CME 302: NUMERICAL LINEAR ALGEBRA
FALL 2005/06
LECTURE 3

GENE H. GOLUB

1. Singular Value Decomposition

Suppose A is an $m \times n$ real matrix with $m \geq n$. Then we can write

$$A = U\Sigma V^\top,$$

where

$$U^\top U = I_m, \quad V^\top V = I_n, \quad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & 0 & \end{bmatrix}.$$

The diagonal elements σ_i, $i = 1, \ldots, n$, are all nonnegative, and can be ordered such that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \quad \sigma_{r+1} = \cdots = \sigma_n = 0$$

where r is the rank of A. This decomposition of A is called the *singular value decomposition*, or SVD. The values σ_i, for $i = 1, 2, \ldots, n$, are the *singular values* of A. The columns of U are the *left singular vectors*, and the columns of V are the *right singular vectors*.

An alternative decomposition of A omits the singular values that are equal to zero:

$$A = \tilde{U}\tilde{\Sigma}\tilde{V}^\top,$$

where $\tilde{U}$ is an $m \times r$ matrix satisfying $\tilde{U}^\top \tilde{U} = I_r$, $\tilde{V}$ is an $n \times r$ matrix satisfying $\tilde{V}^\top \tilde{V} = I_r$, and $\tilde{\Sigma}$ is an $r \times r$ diagonal matrix with diagonal elements $\sigma_1, \ldots, \sigma_r$. The columns of $\tilde{U}$ are the left singular vectors corresponding to the nonzero singular values of A, and form an orthogonal basis for the range of A. The columns of $\tilde{V}$ are the right singular vectors corresponding to the nonzero singular values of A, and are each orthogonal to the null space of A.

Summarizing, the SVD of an $m \times n$ real matrix A is $A = U\Sigma V^\top$ where U is an $m \times m$ orthogonal matrix, V is an $n \times n$ orthogonal matrix, and Σ is an $m \times n$ diagonal matrix with nonnegative diagonal elements; the *condensed* SVD of A is $A = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$ where $\tilde{U}$ is an $m \times r$ matrix with orthogonal columns, $\tilde{V}$ is an $n \times r$ matrix with orthogonal columns, and $\tilde{\Sigma}$ is an $r \times r$ diagonal matrix with positive diagonal elements equal to the nonzero diagonal elements of Σ. The number r is the rank of A.

If A is an $m \times m$ matrix and $\sigma_m > 0$, then

$$A^{-1} = (V^\top)^{-1}\Sigma^{-1}U^{-1} = V\Sigma^{-1}U^\top.$$

We will see that this representation of the inverse can be used to obtain a *pseudo-inverse* (also called *generalized inverse* or *Moore-Penrose inverse*) of a matrix A in the case where A does not actually have an inverse.

We now mention some additional properties of the singular values and singular vectors. We have

$$A^\top A = V\Sigma^\top U^\top U\Sigma V^\top = V(\Sigma^\top \Sigma)V^\top.$$

Date: October 26, 2005, version 1.0.
Notes originally due to James Lambers. Minor editing by Lek-Heng Lim.

The matrix $\Sigma^\top \Sigma$ is a diagonal matrix with diagonal elements σ_i^2, $i = 1, \ldots, n$, which are also the eigenvalues of $A^\top A$, with corresponding eigenvectors v_i, where v_i is the ith column of V. Similarly, $AA^\top = U\Sigma\Sigma^\top U^\top$, from which we can easily see that the columns of U are eigenvectors of $AA^\top$, corresponding to the eigenvalues σ_i^2, $i = 1, \ldots, n$.

Earlier we had shown that

$$\|A\|_2 = [\rho(A^\top A)]^{1/2}.$$

Since the eigenvalues of $A^\top A$ are simply the squares of the singular values of A, we can also say that

$$\|A\|_2 = \sigma_1.$$

Another way to arrive at this same conclusion is to take advantage of the fact that the 2-norm of a vector is invariant under multiplication by an orthogonal matrix, i.e. if $Q^\top Q = I$, then $\|\mathbf{x}\|_2 = \|Q\mathbf{x}\|_2$. Therefore

$$\|A\|_2 = \|U\Sigma V^\top\|_2 = \|\Sigma\|_2 = \sigma_1.$$

2. Least squares, pseudo-inverse, and projection

The singular value decomposition is very useful in studying the linear least squares problem. Suppose that we are given an m-vector $\mathbf{b}$ and an $m \times n$ matrix A, and we wish to find $\mathbf{x}$ such that

$$\|\mathbf{b} - A\mathbf{x}\|_2 = \text{minimum}.$$

From the SVD of A, we can simplify this minimization problem as follows:

$$\begin{aligned}
\|\mathbf{b} - A\mathbf{x}\|_2^2 &= \|\mathbf{b} - U\Sigma V^\top \mathbf{x}\|_2^2 \\
&= \|U^\top \mathbf{b} - \Sigma V^\top \mathbf{x}\|_2^2 \\
&= \|\mathbf{c} - \Sigma \mathbf{y}\|_2^2 \\
&= (c_1 - \sigma_1 y_1)^2 + \cdots + (c_r - \sigma_r y_r)^2 + \\
& c_{r+1}^2 + \cdots + c_m^2
\end{aligned}$$

where $\mathbf{c} = U^\top \mathbf{b}$ and $\mathbf{y} = V^\top \mathbf{x}$. We see that in order to minimize $\|A\mathbf{x} - \mathbf{b}\|_2$, we must set $y_i = c_i/\sigma_i$ for $i = 1, \ldots, r$, but the unknowns y_i, for $i = r+1, \ldots, m$, can have any value, since they do not influence $\|\mathbf{c} - \Sigma\mathbf{y}\|_2$. Therefore, if A does not have full rank, there are infinitely many solutions to the least squares problem. However, we can easily obtain the unique solution of minimum 2-norm by setting $y_{r+1} = \cdots = y_m = 0$.

In summary, the solution of minimum length to the linear least squares problem is

$$\mathbf{x} = V\mathbf{y} = V\Sigma^+\mathbf{c} = V\Sigma^+U^\top\mathbf{b} = A^+\mathbf{b}$$

where Σ^+ is a diagonal matrix with entries

$$\Sigma^+ = \begin{bmatrix} \sigma_1^{-1} & & & & & \\ & \ddots & & & & \\ & & \sigma_r^{-1} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

and $A^+ = V\Sigma^+U^\top$. The matrix A^+ is called the *pseudo-inverse* of A. In the case where A has full rank, the pseudo-inverse is equal to A^{-1}. Note that A^+ is independent of $\mathbf{b}$.

The solution $\mathbf{x}$ of the least-squares problem minimizes $\|A\mathbf{x} - \mathbf{b}\|$, and therefore is the vector that solves the system $A\mathbf{x} = \mathbf{b}$ as closely as possible. However, we can use the SVD to show that $\mathbf{x}$ is the exact solution to a related system of equations.

We write $\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2$, where

$$\mathbf{b}_1 = AA^+\mathbf{b}, \quad \mathbf{b}_2 = (I - AA^+)\mathbf{b}.$$

The matrix AA^+ has the form

$$AA^+ = U\Sigma V^\top V\Sigma^+ U^\top = U\Sigma\Sigma^+ U^\top = U \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} U^\top.$$

It follows that $\mathbf{b}_1$ is a linear combination of $\mathbf{u}_1, \dots, \mathbf{u}_r$, the columns of U that form an orthogonal basis for the range of A. From $\mathbf{x} = A^+\mathbf{b}$ we obtain

$$A\mathbf{x} = AA^+\mathbf{b} = P\mathbf{b} = \mathbf{b}_1,$$

where $P = AA^+$. Therefore, the solution to the least squares problem, is also the exact solution to the system $A\mathbf{x} = P\mathbf{b}$.

It can be shown that the matrix P has the properties

(1) $P = P^\top$
(2) $P^2 = P$

In other words, the matrix P is a *projection.* In particular, it is a projection onto the space spanned by the columns of A, i.e. the range of A.

3. Some properties of the SVD

While the previous discussion assumed that A was a real matrix, the SVD exists for complex matrices as well. In this case, the decomposition takes the form

$$A = U\Sigma V^*$$

where, for general A, $A^* = \bar{A}^\top$, the complex conjugate of the transpose. A^* is often written as A^H, and is equivalent to the transpose for real matrices.

Using the SVD, we can easily establish a lower bound for the largest singular value σ_1 of A, which also happens to be equal to $\|A\|_2$, as previously discussed. First, let us consider the case where A is symmetric and positive definite. Then, we can write $A = U\Lambda U^*$ where U is an orthogonal matrix and Λ is a diagonal matrix with real and positive diagonal elements $\lambda_1 \geq \cdots \geq \lambda_n$ which are the eigenvalues of A. We can then write

$$\max \frac{x^*Ax}{x^*x} = \max \frac{\mathbf{x}^*U\Lambda U^*\mathbf{x}}{\mathbf{x}^*UU^*\mathbf{x}} = \max \frac{\mathbf{y}^*\Lambda\mathbf{y}}{\mathbf{y}^*\mathbf{y}} \leq \lambda_1$$

where $\mathbf{y} = U^*\mathbf{x}$. Now, we consider the case of general A and try to find an upper bound for the expression

$$\max_{\mathbf{u},\mathbf{v}\neq 0} \frac{|\mathbf{u}^*A\mathbf{v}|}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}.$$

We have, by the Cauchy-Schwarz inequality,

$$\max_{\mathbf{u},\mathbf{v}\neq 0} \frac{|\mathbf{u}^*A\mathbf{v}|}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} = \max_{\mathbf{u},\mathbf{v}\neq 0} \frac{|\mathbf{u}^*U\Sigma V^*\mathbf{v}|}{\|U^*\mathbf{u}\|_2\|V^*\mathbf{v}\|_2} = \max_{\mathbf{x},\mathbf{y}\neq 0} \frac{|\mathbf{x}^*\Sigma\mathbf{y}|}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2} \leq \sigma_1 \max_{\mathbf{x},\mathbf{y}\neq 0} \frac{|\mathbf{x}^*\mathbf{y}|}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2} \leq \sigma_1.$$

4. Existence of SVD

We will now prove the existence of the SVD. We define

$$\tilde{A} = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}.$$

It is easy to verify that $\tilde{A} = \tilde{A}^*$, and therefore $\tilde{A}$ has the decomposition $\tilde{A} = Z\Lambda Z^*$ where Z is an orthogonal matrix and Λ is a diagonal matrix with real diagonal elements. If $\mathbf{z}$ is a column of Z, then we can write

$$\tilde{A}\mathbf{z} = \sigma\mathbf{z}, \quad \mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$$

and therefore

$$\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \sigma \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$$

or, equivalently,

$$A\mathbf{y} = \sigma\mathbf{x}, \quad A^*\mathbf{x} = \sigma\mathbf{y}.$$

Now, suppose that we apply $\tilde{A}$ to the vector obtained from $\mathbf{z}$ by negating $\mathbf{y}$. Then we have

$$\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ -\mathbf{y} \end{bmatrix} = \begin{bmatrix} -A\mathbf{y} \\ A^*\mathbf{x} \end{bmatrix} = \begin{bmatrix} -\sigma\mathbf{x} \\ \sigma\mathbf{y} \end{bmatrix} = -\sigma \begin{bmatrix} \mathbf{x} \\ -\mathbf{y} \end{bmatrix}.$$

In other words, if $\sigma \neq 0$ is an eigenvalue, then $-\sigma$ is also an eigenvalue.

Suppose that we normalize the eigenvector $\mathbf{z}$ of $\tilde{A}$ so that $\mathbf{z}^*\mathbf{z} = 2$. Since $\tilde{A}$ is symmetric, eigenvectors corresponding to different eigenvalues are orthogonal, so it follows that

$$\begin{bmatrix} \mathbf{x}^* & \mathbf{y}^* \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = 0.$$

This yields the system of equations

$$\begin{aligned} \mathbf{x}^*\mathbf{x} + \mathbf{y}^*\mathbf{y} &= 2 \\ \mathbf{x}^*\mathbf{x} - \mathbf{y}^*\mathbf{y} &= 0 \end{aligned}$$

which has the unique solution

$$\mathbf{x}^*\mathbf{x} = 1, \quad \mathbf{y}^*\mathbf{y} = 1.$$

From the relationships $A\mathbf{y} = \sigma\mathbf{x}$, $A^*\mathbf{x} = \mathbf{y}$, we obtain

$$A^*A\mathbf{y} = \sigma^2\mathbf{y}, \quad AA^*\mathbf{x} = \sigma^2\mathbf{x}.$$

Therefore, if $\pm\sigma$ are eigenvalues of $\tilde{A}$, then σ^2 is an eigenvalue of both AA^* and A^*A.

To complete the proof, we note that we can represent the matrix of normalized eigenvectors of $\tilde{A}$ corresponding to nonzero eigenvalues as

$$\tilde{Z} = \frac{1}{\sqrt{2}} \begin{bmatrix} X & X \\ Y & -Y \end{bmatrix}.$$

It follows that

$$\begin{aligned} \tilde{A} &= \tilde{Z}\Lambda\tilde{Z}^* \\ &= \frac{1}{2} \begin{bmatrix} X & X \\ Y & -Y \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \\ 0 & -\Sigma_r \end{bmatrix} \begin{bmatrix} X^* & Y^* \\ X^* & -Y^* \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} X\Sigma_r & -X\Sigma_r \\ Y\Sigma_r & Y\Sigma_r \end{bmatrix} \begin{bmatrix} X^* & Y^* \\ X^* & -Y^* \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} 0 & 2X\Sigma_r Y^* \\ 2Y\Sigma_r X^* & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & X\Sigma_r Y^* \\ Y\Sigma_r X^* & 0 \end{bmatrix} \end{aligned}$$

and therefore

$$A = X\Sigma_r Y^*, \quad A^* = Y\Sigma_r X^*$$

where X is an $m \times r$ matrix, Σ is $r \times r$, and Y is $n \times r$, and r is the rank of A. This represents the "condensed" SVD.

Department of Computer Science, Gates Building 2B, Room 280, Stanford, CA 94305-9025
E-mail address: `golub@stanford.edu`