1 Introduction

In this lecture, we see a very popular und useful dimension reduction technique that is based on the *singular value decomposition* (SVD) of a matrix. In contrast to the dimension reduction obtained by the Johnson-Lindenstrauss Lemma, SVD based dimension reductions are not distance preserving. That means that we allow that the distances between pairs of points in our input change. Instead, we want to keep the shape of the point set by fitting it to a subspace according to a least squares error. This preserves most of the 'energy' of the points.

More precisely, the problem that we want to solve is the following. We are given a matrix $A \in \mathbb{R}^{n \times d}$. The points are the rows of A, which we also name $a_1, \ldots, a_n \in \mathbb{R}^d$. Let the rank of A be r, so $r \leq \min\{n, d\}$. Given an integer k, we want to find a subspace V of dimension k that minimizes the sum of the squared distances of all points in A to V. Thus, for each point in A, we square the distance between the point and its projection to V and add these squared errors, and this term should be minimized by our choice of V.

This task can be solved by computing the SVD of A, a decomposition of A into matrices with nice properties. We will see that we can write A as

$$A = \begin{pmatrix} — & a_1 & — \\ & \vdots & \\ — & a_n & — \end{pmatrix} = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix} \begin{pmatrix} — & v_1 & — \\ & \vdots & \\ — & v_d & — \end{pmatrix} = UDV^\intercal$$

where $U \in \mathbb{R}^{n \times r}$ and $V \in \mathbb{R}^{r \times d}$ are matrices with orthonormal columns and $D \in \mathbb{R}^{r \times r}$ is a diagonal matrix. Notice that the columns of V are the d-dimensional points $v_1, \ldots, v_d$ which appear in the rows of the above matrix since it is $V^\intercal$.

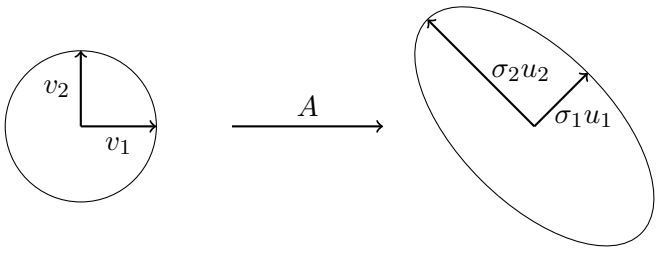

Figure 21.1: A visualization of $AV = UD$ for $r = 2$.

Notice that the SVD can give us an intuition of how A acts as a mapping. We have that

$$AV = UDV^\intercal V = UD$$

because V consists of orthonormal columns. Imagine the r-dimensional sphere that is spanned by $v_1, \ldots, v_r$. The linear mapping defined by A maps this sphere to an ellipsoid with $\sigma_1 u_1, \ldots, \sigma_r u_r$ as the axes, like shown in Figure 21.1.

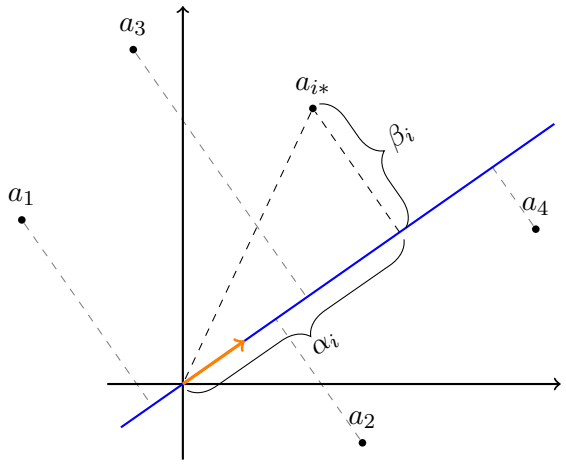

Figure 21.2: Finding the best fit subspace of dimension one.

The singular value decomposition was developed by different mathematicians around the beginning of the 19th century. The survey by Stewart [Ste93] gives an historical overview on its origins. In the following, we see how to obtain the SVD and why it solves our best fit problem. The lecture is partly based on [HK14].

2 Best fit subspaces of dimension k and the SVD

We start with the case that $k = 1$. Thus, we look for the line through the origin that minimizes the sum of the squared errors. See Figure 21.2. It depicts a one-dimensional subspace V in blue. We look at a point a_i, its distance β_i to V, and the length of its projection to V which is named α_i in the picture. Notice that the length of a_i is $\alpha_i^2 + \beta_i^2$. Thus, for our fixed a_i, minimizing β_i is equivalent to maximizing α_i. If we represent V by a unit vector v that spans V (depicted in orange in the picture), then we can compute the projection of a_i to V by the dot product $\langle a_i, v\rangle$. We have just argued that we can find the best fit subspace of dimension one by solving

$$\max_{v\in\mathbb{R}^d,\|v\|=1} \sum_{i=1}^{n} \langle a_i, v\rangle^2 = \min_{v\in\mathbb{R}^d,\|v\|=1} \sum_{i=1}^{n} \operatorname{dist}(a_i, \operatorname{span}(v))^2$$

where we denote the distance between a point a_i and the line spanned by v by $\operatorname{dist}(a_i, \operatorname{span}(v))^2$. Now because $Av = (\langle a_1, v\rangle, \langle a_2, v\rangle, \ldots, \langle a_n v\rangle)^\intercal$, we can rewrite $\sum_{i=1}^{d} \langle a_i, v\rangle^2$ as $\|Av\|^2$. We define the first right singular vector to be a unit vector that maximizes $\|Av\|$.[1] We thus know that the subspaces spanned by it is the best fit subspace of dimension one.

Now we want to generalize this concept to more than one dimension. It turns out that to do so, we can iteratively pick orthogonal unit vectors that span more and more dimensions. Among all unit vectors that are orthogonal to those chosen so far, we pick a vector that maximizes $\|Av\|$. This is formalized in the following definition.

[1] There may be many vectors that achieve the maximum: indeed, for every v that achieves the maximum, $-v$ also has the same maximum. Let us break ties arbitrarily.

Definition 21.1. Let $A \in \mathbb{R}^{n \times d}$ be a matrix. We define

$$\begin{aligned} v_1 &= \arg\max_{\|v\|=1} \|Av\|, & \sigma_1(A) &:= \|Av_1\| \\ v_2 &= \arg\max_{\|v\|=1, \langle v, v_1 \rangle = 0} \|Av\|, & \sigma_2(A) &:= \|Av_2\| \\ &\vdots \\ v_r &= \arg\max_{\|v\|=1, \langle v, v_i \rangle = 0 \ \forall i=1,\ldots,r-1} \|Av\|, & \sigma_r(A) &:= \|Av_r\| \end{aligned}$$

and say that $v_1, \ldots, v_r$ are *right singular vectors* of A and that $\sigma_1 := \sigma_1(A), \ldots, \sigma_r := \sigma_r(A)$ are the *singular values* of A. Then we define the *left singular vectors* by setting

$$u_i := \frac{Av_i}{\|Av_i\|} \quad \text{for all } i = 1, \ldots, r.$$

One worry is that this greedy process picked v_2 after fixing v_1, and hence the span of v_1, v_2 may not be the best two-dimensional subspace. The following claim says that Definition 21.1 indeed gives us the the best fit subspaces.

Claim 21.2. *For any k, the subspace V_k, which is the span of $v_1, \ldots, v_k$, minimizes the sum of the squared distances of all points among all subspaces of dimension k.*

Proof. Let V_2 be the subspace spanned by v_1 and v_2. Let W be any other 2-dimensional subspace and let w_1, w_2 be an orthonormal basis of W. Recall that the squared length of the projection of a point a_i to V decomposes into the squared lengths of the projections to the lines spanned by v_1 and v_2 and the same is true for W, w_1 and w_2.

Since we chose v_1 to maximize $\|Av\|$, we know that $\|Aw_1\| \leq \|Av_1\|$. Similarly, it holds that $\|Aw_2\| \leq \|Av_2\|$, which means that

$$\|Aw_1\|^2 + \|Aw_2\|^2 \leq \|Av_1\|^2 + \|Av_2\|^2.$$

We can extend this argument by induction to show that the space spanned by $v_1, \ldots, v_k$ is the best fit subspace of dimension k. □

We review some properties of the singular values and vectors. Notice that as long as $i < r$, there is always a vector in the row space of A that is linearly independent to $v_1, \ldots, v_i$, which ensures that $\max \|Av\|$ is nonzero. For $i = r$, the vectors $v_1, \ldots, v_r$ span the row space of A. Thus, any vector that is orthogonal to them lies in the kernel of A, meaning that $\arg\max_{\|v\|=1, \langle v, v_i \rangle = 0 \ \forall i=1,\ldots,i-1} \|Av\| = 0$, so we end the process at this point. By construction, we know that the singular values are not increasing. We also see that the right singular vectors form a orthonormal basis of the row space of A. This is true for the left singular vectors and the column space as well (homework). The following fact summarizes the important properties.

Fact 21.3. *The sets $\{u_1, \ldots, u_r\}$ and $\{v_1, \ldots, v_r\}$ as defined in 21.1 are both orthonormal sets and span the column and row space, respectively. The singular values satisfy $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.*

So far, we defined the v_i purely based on the goal to find the best fit subspace. Now we claim that in doing so, we have actually found the decomposition we wanted, i.e. that

$$UDV^\intercal := \begin{pmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix} \begin{pmatrix} — & v_1 & — \\ & \vdots & \\ — & v_d & — \end{pmatrix} = A. \tag{21.1}$$

Claim 21.4. *For any matrix $A \in \mathbb{R}^{n\times d}$ and U, V, D as in (21.1), it holds that*

$$A = UDV^\intercal.$$

Proof. We prove the claim by using the fact that two matrices $A, B \in \mathbb{R}^{n\times d}$ are identical iff for all vectors v, the images are equal, i.e. $Av = Bv$. Notice that it is sufficient to check this for a basis, so it is true if the following subclaim holds (which we do not prove):

Subclaim: Two matrices $A, B \in \mathbb{R}^{n\times d}$ are identical iff $Av = Bv$ for all v in a basis of $\mathbb{R}^d$.

We use the subclaim for $B = UDV^\intercal$. Notice that we can extend $v_1, \dots, v_r$ to a basis of $\mathbb{R}^d$ by adding orthonormal vectors from the kernel of A. These additional vectors are orthogonal to all vectors in the rows of $V^\intercal$, so $V^\intercal v$ is the zero vector for all of them. Since they are in the kernel of A, it holds $\vec{0} = Av = Bv = UD\vec{0} = \vec{0}$ for the additional basis vectors. For $i = 1, \dots, r$, we notice that

$$(UDV^\intercal)v_i = UDe_i = u_i\sigma_i = \frac{Av_i}{\|Av_i\|} \cdot \|Av_i\| = Av_i$$

which completes the proof. □

3 Useful facts about the SVD, including rank-k-approximation

Singular values are a generalization of the concept of eigenvalues for square matrices. Recall that a square symmetric matrix M can be written as $M = \sum_{i=1}^{r} \lambda_i v_i v_i^\intercal$ where λ_i and v_i are eigenvalues and eigenvectors, respectively. This decomposition can be used to define the singular vectors in a different way. In fact, the right singular vectors of A correspond to the eigenvectors of $A^\intercal A$ (notice that this matrix is square and symmetric), and the left singular vectors correspond to the eigenvectors of $AA^\intercal$.

This fact can also be used to compute the SVD. Computing the SVD or eigenvalues and -vectors in a numerically stable way is the topic of a large research area, and there are different ways to obtain algorithms that converge under the assumption of a finite precision.

Fact 21.5. *The SVD can be found (up to arbitrary precision) in time $\mathcal{O}(\min(nd^2, n^2d))$ or even in time $\mathcal{O}(\min(nd^{\omega-1}, dn^{\omega-1}))$ where ω is the matrix multiplication constant. (Here the big-O term hides the dependence on the precision.)*

The SVD is unique in the sense that for any $i \in [r]$, the subspace spanned by unit vectors v that maximize $\|Av\|$ is unique. Aside from the different choices of an orthonormal basis of these subspaces, the singular vectors are uniquely defined. For example, if all singular values are distinct, then the subspace of unit vectors that maximize $\|Av\|$ is one-dimensional and the singular vector is unique (up to sign changes, i.e., up to multiplication by -1).

Sometimes, it is helpful to observe that the matrix product $UDV^\intercal$ can also be written as the sum of outer products of the singular vectors. This formulation has the advantage that we can write the projection of A to the best fit subspaces of dimension k as the sum of the first k terms.

Remark 21.6. The SVD can equivalently be written as

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

where $u_i v_i^\intercal$ is the outer product. For $k \leq r$, the projection of A to V_k is

$$A_k := \sum_{i=1}^{k} \sigma_i u_i v_i^\intercal.$$

Recall that the Frobenius norm of a matrix A is the square root of the sum of its squared entries, i.e. it is defined by $\|A\|_F := \sqrt{\sum_{i,j} a_{ij}^2}$. This means that $\|A - B\|_F^2$ is equal to the sum of the squared distances between each row in A and the corresponding row in B for matrices of equal dimensions. Imagine that B is a rank k matrix. Then its points lie within a k-dimensional subspace, and $\|A - B\|_F^2$ cannot be smaller than the distance between A and this subspace. Since A_k is the projection to the best fit subspace of dimension k, A_k minimizes $\|A - B\|_F$ (notice that A_k has rank at most k). It is therefore also called the best rank k-approximation of A.

Theorem 21.7. *Let $A \in \mathbb{R}^{n \times d}$ be a matrix of rank r and let $k \leq r$ be given. It holds that*

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

for any matrix $B \in \mathbb{R}^{n \times d}$ of rank at most k.

The theorem is also true if the Frobenius norm is replaced by the *spectral norm*.[2] For a matrix A, the spectral norm is equal to the maximum singular value, i.e. $\|A\|_2 := \max_{v \in \mathbb{R}^d, \|v\|=1} \|Av\| = \sigma_1$.

4 Applications

Topic modeling. Replacing A by A_k is a great compression idea. For example, for *topic modeling*, we imagine A to be a matrix that stores the number of times that any of d words appears in any of n documents. Then we assume that the rank $r$of A corresponds to r *topics*. Recall that

$$A = \begin{pmatrix} — & a_1 & — \\ & \vdots & \\ — & a_n & — \end{pmatrix} = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{pmatrix} \begin{pmatrix} — & v_1 & — \\ & \vdots & \\ — & v_d & — \end{pmatrix}.$$

Assume that the entries in U and V are positive. Since the column vectors are unit vectors, they define a convex combination of the r topics. We can thus imagine U to contain information on how much each of the documents consists of each topic. Then, D assigns a weight to each of the topics. Finally, we $V^\intercal$ gives information on how much each topic consists of each of the words. The combination of the three matrices generates the actual documents. By using the SVD, we can represent a set of documents based on fewer topics, thus obtaining an easier model of how they are generated.

Notice that this interpretation of the SVD needs that the entries are non negative, and that obtaining such a decomposition is an NP-hard problem.

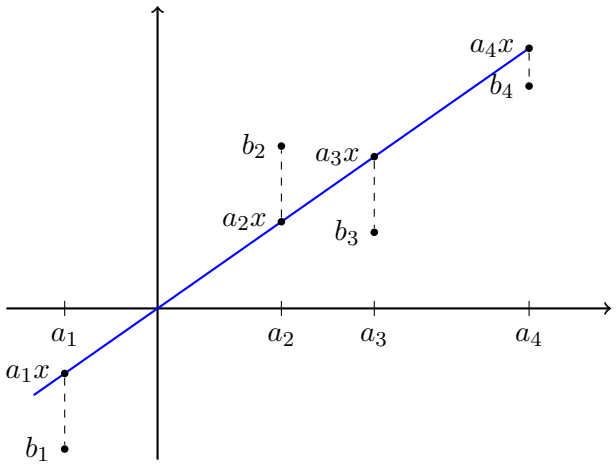

Figure 21.3

4.1 Pseudoinverse and least squares regression.

For any diagonal matrix $M = \text{diag}(d_1, \ldots, d_\ell)$, define $M^+ := \text{diag}(1/d_1, \ldots, 1/d_\ell)$. We notice that for the matrices from the SVD, it holds that

$$VD^+U^\intercal UDV = \text{diag}(\underbrace{1, \ldots, 1}_{r \text{ times}}, 0, \ldots, 0).$$

If A is an $n \times n$-matrix of rank n, then $r = n$ and the result of this product is I. Thus, $A^+ := VD^+U^\intercal$ is then the inverse of A. In general, A^+ is the (Moore Penrose) *pseudoinverse* of A. It satisfies that

$$A(A^+b) = b \quad \forall b \text{ in the image of } A$$

The pseudoinverse helps to find the solution to another popular minimization problem, *least squares regression*. Given an overconstrained system of equations $Ax = b$, least squares regression asks for a point x that minimizes the squared error $\|Ax - b\|_2^2$. I.e., we want

$$x^* := \arg\min \|Ax - b\|_2^2.$$

Notice that if there is an x' with $Ax' = b$, then it also minimizes $\|Ax' - b\|_2^2$, and if A had full rank this x' would be obtained by computing $A^{-1}b$. If A does not have full rank, an optimal solution is obtained by using the pseudoinverse:

$$x^* = A^+b$$

(This is often used as another definition for the pseudoinverse.)

Here's a proof: for any choice of $x \in \mathbb{R}^d$, Ax is some point in the column span of A. So x^*, the minimizer, must be the projection of b onto colspan(A). One orthonormal basis for colspan(A) is the columns of U. Hence the projection Πb of b onto colspan(A) is given by $UU^\intercal b$. (Why? Extend U to a basis for all of $\mathbb{R}^d$, write b in this basis, and consider what it's projection must be.) Hence we want $Ax^* = UU^\intercal b$. For this, it suffices to set $x^* = VD^+U^\intercal b = A^+b$.

[2] In fact, this theorem holds for any *unitarily invariant matrix norm*; a matrix norm $\|\cdot\|$ is unitarily invariant if $\|A\| = \|UAV\|$ for any unitary matrices U, V. Other examples of unitarily invariant norms are the Schatten norms, and the Ky Fan norms. J. von Neumann characterized all unitarily invariant matrix norms as those obtained by taking a "symmetric" (vector) norm of the vector of singular values — here symmetric means $\|x\| = \|y\|$ when y is obtained by flipping the signs of some entries of x and then permuting them around. See [HJ85, Theorem 7.4.24].

5 Symmetric Matrices

For a (square) symmetric matrix A, the (normalized) eigenvectors v_i and the eigenvalues λ_i satisfy the following properties: the v_is form an orthonormal basis, and $A = V\Lambda V^\intercal$, where the columns of V are the v_i vectors, and Λ is a diagonal matrix with λ_is on the diagonal. It is no longer the case that the eigenvalues are all non-negative. (In fact, we can match up the eigenvalues and singular values such that they differ only in sign.)

Given a function $f : \mathbb{R} \to \mathbb{R}$, we can extend this to a function on symmetric matrices as follows:

$$f(A) = V \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_n)) V^\intercal.$$

For instance, you can check that A^k or e^A defined this way indeed correspond to what you think they might mean. (The other way to define e^A would be $\sum_{k\geq 0} \frac{A^k}{k!}$.)

References

[HJ85] Roger A. Horn and Charles R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985. MR 832183 (87e:15001) 2

[HK14] John Hopcroft and Ravindran Kannan, *Foundations of data science (version 21/8/20014), chapter 3*, http://research.microsoft.com/en-US/people/kannan/book-no-solutions-aug-21-2014.pdf#page=52, 2014, accessed: 03/04/2015. 1

[Ste93] Gilbert W. Stewart, *On the early history of the singular value decomposition*, SIAM Review **35** (1993), 551 – 566. 1