

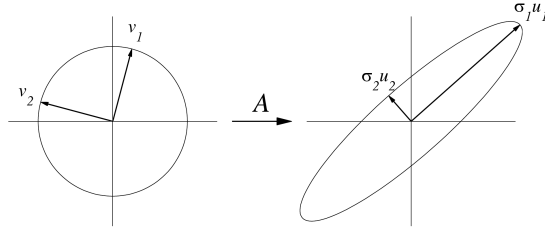
# Data mining: lecture 8

Edo liberty, adapted from class notes by Yoel Shkolnisky

We will see that any matrix  $A \in \mathbb{R}^{m \times n}$  can be written as  $A = U\Sigma V^T$  such that  $U \in \mathbb{R}^{m \times m}$  is unitary,  $V \in \mathbb{R}^{n \times n}$  is unitary, and  $\Sigma \in \mathbb{R}^{m \times n}$  is a non-negative real diagonal matrix.  $\Sigma(i, i)$ , denoted  $\sigma_i$ , are unique. If  $A$  the singular values are distinct, then the singular vectors are unique up to a multiplication by  $z \in \mathbb{C}$  with  $|z| = 1$ .

**Remark 0.1.** Note the difference in notation from what we saw in class. The matrices  $V$  and  $U$  are what we denoted by  $[V; \bar{V}]$  and  $[U; \bar{U}]$  respectively. This makes the proofs a little cleaner and hopefully more easy to follow. Note also that  $\Sigma$ , unlike the matrix we denoted by  $S$ , is not square. The non square matrix  $\Sigma$  is still diagonal though, i.e.  $\Sigma(i, j) = 0$  for all  $i \neq j$ .

## 1 The geometry of SVD



## 2 Proof of existence

Set  $\sigma_1 = \|A\|_2$ . Let  $u_1 \in \mathbb{R}^n$  and  $v_1 \in \mathbb{R}^m$  be unit 2-norm vectors such that  $Av_1 = \sigma_1 u_1$ . To find these vectors, find the unit vector  $v_1$  that brings to maximum the expression

$$\max_{\|x\|=1} \|Ax\|.$$

Then  $Av_1 = \mu u_1$  for some  $\mu$  and a unit vector  $u_1$ . Since  $\|Av_1\| = \sigma_1$ , we get that  $\sigma_1 = \|Av_1\| = |\mu| \|u_1\| = |\mu|$ . Set  $\mu = \sigma_1$  to be positive, by flipping the sign of  $u_1$  if needed.

Complete  $v_1$  into an orthonormal basis of  $\mathbb{C}^n$ , denote  $V_1$ . Complete  $u_1$  into

an orthonormal basis of  $\mathbb{C}^m$ , denoted  $U_1$ .

$$S = U_1^T A V_1 = U_1^T [\sigma_1 u_1, A v_2, \dots, A v_n] = \begin{pmatrix} \sigma_1 & w^T \\ 0 & B \end{pmatrix}.$$

We will show that  $w^T = 0$ .

$$\left\| S \begin{pmatrix} \sigma_1 \\ w \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \sigma_1 & w^T \\ 0 & B \end{pmatrix} \begin{pmatrix} \sigma_1 \\ w \end{pmatrix} \right\|_2 \geq \sigma_1^2 + w^T w = \sqrt{\sigma_1^2 + w^T w} \left\| \begin{pmatrix} \sigma_1 \\ w \end{pmatrix} \right\|_2.$$

That is  $\|S\| \geq \sqrt{\sigma_1^2 + w^T w}$ . But  $\|S\|_2 = \|A\|_2 = \sigma_1$  and so  $w = 0$ .

By induction,  $B = U_2 \Sigma_2 V_2^T$  and

$$A = U_1 S V_1^T = U_1 \begin{pmatrix} 1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & V_2^T \end{pmatrix} V_1^T.$$

The matrices

$$U = U_1 \begin{pmatrix} 1 & 0 \\ 0 & U_2 \end{pmatrix}, \quad V = V_1 \begin{pmatrix} 1 & 0 \\ 0 & V_2^T \end{pmatrix}$$

are unitary and the proof is complete.

### 3 More properties of SVD

**Lemma 3.1.** *The rank of  $A$  equals the number of nonzero singular values.*

*Proof.* Recall that if  $B$  is  $n \times k$  with rank  $n$  then  $\text{rank}(AB) = \text{rank}(A)$ , and if  $C$  is  $l \times m$  with rank  $m$  then  $\text{rank}(CA) = \text{rank}(A)$ . Thus,

$$\text{rank}(A) = \text{rank}(U \Sigma V^T) = \text{rank}(\Sigma V^T) = \text{rank}(\Sigma).$$

Since  $\Sigma$  is diagonal, its rank is the number of its nonzero elements. □

**Lemma 3.2.** *Let  $\text{rank}(A) = r$ . Then,*

$$\begin{aligned} \text{range}(A) &= \text{span}(u_1, \dots, u_r), \\ \text{null}(A) &= \text{span}(v_{r+1}, \dots, v_n). \end{aligned}$$

*Proof.*

$$\begin{aligned} y \in \text{range}(A) &\iff \exists x \text{ such that } y = Ax \\ &\iff y = U \Sigma V^T x \\ &\iff y = U \Sigma z, \text{ where } z = V^T x \\ &\iff y = U (\sigma_1 z_1, \dots, \sigma_r z_r, 0, \dots, 0)^T \\ &\iff y = \sum_{i=1}^r (\sigma_i z_i) u_i \\ &\iff y \in \text{span}(u_1, \dots, u_r). \end{aligned}$$

$$\begin{aligned}
x \in \text{null}(A) &\iff \|Ax\|_2 = 0 \iff \|U\Sigma V^T x\|_2 = 0 \\
&\iff \|\Sigma V^T x\|_2 = 0 \iff \|\Sigma y\|_2 = 0 \text{ where } y = V^T x \\
&\iff y = (0, \dots, 0, y_{r+1}, \dots, y_n)^T \text{ where } y = V^T x \\
&\iff x = Vy, \ y = (0, \dots, 0, y_{r+1}, \dots, y_n)^T \\
&\iff x = \sum_{i=r+1}^n y_i v_i \\
&\iff x \in \text{span}(v_{r+1}, \dots, v_n).
\end{aligned}$$

□

**Lemma 3.3.**  $\|A\|_2 = \sigma_1$  (even if you don't know the above proof).

*Proof.* Immediate from the invariance of  $\|\cdot\|_2$  under unitary transformations.

□

Similarly,  $\|A\|_F = (\sigma_1^2 + \dots + \sigma_r^2)^{1/2}$ .

## 4 Relation between singular values and eigenvalues

**Lemma 4.1.** *The singular values of  $A$  are the square roots of the nonzero eigenvalues of  $A^T A$  and  $AA^T$ .*

*Proof.* If  $A = U\Sigma V^T$ , then  $A^T = V\Sigma U^T$  and

$$AA^T = (U\Sigma V^T)(V\Sigma U^T) = U\Sigma\Sigma U^T = U\Sigma^2 U^{-1}.$$

$AA^T$  is positive semi-definite and therefore all eigenvalues are non-negative and there is no problem with the square root. □

Do not use this observation to compute the SVD! Reason: Assume for simplicity that we have a  $2 \times 2$  matrix  $A$  (not diagonal) whose SVD is given by  $A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$  (See 5.1 below). If  $\sigma_2/\sigma_1 < 10^{-15}$ , then the second term will disappear due to round-off, that is, we cannot represent such a matrix in double precision. Now, if the matrix  $A$  has  $\sigma_2/\sigma_1 < 10^{-8}$ , then  $A^T A$  and  $AA^T$  have ratio of singular values that is smaller than  $10^{-15}$ , and so those matrices cannot be represented, and will be approximated as rank-1 matrices with the second singular value being due to round-off. In other words, although  $A$  is not terribly conditioned, we lose the small eigenvalues if we try to compute the SVD by computing the eigenvalues of  $A^T A$  or  $AA^T$ .

**Lemma 4.2.** *If  $A$  is hermitian, then the singular values of  $A$  are the absolute values of its eigenvalues.*

*Proof.* A hermitian matrix is diagonalized by a unitary matrix with real eigenvalues. That is,

$$A = Q\Lambda Q^T = Q|\Lambda| \text{sign}(\Lambda)Q^T.$$

Now set  $U = Q$ ,  $\Sigma = |\Lambda|$ ,  $V^T = \text{sign}(\Lambda)Q^T$ . □

## 5 Approximation properties

### 5.1 Rank-k approximation in the spectral norm

**Lemma 5.1.** *A can be written as a sum of rank-1 matrices. Explicitly,*

$$A = \sum_{j=1}^r \sigma_j u_j v_j^T.$$

**Theorem 5.1.** *Set*

$$A_k = \sum_{j=1}^k \sigma_j u_j v_j^T.$$

*Then,*

$$\min_{\substack{B \in \mathbb{C}^{m \times n} \\ \text{rank}(B) \leq k}} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}.$$

*Proof.*

$$A - A_k = \sum_{j=1}^r \sigma_j u_j v_j^T - \sum_{j=1}^k \sigma_j u_j v_j^T = \sum_{j=k+1}^r \sigma_j u_j v_j^T$$

and thus  $\sigma_{k+1}$  is the largest singular value of  $A - A_k$ . Alternatively, look at  $U^T A_k V = \text{diag}(\sigma_1, \dots, \sigma_k, 0, \dots, 0)$ , which means that  $\text{rank}(A_k) = k$ , and that

$$\|A - A_k\|_2 = \|U^T(A - A_k)V\|_2 = \|\text{diag}(0, \dots, 0, \sigma_{k+1}, \dots, \sigma_r)\|_2 = \sigma_{k+1}.$$

Let  $B$  be an arbitrary matrix with  $\text{rank}(B_k) = k$ . Then, it has a null space of dimension  $n - k$ , that is,

$$\text{null}(B) = \text{span}(w_1, \dots, w_{n-k}).$$

A dimension argument shows that

$$\text{span}(w_1, \dots, w_{n-k}) \cap \text{span}(v_1, \dots, v_{k+1}) \neq \{0\}.$$

Let  $w$  be a unit vector from the intersection. Since

$$Aw = \sum_{j=1}^{k+1} \sigma_j (v_j^T w) u_j,$$

we have

$$\|A - B\|_2^2 \geq \|(A - B)w\|_2^2 = \|Aw\|_2^2 = \sum_{j=1}^{k+1} \sigma_j^2 |v_j^T w|^2 \geq \sigma_{k+1}^2 \sum_{j=1}^{k+1} |v_j^T w|^2 = \sigma_{k+1}^2,$$

since  $w \in \text{span}\{v_1, \dots, v_{k+1}\}$ , and the  $v_j$  are orthogonal. □

## 5.2 Rank-k approximation in the Frobenius norm

The same theorem holds with the Frobenius norm.

**Theorem 5.2.** *Set*

$$A_k = \sum_{j=1}^k \sigma_j u_j v_j^T.$$

*Then,*

$$\min_{\substack{B \in \mathbb{C}^{m \times n} \\ \text{rank}(B) \leq k}} \|A - B\|_F = \|A - A_k\|_F = \sqrt{\sum_{i=k+1}^n \sigma_i^2}.$$

*Proof.* Suppose  $A = U\Sigma V^T$ . Then

$$\min_{\text{rank}(B) \leq k} \|A - B\|_F^2 = \min_{\text{rank}(B) \leq k} \|U\Sigma V^T - UU^T B V V^T\|_F^2 = \min_{\text{rank}(B) \leq k} \|\Sigma - U^T B V\|_F^2.$$

Now,

$$\|\Sigma - U^T B V\|_F^2 = \sum_{i=1}^n (\sigma_{ii} - (U^T B V)_{ii})^2 + \text{off-diagonal terms}.$$

If  $B$  is the best approximation matrix and  $U^T B V$  is not diagonal, then write  $U^T B V = D + O$ , where  $D$  is diagonal and  $O$  contains the off-diagonal elements. Then the matrix  $B = U D V^T$  is a better approximation, which is a contradiction.

Thus,  $U^T B V$  must be diagonal. Hence,

$$\|\Sigma - D\|_F^2 = \sum_{i=1}^n (\sigma_i - d_i)^2 = \sum_{i=1}^k (\sigma_i - d_i)^2 + \sum_{i=k+1}^n \sigma_i^2,$$

and this is minimal when  $d_i = \sigma_i$ ,  $i = 1, \dots, k$ . The best approximating matrix is  $A_k = U D V^T$ , and the approximation error is  $\sqrt{\sum_{i=k+1}^n \sigma_i^2}$ .  $\square$

## 5.3 Closest orthogonal matrix

The SVD also allows to find the orthogonal matrix that is closest to a given matrix. Again, suppose that  $A = U\Sigma V^T$  and  $W$  is an orthogonal matrix that minimizes  $\|A - W\|_F^2$  among all orthogonal matrices. Now,

$$\|U\Sigma V^T - W\|_F^2 = \|U\Sigma V^T - UU^T W V V^T\| = \|\Sigma - \tilde{W}\|,$$

where  $\tilde{W} = U^T W V$  is another orthogonal matrix. We need to find the orthogonal matrix  $\tilde{W}$  that is closest to  $\Sigma$ . Alternatively, we need to minimize  $\|\tilde{W}^T \Sigma - I\|_F^2$ .

If  $U$  is orthogonal and  $D$  is diagonal and positive, then

$$\begin{aligned} \text{trace}(UD) &= \sum_{i,k} u_{ik} d_{ki} \leq \sum_i \left( \left( \sum_k u_{ik}^2 \right)^{1/2} \left( \sum_k d_{ik}^2 \right)^{1/2} \right) \\ &= \sum_i \left( \sum_k d_{ki}^2 \right)^{1/2} = \sum_i (d_{ii}^2)^{1/2} = \sum_i d_{ii} = \text{trace}(D). \end{aligned} \quad (1)$$

Now

$$\begin{aligned} \left\| \tilde{W}^T \Sigma - I \right\|_F^2 &= \text{trace} \left( (\tilde{W}^T \Sigma - I) (\tilde{W}^T \Sigma - I)^T \right) \\ &= \text{trace} \left( (\tilde{W}^T \Sigma - I) (\Sigma \tilde{W} - I) \right) \\ &= \text{trace} (\tilde{W}^T \Sigma^2 \tilde{W}) - \text{trace} (\tilde{W}^T \Sigma) - \text{trace} (\Sigma \tilde{W}) + n \\ &= \text{trace} \left( (\Sigma \tilde{W})^T (\Sigma \tilde{W}) \right) - 2 \text{trace} (\Sigma \tilde{W}) + n \\ &= \left\| \Sigma \tilde{W} \right\|_F^2 - 2 \text{trace} (\Sigma \tilde{W}) + n \\ &= \left\| \Sigma \right\|_F^2 - 2 \text{trace} (\Sigma \tilde{W}) + n. \end{aligned}$$

Thus, we need to maximize  $\text{trace}(\Sigma \tilde{W})$ . But this is maximized by  $\tilde{W} = I$  by (1). Thus, the best approximating matrix is  $W = UV^T$ .

## 6 The “Thin” SVD

Also called “economy size” SVD. If  $A \in \mathbb{C}^{m \times n}$ ,  $A = U\Sigma V^T$ , and  $m \geq n$ , then the “thin” SVD is  $A = U_1 \Sigma_1 V^T$  where

$$U_1 = [u_1, \dots, u_n] \in \mathbb{C}^{m \times n}$$

and

$$\Sigma_1 = \text{diag}(\sigma_1, \dots, \sigma_n) \in \mathbb{R}^{n \times n}.$$

## 7 Applications of the SVD

1. Determining range, null space and rank (also numerical rank).
2. Matrix approximation.
3. Inverse and Pseudo-inverse: If  $A = U\Sigma V^T$  and  $\Sigma$  is full rank, then  $A^{-1} = V\Sigma^{-1}U^T$ . If  $\Sigma$  is singular, then its pseudo-inverse is given by  $A^\dagger = V\Sigma^\dagger U^T$ , where  $\Sigma^\dagger$  is formed by replacing every nonzero entry by its reciprocal.

4. Least squares: If we need to solve  $Ax = b$  in the least-squares sense, then  $x_{LS} = V\Sigma^\dagger U^T b$ .
5. Denoising – Small singular values typically correspond to noise. Take the matrix whose columns are the signals, compute SVD, zero small singular values, and reconstruct.
6. Compression – We have signals as the columns of the matrix  $S$ , that is, the  $i$  signal is given by

$$S_i = \sum_{j=1}^r (\sigma_j v_{ij}) u_j.$$

If some of the  $\sigma_i$  are small, we can discard them with small error, thus obtaining a compressed representation of each signal. We have to keep the coefficients  $\sigma_j v_{ij}$  for each signal and the dictionary, that is, the vectors  $u_i$  that correspond to the retained coefficients.

## 8 Differences between SVD and eigen-decomposition

1. Not every matrix has an eigen-decomposition (not even any square matrix). Any matrix (even rectangular) has an SVD.
2. In eigen-decomposition  $A = X\Lambda X^{-1}$ , that is, the eigen-basis is not always orthogonal. The basis of singular vectors is always orthogonal.
3. In SVD we have two singular-bases (right and left).
4. SVD tells everything on a matrix.
5. SVD as no numerical problems.
6. Relation to condition number; the numerical problems with eigen-decomposition; multiplication by an orthogonal matrix is perfectly conditioned.

## 9 Linear regression in the least-squared loss

In Linear regression we aim to find the best linear approximation to a set of observed data. For the  $m$  data points  $\{x_1, \dots, x_m\}$ ,  $x_i \in \mathbb{R}^n$ , each receiving the value  $y_i$ , we look for the weight vector  $w$  that minimizes:

$$\sum_{i=1}^m (x_i^T w - y_i)^2 = \|Aw - y\|_2^2$$

Where  $A$  is a matrix that holds the data points as rows  $A_i = x_i^T$ .

**Proposition 9.1.** *The vector  $w$  that minimizes  $\|Aw - y\|_2^2$  is  $w = A^\dagger y = V\Sigma^\dagger U^T y$  for  $A = U\Sigma V^T$  and  $\Sigma_{ii}^\dagger = 1/\Sigma_{ii}$  if  $\Sigma_{ii} > 0$  and 0 else.*

Let us define  $U_{\parallel}$  and  $U_{\perp}$  as the parts of  $U$  corresponding to positive and zero singular values of  $A$  respectively. Also let  $y_{\parallel} = 0$  and  $y_{\perp}$  be two vectors such that  $y = y_{\parallel} + y_{\perp}$  and  $U_{\parallel}y_{\perp} = 0$  and  $U_{\perp}y_{\parallel} = 0$ .

Since  $y_{\parallel}$  and  $y_{\perp}$  are orthogonal we have that  $\|Aw - y\|_2^2 = \|Aw - y_{\parallel} - y_{\perp}\|_2^2 = \|Aw - y_{\parallel}\|_2^2 + \|y_{\perp}\|_2^2$ . Now, since  $y_{\parallel}$  is in the range of  $A$  there is a solution  $w$  for which  $\|Aw - y_{\parallel}\|_2^2 = 0$ . Namely,  $w = A^{\dagger}y = V\Sigma^{\dagger}U^T y$  for  $A = U\Sigma V^T$ . This is because  $U\Sigma V^T V\Sigma^{\dagger}U^T y = y_{\parallel}$ . Moreover, we get that the minimal cost is exactly  $\|y_{\perp}\|_2^2$  which is independent of  $w$ .

## 10 Optimal squared loss dimension reduction

Given a set of  $n$  vectors  $x_1, \dots, x_n$  in  $\mathbb{R}^m$ . We look for a rank  $k$  projection matrix  $P \in \mathbb{R}^{m \times m}$  that minimizes:

$$\sum_{i=1}^n \|Px_i - x_i\|_2^2$$

If we denote by  $A$  the matrix whose  $i$ 'th column is  $x_i$  then this is equivalent to minimizing  $\|PA - A\|_{Fro}^2$ . Since the best possible rank  $k$  approximation to the matrix  $A$  is  $A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$  the best possible solution would be a projection  $P$  for which  $PA = A_k$ . This is achieved by  $P = U_k U_k^T$  where  $U_k$  is the matrix corresponding to the first  $k$  left singular vectors of  $A$ .

If we define  $y_i = U_k^T x_i$  we see that the values of  $y_i \in \mathbb{R}^k$  are optimally fitted to the set of points  $x_i$  in the sense that they minimize:

$$\min_{y_1, \dots, y_n} \min_{\Psi \in \mathbb{R}^{k \times m}} \sum_{i=1}^n \|\Psi y_i - x_i\|_2^2$$

The mapping of  $x_i \rightarrow U_k^T x_i = y_i$  thus reduces the dimension of any set of points  $x_1, \dots, x_n$  in  $\mathbb{R}^m$  to a set of points  $y_1, \dots, y_n$  in  $\mathbb{R}^k$  optimally in the squared loss sense. This is commonly referred to as Principal Component Analysis (PCA).