Singular Value Decomposition (奇异值分解)

Lecture 29

Dept. of Math., SUSTech

2023.12

Singular Value Decomposition

Introduction

$U\Sigma V^T$ joins with LU from elimination and QR from orthogonalization (Gauss and Gram-Schmidt). Nobody's name is attached: $A = U\Sigma V^T$ is known as the "SVD" or the singular value decomposition. We want to describe it , to prove it, and to discuss its applications – which are many and growing.

The SVD is closely associated with the eigenvalue-eigenvector factorization $Q\Lambda Q^T$ of a positive definite matrix.

- The diagonal matrix Σ has eigenvalues from A^TA, not from A! Those positive entries will be $\sigma_1, \sigma_2, \cdots, \sigma_r$. They are the singular values of A.
- They fill the first r places on the main diagonal of Σ–when A has rank r. The rest of Σ is zero.

Singular Value Decomposition

Every matrix can split into $A = U\Sigma V^T$. With rectangular matrices, the key is almost always to consider A^TA and AA^T.

Theorem

Any m by n matrix A can be factored into

$$A = U\Sigma V^T = (orthogonal)(diagonal)(orthogonal)$$

The columns of U (m by m) are eigenvectors of AA^T, and the columns of V (n by n) are eigenvectors of A^TA. The r singular values on the diagonal of Σ (m by n) are the square roots of the nonzero eigenvalues of both AA^T and A^TA.

Proof

Proof. A^TA is a symmetric $n \times n$ matrix. Therefore, its eigenvalues are all real and it has an orthogonal diagonalizing matrix V. Furthermore, its eigenvalues must all be nonnegative. To see this, let x be an eigenvector belonging to λ. It follows that

$$||Ax||^2 = x^T A^T Ax = \lambda x^T x = \lambda ||x||^2$$

Hence,

$$\lambda = \frac{||Ax||^2}{||x||^2} \geq 0$$

We may assume that the columns of V have been ordered so that the corresponding eigenvalues satisfy

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$$

The singular values of A are given by

$$\sigma_j = \sqrt{\lambda_j},\ j = 1, \cdots, n.$$

Let r denote the rank of A. The matrix $A^T A$ will also have rank r. Since $A^T A$ is symmetric, its rank equals the number of nonzero eigenvalues. Thus,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0 \text{ and } \lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_n = 0$$

The same relation holds for the singular values

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \text{ and } \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$$

Now let

$$V_1 = (v_1, v_2, \cdots, v_r),\ V_2 = (v_{r+1}, v_{r+2}, \cdots, v_n)$$

and

$$\Sigma_1 = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix}$$

Hence, Σ_1 is an $r \times r$ diagonal matrix whose diagonal entries are the nonzero singular values $\sigma_1, \sigma_2, \cdots, \sigma_r$. The the $m \times n$ matrix Σ is then given by

$$\begin{bmatrix} \Sigma_1 & O \\ O & O \end{bmatrix}.$$

Proof.

The column vectors of V_2 are eigenvectors of A^TA belonging to $\lambda = 0$. Thus

$$A^TAv_j = 0 \quad j = r+1, \cdots, n.$$

and, consequently, the column vectors of V_2 form an orthonormal basis for $N(A^TA) = N(A)$. Therefore,

$$AV_2 = O$$

And since V is an orthogonal matrix, it follows that

$$I = VV^T = V_1 V_1^T + V_2 V_2^T \tag{1}$$

$$A = AI = AV_1 V_1^T + AV_2 V_2^T = AV_1 V_1^T \tag{2}$$

So far we have shown how to construct the matrices V and Σ of the singular value decomposition. To complete the proof, we must show how to construct an $m \times m$ orthogonal matrix U such that

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

or, equivalently,

$$AV = U\Sigma \tag{3}$$

Comparing the first r columns of each side of (3), we see that

$$Av_j = \sigma_j u_j \qquad j = 1, \cdots, r$$

Thus, if we define

$$u_j = \frac{1}{\sigma_j} A v_j \qquad j = 1, \cdots, r \tag{4}$$

and

$$U_1 = (u_1, \cdots, u_r)$$

then it follows that

$$AV_1 = U_1 \Sigma_1 \tag{5}$$

The column vectors of U_1 form an orthonormal set, since

$$\begin{aligned} u_i^T u_j &= \left(\frac{1}{\sigma_i} v_i^T A^T\right)\left(\frac{1}{\sigma_j} A v_j\right) \qquad 1 \le i \le r, 1 \le j \le r \\ &= \frac{1}{\sigma_i \sigma_j} v_i^T (A^T A v_j) = \frac{\sigma_j}{\sigma_i} v_i^T v_j = \delta_{ij} \end{aligned}$$

It follows from (4) that each $u_j, 1 \le j \le r$, is in the column space of A. The dimension of the column space is r, so $u_1, \cdots, u_r$ form an orthonormal basis for $C(A)$. The vector space $C(A)^{\perp} = N(A^T)$ has dimension $m-r$. Let $\{u_{r+1}, u_{r+2}, \cdots, u_m\}$ be an orthonormal basis for $N(A^T)$ and set

$$U_2 = (u_{r+1}, u_{r+2}, \cdots, u_m),$$

$$U = [U_1 \ U_2]$$

It is easily can be seen that $u_1, \cdots, u_m$ form an orthonormal basis for $\mathbb{R}^m$. Hence, U is an orthogonal matrix. We still must show that $U\Sigma V^T$ actually equals A. This follows from (5) and (2), since

$$U\Sigma V^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & O \\ O & O \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 \Sigma_1 V_1^T = AV_1 V_1^T = A.$$

The proof is thus complete.

Observations

1: Special cases: For positive definite matrices, Σ is Λ and $U\Sigma V^T$ is identical to $Q\Lambda Q^T$. For other symmetric matrices, any negative eigenvalues in Λ become positive in Σ. For complex matrices, Σ remains real but U and V become unitary.

2: U and V give orthonormal bases for all four fundamental subspaces:

first	r	columns of U : column space of A
last	$m-r$	columns of U : left nullspace of A
first	r	columns of V : row space of A
last	$n-r$	columns of V : nullspace of A.

3: The SVD chooses those bases in an extremely special way. They are more than just orthonormal. When A multiplies a column v_j of V, it produces σ_j times a column of U. That comes directly from $AV = U\Sigma$, looked at a column at a time.

4: Eigenvectors of AA^T or A^TA must go into the columns of U and V:

$$AA^T = (U\Sigma V^T)(V\Sigma^T U^T) = U\Sigma\Sigma^T U^T$$

U must be the eigenvector matrix for AA^T. Similarly, $A^TA = V\Sigma^T\Sigma V^T$. The eigenvalue matrix in the middle is $\Sigma\Sigma^T$–which is m by m with $\sigma_1^2, \cdots, \sigma_r^2$ on the diagonal. Similarly, the V matrix must be the eigenvector matrix for A^TA.

5: We also see that $Av_j = \sigma_j u_j$. Here is the reason:

$$AA^T Av_j = \sigma_j^2 Av_j$$

Therefore, Av_j is an eigenvector of AA^T. Indeed, we have $AV = U\Sigma$.

Example

Example 1 This A has only one column: $\text{rank}(A) = r = 1$. Then Σ has only $\sigma_1 = 3$. The SVD is as follows:

$$A = \begin{bmatrix} -1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} = U_{3\times 3}\Sigma_{3\times 1}V_{1\times 1}^T$$

A^TA is 1 by 1, whereas AA^T is 3 by 3. They both have eigenvalue 9 (whose square root is the 3 in Σ). The two zero eigenvalues of AA^T leave some freedom for the eigenvectors in columns 2 and 3 of U. We kept that matrix orthogonal.

Example

Example 2 Now A has rank 2, and $AA^T = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ with $\lambda = 3$ and 1. The SVD is as follows:

$$A = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} = U\Sigma V^T$$

$$= \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

Notice $\sqrt{3}$ and $\sqrt{1}$. The columns of U are left singular vectors (unit eigenvectors of AA^T). The columns of V are right singular vectors (unit eigenvectors of A^TA).

Image Processing

Suppose a satellite takes a picture, and wants to send it to Earth. The picture may contain 1000 by 1000 “pixels”–a million little squares, each with a definite color. We can code the colors, and send back 1,000,000 numbers. It is better to find the essential information inside the 1000 by 1000 matrix, and send only that. Suppose we know the SVD. The key is in the singular values (in Σ). Typically, some σ’s are significant and others are extremely small. If we keep 20 and throw away 980, then we send only the corresponding 20 columns of U and V. It is a 25 to 1 compression.

We can do the matrix multiplication as columns time rows:

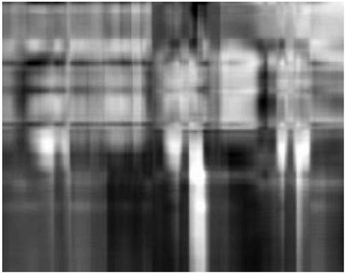
$$A = U\Sigma V^T = u_1\sigma_1 v_1^T + u_2\sigma_2 v_2^T + \cdots + u_r\sigma_r v_r^T$$

Digital Image Processing

The following figure shows an image corresponding to a 176×260 matrix A and three images corresponding to lower rank approximations of A. The gentlemen in the picture are (from left to right) James H. Wilkinson, Wallace Givens, and George Forsythe (three pioneers in the field of numerical linear algebra).

Original 176 by 260 Image

Rank 5 Approximation to Image



Rank 15 Approximation to Image

Rank 30 Approximation to Image

Courtesy Oak Ridge National Laboratory.

The pictures are really striking, as more and more singular values are included. At first you see nothing, and suddenly you recognize everything. The cost is in computing the SVD–this had become much more efficient, but it is expensive for a big matrix.

The Effective Rank

The rank of a matrix is the number of independent rows, and the number of independent columns. That can be hard to decide in computations!

Consider the following:

$$\begin{bmatrix} \varepsilon & 2\varepsilon \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \varepsilon & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \varepsilon & 1 \\ \varepsilon & 1+\varepsilon \end{bmatrix}$$

We go to a more stable measure of rank. The first step is to use A^TA or AA^T, which are symmetric but share the same rank of A. Their eigenvalues– the singular values squared–are not misleading. Based on the accuracy of the data, we decide on a tolerance like 10^{-6} and count the singular values above it–that is the effective rank. The examples above have effective rank 1 when ε is very small.

Polar Decomposition

- Every nonzero complex number z is a positive number r times a number $e^{i\theta}$ on the unit circle: $z = re^{i\theta}$. That expresses z in "polar coordinates."
- If we think of z as a 1 by 1 matrix, r corresponds to a positive definite matrix and $e^{i\theta}$ corresponds to an orthonormal matrix.
- More exactly, since $e^{i\theta}$ is complex and satisfies $e^{i\theta}e^{-i\theta} = 1$, it forms a 1 by 1 unitary matrix: $U^H U = I$.

Polar Factorization

The SVD extends the “polar factorization” to matrices of any size:

Theorem

Every real square matrix can be factored into $A = QS$, where Q is orthogonal and S is symmetric positive semidefinite. If A is invertible then S is positive definite.

Application of $A = QS$: A major use of the polar decomposition is in continuum mechanics. In any deformation, it is important to separate stretching from rotation, that is exactly what QS achieves. The orthogonal matrix Q is a rotation, and possibly a reflection.

The material feels no strain. The symmetric matrix S has eigenvalues $\sigma_1, \sigma_2, \cdots, \sigma_r$, which are the stretching factors or compression factors.

Examples

Example 3 Polar decomposition:

$$A = QS \qquad \begin{bmatrix} 1 & -2 \\ 3 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}$$

Example 4 Reverse polar decomposition:

$$A = S'Q \qquad \begin{bmatrix} 1 & -2 \\ 3 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

The exercises show how, in the reverse order, S changes but Q remains the same. Both S and S' are symmetric positive definite because this A is invertible.

Least Squares

For a rectangular system $Ax = b$, the least-squares solution comes from the normal equations $A^T A\hat{x} = A^T b$. If A has dependent columns then $A^T A$ is not invertible and $\hat{x}$ is not determined. We can now complete Chapter 3, by choosing a “best” (shortest) $\hat{x}$ for every $Ax = b$.

Proposition

The optimal solution of $Ax = b$ is the minimum length solution of $A^T A\hat{x} = A^T b$.

That minimum length solution will be called x^+. It is our preferred choice as the best solution to $Ax = b$ (which had no solution), and also to $A^T A\hat{x} = A^T b$ (which had too many).

Example

Example 5 A is diagonal, with dependent rows and dependent columns:

$$A\hat{x} = p \qquad \text{is} \qquad \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ 0 \end{bmatrix}$$

The columns all end with zero. In the column space, the closest vector to $b = (b_1, b_2, b_3)$ is $p = (b_1, b_2, 0)$. The best we can do with $Ax = b$ it to solve the first two equations, since the third equation is $0 = b_3$.

Minimum Length Solution

The minimum length solution is x^+:

$$x^+ = \begin{bmatrix} b_1/\sigma_1 \\ b_2/\sigma_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/\sigma_1 & 0 & 0 \\ 0 & 1/\sigma_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

Proof.

This equation finds x^+, and it also displays the matrix that produces x^+ from b. That matrix is the pseudoinverse A^+ of our diagonal A.

$$\Sigma_{m\times n} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \\ & & \end{bmatrix}, \quad \Sigma^+_{n\times m} = \begin{bmatrix} 1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_r \\ & & \end{bmatrix},$$

$$\Sigma^+ b = \begin{bmatrix} b_1/\sigma_1 \\ \vdots \\ b_r/\sigma_r \\ \end{bmatrix}$$

Figure 6.3

Remarks:

- The shortest solution x^+ is always in the row space of A.
- All we are doing is to choose that vector, $x^+ = x_r$, as the best solution to $Ax = b$.

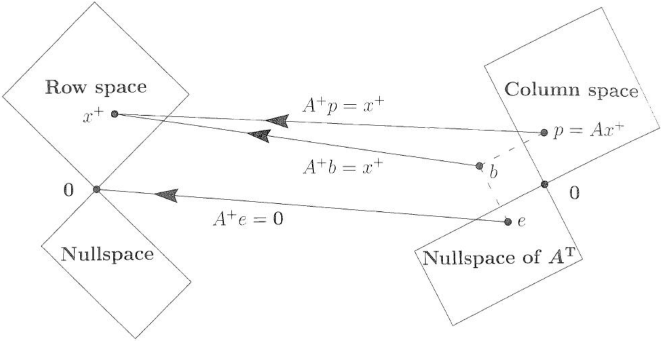


Figure 6.3: The pseudoinverse A^+ inverts A where it can on the column space.

Example

Example 6 $Ax = b$ is $-x_1 + 2x_2 + 2x_3 = 18$, with a whole plane of solutions.

According to our theory, the shortest solution should be in the row space of $A = [-1 \quad 2 \quad 2]$. The multiple of that row that satisfies the equation is $x^+ = (-2, 4, 4)$. The matrix that produces x^+ from $b = [18]$ is the pseudoinverse A^+. Whereas A was 1 by 3, this A^+ is 3 by 1:

$$A^+ = \begin{bmatrix} -1 & 2 & 2 \end{bmatrix}^+ = \begin{bmatrix} -\frac{1}{9} \\ \frac{2}{9} \\ \frac{2}{9} \end{bmatrix} \text{ and } A^+[18] = \begin{bmatrix} -2 \\ 4 \\ 4 \end{bmatrix}$$

The row space of A is the column space of A^+.

A Formula for A^+

Here is a formula for A^+:

Proposition

If $A = U\Sigma V^T$ (the SVD), then its pseudoinverse is $A^+ = V\Sigma^+ U^T$. The minimum length least-square solution is $x^+ = A^+ b = V\Sigma^+ U^T b$.

Proof.

Multiplication by the orthogonal matrix U^T leaves lengths unchanged:

$$||Ax - b|| = ||U\Sigma V^T x - b|| = |||\Sigma V^T x - U^T b|$$

Let $y = V^T x = V^{-1}x$, y has the same length as x. Then minimizing $||Ax - b||$ is the same as minimizing $||\Sigma y - U^T b||$. Now Σ is diagonal and we know the best y^+. It is $y^+ = \Sigma^+ U^T b$, so the best x^+ is $Vy^+ = A^+ b$. Vy^+ is in the row space, and $A^T A x^+ = A^T b$ from the SVD. □

Homework Assignment 29

6.3: 1, 3, 5, 10, 17, 18, 20.