Lecture 4. The Singular Value Decomposition

The singular value decomposition (SVD) is a matrix factorization whose computation is a step in many algorithms. Equally important is the use of the SVD for conceptual purposes. Many problems of linear algebra can be better understood if we first ask the question, what if we take the SVD?

A Geometric Observation

The SVD is motivated by the following geometric fact:

The image of the unit sphere under any $m \times n$ matrix is a hyperellipse.

The SVD is applicable to both real and complex matrices. However, in describing the geometric interpretation, we assume as usual that the matrix is real.

The term "hyperellipse" may be unfamiliar, but this is just the m-dimensional generalization of an ellipse. We may define a hyperellipse in $\mathbb{R}^m$ as the surface obtained by stretching the unit sphere in $\mathbb{R}^m$ by some factors $\sigma_1, \ldots, \sigma_m$, possibly zero, in some orthogonal directions $u_1, \ldots, u_m \in \mathbb{R}^m$. For convenience, let us take the u_i to be unit vectors. The vectors $\{\sigma_i u_i\}$ are the *principal semiaxes* of the hyperellipse, with lengths $\sigma_1, \ldots, \sigma_m$. If A has rank r, exactly r of the lengths σ_i will turn out to be nonzero, and in particular, if $m \geq n$, at most n of them will be nonzero.

The italicized statement above has the following meaning. By the unit sphere, we mean the usual Euclidean sphere in n-space, i.e., the unit sphere in the 2-norm; let us denote it by S. Then AS, the image of S under the mapping A, is a hyperellipse as just defined.

This geometric fact is not obvious. We shall restate it in the language of linear algebra and prove it later. For the moment, assume it is true.

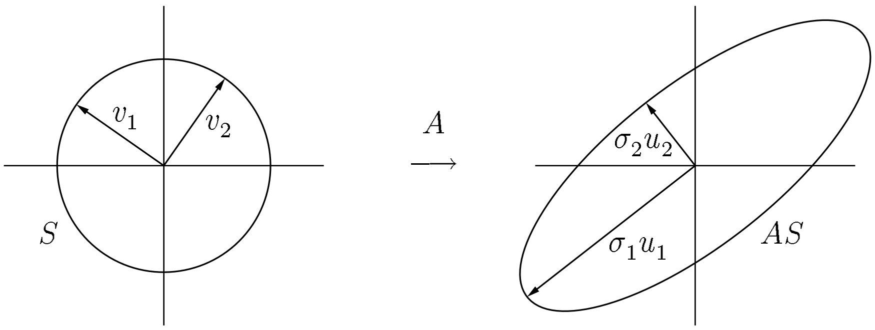


Figure 4.1. *SVD of a* 2×2 *matrix.*

Let S be the unit sphere in $\mathbb{R}^n$, and take any $A \in \mathbb{R}^{m \times n}$ with $m \geq n$. For simplicity, suppose for the moment that A has full rank n. The image AS is a hyperellipse in $\mathbb{R}^m$. We now define some properties of A in terms of the shape of AS. The key ideas are indicated in Figure 4.1.

First, we define the n *singular values* of A. These are the lengths of the n principal semiaxes of AS, written $\sigma_1, \sigma_2, \ldots, \sigma_n$. It is conventional to assume that the singular values are numbered in descending order, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$.

Next, we define the n *left singular vectors* of A. These are the unit vectors $\{u_1, u_2, \ldots, u_n\}$ oriented in the directions of the principal semiaxes of AS, numbered to correspond with the singular values. Thus the vector $\sigma_i u_i$ is the ith largest principal semiaxis of AS.

Finally, we define the n *right singular vectors* of A. These are the unit vectors $\{v_1, v_2, \ldots, v_n\} \in S$ that are the preimages of the principal semiaxes of AS, numbered so that $Av_j = \sigma_j u_j$.

The terms "left" and "right" in the definitions above are decidedly awkward. They come from the positions of the factors U and V in (4.2) and (4.3), below. What is awkward is that in a sketch like Figure 4.1, the left singular vectors correspond to the space on the right and the right singular vectors correspond to the space on the left! One could resolve this problem by interchanging the two halves of the figure, with the map A pointing from right to left, but that would go against deeply-ingrained habits.

Reduced SVD

We have just mentioned that the equations relating right singular vectors $\{v_j\}$ and left singular vectors $\{u_j\}$ can be written

$$Av_j = \sigma_j u_j, \qquad 1 \le j \le n. \tag{4.1}$$

This collection of vector equations can be expressed as a matrix equation,

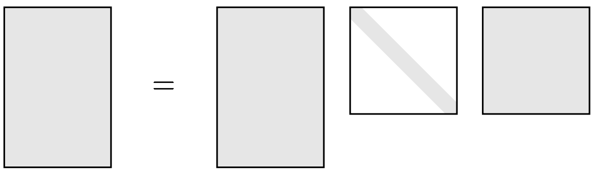
$$\left[\begin{array}{c} \\ \\ \quad\quad A \quad\quad \\ \\ \\ \end{array}\right] \left[\begin{array}{c|c|c|c} & & & \\ v_1 & v_2 & \cdots & v_n \\ & & & \end{array}\right] = \left[\begin{array}{c|c|c|c} & & & \\ & & & \\ u_1 & u_2 & \cdots & u_n \\ & & & \\ & & & \end{array}\right] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix},$$

or, more compactly, $AV = \hat{U}\hat{\Sigma}$. In this matrix equation, $\hat{\Sigma}$ is an $n \times n$ diagonal matrix with positive real entries (since A was assumed to have full rank n), $\hat{U}$ is an $m \times n$ matrix with orthonormal columns, and V is an $n \times n$ matrix with orthonormal columns. Thus V is unitary, and we can multiply on the right by its inverse V^* to obtain

$$A = \hat{U}\hat{\Sigma}V^*. \tag{4.2}$$

This factorization of A is called a *reduced singular value decomposition*, or *reduced SVD*, of A. Schematically, it looks like this:

Reduced SVD $(m \ge n)$



A $\qquad$ $\hat{U}$ $\qquad$ $\hat{\Sigma}$ $\qquad$ V^*

Full SVD

In most applications, the SVD is used in exactly the form just described. However, this is not the standard way in which the idea of an SVD is usually formulated. We have introduced the awkward term "reduced" and the unsightly hats on U and Σ in order to distinguish the factorization (4.2) from the more standard "full" SVD. This "reduced" vs. "full" terminology and hatted notation will be maintained throughout the book, and we shall make a similar distinction between reduced and full QR factorizations.

The idea is as follows. The columns of $\hat{U}$ are n orthonormal vectors in the m-dimensional space $\mathbb{C}^m$. Unless $m = n$, they do not form a basis of

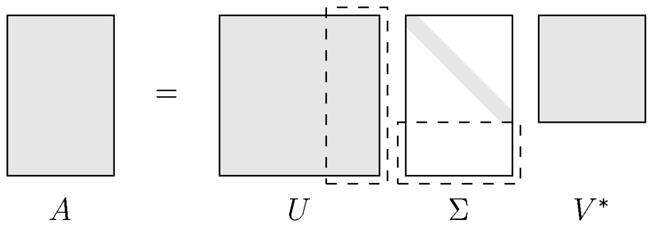
$\mathbb{C}^m$, nor is $\hat{U}$ a unitary matrix. However, by adjoining an additional $m - n$ orthonormal columns, $\hat{U}$ can be extended to a unitary matrix. Let us do this in an arbitrary fashion, and call the result U.

If $\hat{U}$ is replaced by U in (4.2), then $\hat{\Sigma}$ will have to change too. For the product to remain unaltered, the last $m-n$ columns of U should be multiplied by zero. Accordingly, let Σ be the $m \times n$ matrix consisting of $\hat{\Sigma}$ in the upper $n \times n$ block together with $m - n$ rows of zeros below. We now have a new factorization, the *full SVD* of A:

$$A = U\Sigma V^*. \tag{4.3}$$

Here U is $m \times m$ and unitary, V is $n \times n$ and unitary, and Σ is $m \times n$ and diagonal with positive real entries. Schematically:

Full SVD ($m \geq n$)



The dashed lines indicate the "silent" columns of U and rows of Σ that are discarded in passing from (4.3) to (4.2).

Having described the full SVD, we can now discard the simplifying assumption that A has full rank. If A is rank-deficient, the factorization (4.3) is still appropriate. All that changes is that now not n but only r of the left singular vectors of A are determined by the geometry of the hyperellipse. To construct the unitary matrix U, we introduce $m - r$ instead of just $m - n$ additional arbitrary orthonormal columns. The matrix V will also need $n - r$ arbitrary orthonormal columns to extend the r columns determined by the geometry. The matrix Σ will now have r positive diagonal entries, with the remaining $n - r$ equal to zero.

By the same token, the reduced SVD (4.2) also makes sense for matrices A of less than full rank. One can take $\hat{U}$ to be $m \times n$, with $\hat{\Sigma}$ of dimensions $n \times n$ with some zeros on the diagonal, or further compress the representation so that $\hat{U}$ is $m \times r$ and $\hat{\Sigma}$ is $r \times r$ and strictly positive on the diagonal.

Formal Definition

Let m and n be arbitrary; we do not require $m \geq n$. Given $A \in \mathbb{C}^{m \times n}$, not necessarily of full rank, a *singular value decomposition* (SVD) of A is a factorization

$$A = U\Sigma V^* \tag{4.4}$$

where

$$\begin{aligned} U &\in \mathbb{C}^{m\times m} && \text{is unitary,}\\ V &\in \mathbb{C}^{n\times n} && \text{is unitary,}\\ \Sigma &\in \mathbb{R}^{m\times n} && \text{is diagonal.} \end{aligned}$$

In addition, Σ is assumed to have its diagonal entries σ_j nonnegative and in nonincreasing order; that is, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$, where $p = \min(m, n)$.

Note that the diagonal matrix Σ has the same shape as A even when A is not square, but U and V are always square unitary matrices.

It is clear that the image of the unit sphere in $\mathbb{R}^n$ under a map $A = U\Sigma V^*$ must be a hyperellipse in $\mathbb{R}^m$. The unitary map V^* preserves the sphere, the diagonal matrix Σ stretches the sphere into a hyperellipse aligned with the canonical basis, and the final unitary map U rotates or reflects the hyperellipse without changing its shape. Thus, if we can prove that every matrix has an SVD, we shall have proved that the image of the unit sphere under any linear map is a hyperellipse, as claimed at the outset of this lecture.

Existence and Uniqueness

Theorem 4.1. *Every matrix $A \in \mathbb{C}^{m\times n}$ has a singular value decomposition (4.4). Furthermore, the singular values $\{\sigma_j\}$ are uniquely determined, and, if A is square and the σ_j are distinct, the left and right singular vectors $\{u_j\}$ and $\{v_j\}$ are uniquely determined up to complex signs (i.e., complex scalar factors of absolute value 1).*

Proof. To prove existence of the SVD, we isolate the direction of the largest action of A, and then proceed by induction on the dimension of A.

Set $\sigma_1 = \|A\|_2$. By a compactness argument, there must be a vector $v_1 \in \mathbb{C}^n$ with $\|v_1\|_2 = 1$ and $\|u_1\|_2 = \sigma_1$, where $u_1 = Av_1$. Consider any extensions of v_1 to an orthonormal basis $\{v_j\}$ of $\mathbb{C}^n$ and of u_1 to an orthonormal basis $\{u_j\}$ of $\mathbb{C}^m$, and let U_1 and V_1 denote the unitary matrices with columns u_j and v_j, respectively. Then we have

$$U_1^* A V_1 = S = \begin{bmatrix} \sigma_1 & w^* \\ 0 & B \end{bmatrix}, \tag{4.5}$$

where 0 is a column vector of dimension $m-1$, w^* is a row vector of dimension $n-1$, and B has dimensions $(m-1)\times(n-1)$. Furthermore,

$$\left\| \begin{bmatrix} \sigma_1 & w^* \\ 0 & B \end{bmatrix} \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2 \geq \sigma_1^2 + w^*w = (\sigma_1^2 + w^*w)^{1/2} \left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|_2,$$

implying $\|S\|_2 \geq (\sigma_1^2 + w^*w)^{1/2}$. Since U_1 and V_1 are unitary, we know that $\|S\|_2 = \|A\|_2 = \sigma_1$, so this implies $w = 0$.

If $n = 1$ or $m = 1$, we are done. Otherwise, the submatrix B describes the action of A on the subspace orthogonal to v_1. By the induction hypothesis, B has an SVD $B = U_2\Sigma_2V_2^*$. Now it is easily verified that

$$A = U_1 \begin{bmatrix} 1 & 0 \\ 0 & U_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V_2 \end{bmatrix}^* V_1^*$$

is an SVD of A, completing the proof of existence.

For the uniqueness claim, the geometric justification is straightforward: if the semiaxis lengths of a hyperellipse are distinct, then the semiaxes themselves are determined by the geometry, up to signs. Algebraically, we can argue as follows. First we note that σ_1 is uniquely determined by the condition that it is equal to $\|A\|_2$, as follows from (4.4). Now suppose that in addition to v_1, there is another linearly independent vector w with $\|w\|_2 = 1$ and $\|Aw\|_2 = \sigma_1$. Define a unit vector v_2, orthogonal to v_1, as a linear combination of v_1 and w,

$$v_2 = \frac{w - (v_1^*w)v_1}{\|w - (v_1^*w)v_1\|_2}.$$

Since $\|A\|_2 = \sigma_1$, $\|Av_2\|_2 \le \sigma_1$; but this must be an equality, for otherwise, since $w = v_1c + v_2s$ for some constants c and s with $|c|^2 + |s|^2 = 1$, we would have $\|Aw\|_2 < \sigma_1$. This vector v_2 is a second right singular vector of A corresponding to the singular value σ_1; it will lead to the appearance of a vector y (equal to the last $n-1$ components of $V_1^*v_2$) with $\|y\|_2 = 1$ and $\|By\|_2 = \sigma_1$. We conclude that, if the singular vector v_1 is not unique, then the corresponding singular value σ_1 is not simple. To complete the uniqueness proof we note that as indicated above, once σ_1 and v_1 and u_1 are determined, the remainder of the SVD is determined by the action of A on the space orthogonal to v_1. Since v_1 is unique up to sign, this orthogonal space is uniquely defined, and the uniqueness of the remaining singular values and vectors now follows by induction. □

Exercises

1. Determine SVDs of the following matrices (by hand calculation):

(a) $\begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}$, (b) $\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$, (c) $\begin{bmatrix} 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, (d) $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$, (e) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

2. Suppose A is an $m \times n$ matrix and B is the $n \times m$ matrix obtained by rotating A ninety degrees clockwise on paper (not exactly a standard mathematical

transformation!). Do A and B have the same singular values? Prove it or give a counterexample.

3. Write a MATLAB program (see Lecture 9) which, given a real 2×2 matrix A, plots the right singular vectors v_1 and v_2 in the unit circle and also the left singular vectors u_1 and u_2 in the appropriate ellipse, as in Figure 4.1. Apply your program to the matrix (3.7) and also to the 2×2 matrices in Exercise 1.

4. Two matrices $A, B \in \mathbb{C}^{m\times m}$ are *unitarily equivalent* if $A = QBQ^*$ for some unitary $Q \in \mathbb{C}^{m\times m}$. Is it true or false that A and B are unitarily equivalent if and only if they have the same singular values?

5. Theorem 4.1 asserts that every $A \in \mathbb{C}^{m\times n}$ has an SVD $A = U\Sigma V^*$. Show that if A is real, then it has a real SVD ($U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$).