the Moore-Penrose generalized inverse of X.

Claim $\hat{b} = X^+ y$ is a solution to

$$\min_b \|y - Xb\|^2$$

Note: U_1 is $n \times r$, $\tilde{\Sigma}$ is $r \times r$, V_1 is $p \times r$.

The columns of U_1 are an orthonormal basis for the column space of X we like to project y onto.

The columns of V_1 are an orthonormal basis for the space of coefficient vectors that do NOT map to 0.

Have to check

$$X^T(y - Xb^o) = 0 \quad \text{or} \quad X^T y = X^T X b^o$$

$$X = U_1 \tilde{\Sigma} V_1^T \qquad X^T = V_1 \tilde{\Sigma} U_1^T$$

$$X^+ = V_1 \tilde{\Sigma}^{-1} U_1^T \qquad b^o = X^+ y$$

$$X^T X b^o = X^T X X^+ y$$

$$X^T X X^+ = [V_1 \tilde{\Sigma} U_1^T][U_1 \tilde{\Sigma} V_1^T]\{V_1 \tilde{\Sigma}^{-1} U_1^T\}$$

$$= V_1 \tilde{\Sigma} U_1^T = X^T$$

so $\quad X^T X b^o = X^T X X^+ y = X^T y$

Clearly, $XX^+ y = Xb^o$ projects y onto the column space of X.

$$X^+ = V_1 \tilde{\Sigma}^{-1} U_1^T$$

$$X^+ y = V_1 \tilde{\Sigma}^{-1} \begin{bmatrix} \langle u_1, y \rangle \\ \langle u_2, y \rangle \\ \vdots \\ \langle u_r, y \rangle \end{bmatrix}$$

$$= V_1 \begin{bmatrix} \langle u_1, y \rangle / \sigma_{11} \\ \langle u_2, y \rangle / \sigma_{22} \\ \vdots \\ \langle u_r, y \rangle / \sigma_{rr} \end{bmatrix}$$

$$= \sum_{i=1}^{r} v_i \frac{\langle u_i, y \rangle}{\sigma_{ii}}$$

XX^+ projects onto the column space of X.

$$X = U_1 \, \tilde{\Sigma} \, V_1', \quad X^+ = V_1 \, \tilde{\Sigma}^{-1} \, U_1'$$

$$X \, X^+ = [U_1 \, \tilde{\Sigma} \, V_1'][V_1 \, \tilde{\Sigma}^{-1} \, U_1] = U_1 \, U_1'$$

$X^+ \, X$ projects onto the row space of X.

$$X^+ \, X = [V_1 \, \tilde{\Sigma}^{-1} \, U_1'][U_1 \, \tilde{\Sigma} \, V_1'] = V_1 \, V_1'$$

$X^+ X$ projects y onto the row space of X. gives us a characterization of the MP choice of solution.

$$V = [\underbrace{V_1}_{\text{ON basis for row space of } X}, \underbrace{V_2}_{\text{ON basis for subspace orthogonal to the row space}}]$$

Any b in $\mathbb{R}^p$ can be written

$$b = V_1 \alpha + V_2 \gamma$$

$$X^+ X b = [V_1 \tilde{\Sigma}^{-1} U_1^T][U_1 \tilde{\Sigma} V_1^T][V_1 \alpha + V_2 \gamma]$$

$$= V_1 \alpha$$

$X^+ X$ is a projection onto the row space of X.

The column space and row space of X have the same dimension so we can define a 1-1 map between them.

Everthing else gets projected away.

9. Matrix Approximation

Suppose $\sigma_{11} \geq \sigma_{22} \geq \ldots \sigma_{rr}$ and after s they are small, $\sigma_{ii} \approx 0$, $i > s$.

$$\underset{m \times n}{A} = U_1 \tilde{\Sigma} V_1^T$$

$$= [u_1, u_2, \ldots u_r] \begin{bmatrix} \sigma_{11} & & & \\ & \sigma_{22} & & \\ & & \ddots & \\ & & & \sigma_{rr} \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_r^T \end{bmatrix}$$

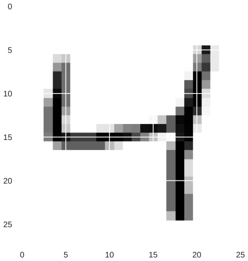
$$= [u_1, u_2, \ldots u_r] \begin{bmatrix} \sigma_{11} v_1^T \\ \sigma_{22} v_2^T \\ \vdots \\ \sigma_{rr} v_r^T \end{bmatrix} = \sum_{i=1}^{r} \sigma_{ii} \underbrace{u_i}_{m \times 1} \underbrace{v_i^T}_{1 \times n}$$

$$\approx \sum_{i=1}^{s} \sigma_{ii} u_i v_i^T = [u_1, u_2, \ldots u_s] \begin{bmatrix} \sigma_{11} v_1^T \\ \sigma_{22} v_2^T \\ \vdots \\ \sigma_{ss} v_s^T \end{bmatrix}$$

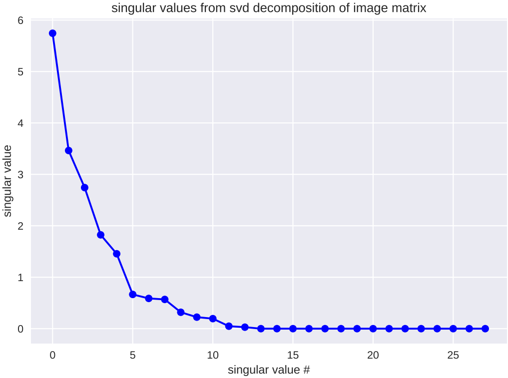
$$= \hat{U} \hat{V}^T$$

Example:

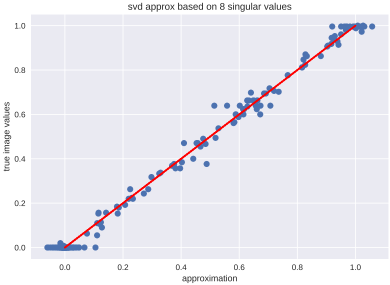
Here is a 28x28 grey scale image of a digit.
The (i,j) element of the matrix is 0:255 indicating the grayscale.
I divided by 255.



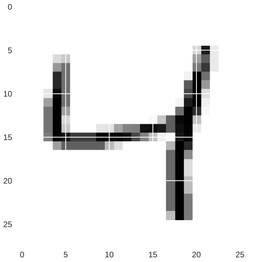
Here are the singular values from the 28x28 matrix.
This is called a scree plot.



Here is are the matrix values for the image plotted against the values from the approximation using 8 singular values.



Here is a 28x28 grey scale image of a digit.



and the approx: