6.3 Singular Value Decomposition

A great matrix factorization has been saved for the end of the basic course. $U\Sigma V^{\mathrm{T}}$ joins with LU from elimination and QR from orthogonalization (Gauss and Gram-Schmidt). Nobody's name is attached; $A = U\Sigma V^{\mathrm{T}}$ 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 $Q\Lambda Q^{\mathrm{T}}$ of a positive definite matrix. The eigenvalues are in the diagonal matrix Λ. The eigenvector matrix Q *is orthogonal* ($Q^{\mathrm{T}}Q = I$) because eigenvectors of a symmetric matrix can be chosen to be orthonormal. For most matrices that is not true, and for rectangular matrices it is ridiculous (eigenvalues undefined). But now we allow the Q on the left and the Q^{T} on the right to be *any two orthogonal matrices* U and V^{T}—not necessarily transposes of each other. Then every matrix will split into $A = U\Sigma V^{\mathrm{T}}$.

The diagonal (but rectangular) matrix Σ has eigenvalues from $A^{\mathrm{T}}A$, not from A! Those positive entries (also called sigma) will be $\sigma_1, \ldots, \sigma_r$. 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.

With rectangular matrices, the key is almost always to consider $A^{\mathrm{T}}A$ and AA^{T}.

Singular Value Decomposition: Any m by n matrix A can be factored into

$$A = U\Sigma V^{\mathrm{T}} = (\textbf{orthogonal})(\textbf{diagonal})(\textbf{orthogonal}).$$

The columns of U (m by m) are eigenvectors of AA^{T}, and the columns of V (n by n) are eigenvectors of $A^{\mathrm{T}}A$. 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 $A^{\mathrm{T}}A$.

***Remark* 1.** For positive definite matrices, Σ is Λ and $U\Sigma V^{\mathrm{T}}$ is identical to $Q\Lambda Q^{\mathrm{T}}$. 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* (the complex version of orthogonal). We take complex conjugates in $U^{\mathrm{H}}U = I$ and $V^{\mathrm{H}}V = I$ and $A = U\Sigma V^{\mathrm{H}}$.

***Remark* 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

***Remark* 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.

***Remark* 4.** Eigenvectors of AA^{T} and $A^{\mathrm{T}}A$ must go into the columns of U and V:

$$AA^{\mathrm{T}} = (U\Sigma V^{\mathrm{T}})(V\Sigma^{\mathrm{T}}U^{\mathrm{T}}) = U\Sigma\Sigma^{\mathrm{T}}U^{\mathrm{T}} \quad \text{and, similarly,} \quad A^{\mathrm{T}}A = V\Sigma^{\mathrm{T}}\Sigma V^{\mathrm{T}}. \tag{1}$$

U must be the eigenvector matrix for AA^{T}. The eigenvalue matrix in the middle is $\Sigma\Sigma^{\mathrm{T}}$—which is m by m with $\sigma_1^2, \ldots, \sigma_r^2$ on the diagonal.

From the $A^{\mathrm{T}}A = V\Sigma^{\mathrm{T}}\Sigma V^{\mathrm{T}}$, the *$V$ matrix must be the eigenvector matrix for $A^{\mathrm{T}}A$.* The diagonal matrix $\Sigma^{\mathrm{T}}\Sigma$ has the same $\sigma_1^2, \ldots, \sigma_r^2$, but it is n by n.

***Remark* 5.** Here is the reason that $Av_j = \sigma_j u_j$. Start with $A^{\mathrm{T}}Av_j = \sigma_j^2 v_j$:

$$\textbf{Multiply by } A \qquad AA^{\mathrm{T}}Av_j = \sigma_j^2 Av_j \tag{2}$$

This says that Av_j is an eigenvector of AA^{T}! We just moved parentheses to $(AA^{\mathrm{T}})(Av_j)$. The length of this eigenvector Av_j is σ_j, because

$$v^{\mathrm{T}}A^{\mathrm{T}}Av_j = \sigma_j^2 v_j^{\mathrm{T}}v_j \qquad \text{gives} \qquad \|Av_j\|^2 = \sigma_j^2.$$

So the unit eigenvector is $Av_j/\sigma_j = u_j$. **In other words,** $AV = U\Sigma$.

Example 1. This A has only one column: rank $r = 1$. Then Σ has only $\sigma_1 = 3$:

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

$A^{\mathrm{T}}A$ 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 2. Now A has rank 2, and $AA^{\mathrm{T}} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ with $\lambda = 3$ and 1:

$$\begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} = U\Sigma V^{\mathrm{T}} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -2 & 1 \\ -1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{matrix} /\sqrt{6} \\ /\sqrt{2} \\ /\sqrt{3} \end{matrix}.$$

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 $A^{\mathrm{T}}A$).

Application of the SVD

We will pick a few important applications, after emphasizing one key point. The SVD is terrific for numerically stable computations. because U and V are orthogonal matrices. They never change the length of a vector. Since $\|Ux\|^2 = x^{\mathrm{T}}U^{\mathrm{T}}Ux = \|x\|^2$, multiplication by U cannot destroy the scaling.

Of course Σ could multiply by a large σ or (more commonly) divide by a small σ, and overflow the computer. But still Σ *is as good as possible*. It reveals exactly what is large and what is small. The ratio $\sigma_{\max}/\sigma_{\min}$ is the **condition number** of an invertible n by n matrix. The availability of that information is another reason for the popularity of the SVD. We come back to this in the second application.

1. 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. The other 980 columns are multiplied in $U\Sigma V^{\mathrm{T}}$ by the small σ's that are being ignored. *We can do the matrix multiplication as columns times rows*:

$$A = U\Sigma V^{\mathrm{T}} = u_1\sigma_1 v_1^{\mathrm{T}} + u_2\sigma_2 v_2^{\mathrm{T}} + \cdots + u_r\sigma_r v_r^{\mathrm{T}}. \tag{3}$$

Any matrix is the sum of r matrices of rank 1. If only 20 terms are kept, we send 20 times 2000 numbers instead of a million (25 to 1 compression).

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 has become much more efficient, but it is expensive for a big matrix.

2. 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! In exact arithmetic, counting the pivots is correct. Real arithmetic can be misleading—but discarding small pivots is not the answer. Consider the following:

$$\textbf{ε is small} \qquad \begin{bmatrix} \varepsilon & 2\varepsilon \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \varepsilon & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \varepsilon & 1 \\ \varepsilon & 1+\varepsilon \end{bmatrix}.$$

The first has rank 1, although roundoff error will probably produce a second pivot. Both pivots will be small; how many do we ignore? The second has one small pivot, but we cannot pretend that its row is insignificant. The third has two pivots and its rank is 2, but its "effective rank" ought to be 1.

We go to a more stable measure of rank. The first step is to use $A^{\mathrm{T}}A$ or AA^{T}, which are symmetric but share the same rank as 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).

3. 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 *orthogonal 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*: $U^{\mathrm{H}}U = I$. We take the complex conjugate as well as the transpose, for U^{H}.

The SVD extends this "polar factorization" to matrices of any size:

> 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.

For proof we just insert $V^{\mathrm{T}}V = I$ into the middle of the SVD:

$$A = U\Sigma V^{\mathrm{T}} = (UV^{\mathrm{T}})(V\Sigma V^{\mathrm{T}}). \tag{4}$$

The factor $S = V\Sigma V^{\mathrm{T}}$ is symmetric and semidefinite (because Σ is). The factor $Q = UV^{\mathrm{T}}$ is an orthogonal matrix (because $Q^{\mathrm{T}}Q = VU^{\mathrm{T}}UV^{\mathrm{T}} = I$). In the complex case, S becomes Hermitian instead of symmetric and Q becomes unitary instead of orthogonal. In the invertible case Σ is definite and so is S.

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.

Application of $A = QS$: A major use of the polar decomposition is in continuum mechanics (and recently in robotics). In any deformation, it is important to separate stretching from rotation, and 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 $\sigma_1, \ldots, \sigma_r$, which are the stretching factors (or compression factors). The diagonalization that displays those eigenvalues is the natural choice of axes—called ***principal axes***: as in the ellipses of Section 6.2. It is S that requires work on the material, and stores up elastic energy.

We note that S^2 is $A^{\mathrm{T}}A$, which is symmetric positive definite when A is invertible. S is the symmetric positive definite square root of $A^{\mathrm{T}}A$, and Q is AS^{-1}. In fact, *A could be rectangular, as long as $A^{\mathrm{T}}A$ is positive definite*. (That is the condition we keep meeting,

that A must have independent columns.) In the reverse order $A = S'Q$, the matrix S' is the symmetric positive definite square root of AA^{T}.

4. Least Squares For a rectangular system $Ax = b$. the least-squares solution comes from the normal equations $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b$. ***If A has dependent columns then $A^{\mathrm{T}}A$ is not invertible and $\widehat{x}$ is not determined***. Any vector in the nullspace could be added to $\widehat{x}$. We can now complete Chapter 3, by choosing a "best" (*shortest*) $\widehat{x}$ for every $Ax = b$.

$Ax = b$ has two possible difficulties: *Dependent rows or dependent columns*. With dependent rows, $Ax = b$ may have no solution. That happens when b is outside the column space of A. Instead of $Ax = b$. we solve $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b$. But if A has *dependent columns*, this $\widehat{x}$ will not be unique. We have to choose a particular solution of $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b$, and we choose the shortest.

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

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 $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b$ (which had too many). We start with a diagonal example.

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

$$A\widehat{x} = p \quad \text{is} \quad \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \\ \widehat{x}_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ 0 \end{bmatrix}.$$

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 $Ax = b$ is to solve the first two equations, since the third equation is $0 = b_3$. That error cannot be reduced, but the errors in the first two equations will be zero. Then

$$\widehat{x}_1 = b_1/\sigma_1 \qquad \text{and} \qquad \widehat{x}_2 = b_2/\sigma_2.$$

Now we face the second difficulty. To make $\widehat{x}$ as short as possible, we choose the totally arbitrary $\widehat{x}_3$ and $\widehat{x}_4$ to be zero. ***The minimum length solution is*** x^+:

A^+ **is pseudoinverse**
$x^+ = A^+b$ **is shortest**

$$x^+ = \begin{bmatrix} b_1/\sigma_1 \\ b_2/\sigma_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/\sigma_1 & 0 & 0 \\ 0 & 1/\sigma_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}. \tag{5}$$

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. Based on this example, we know Σ^+

and x^+ for any diagonal matrix Σ:

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \\ & & \end{bmatrix} \qquad \Sigma^+ = \begin{bmatrix} 1/\sigma_1 & & \\ & \ddots & \\ & & 1/\sigma_r \\ & & \end{bmatrix} \qquad \Sigma^+ b = \begin{bmatrix} b_1/\sigma_1 \\ \vdots \\ b_r/\sigma_r \\ \end{bmatrix}.$$

The matrix Σ is m by n, with r nonzero entries σ_i. Its pseudoinverse Σ^+ is n by m, with r nonzero entries $1/\sigma_i$. **All the blank spaces are zeros**. Notice that $(\Sigma^+)^+$ is Σ again. That is like $(A^{-1})^{-1} = A$, but here A is not invertible.

Now we find x^+ in the general case. We claim that ***the shortest solution x^+ is always in the row space of A***. Remember that any vector $\widehat{x}$ can be split into a row space component x_r and a nullspace component: $\widehat{x} = x_r + x_n$. There are three important points about that splitting:

1. The row space component also solves $A^{\mathrm{T}}A\widehat{x}_r = A^{\mathrm{T}}b$, because $Ax_n = 0$.

2. The components are orthogonal, and they obey Pythagoras's law:

$$\|\widehat{x}\|^2 = \|x_r\|^2 + \|x_n\|^2, \quad \text{so } \widehat{x} \text{ is shortest when } x_n = 0.$$

3. All solutions of $A^{\mathrm{T}}A\widehat{x} = A^{\mathrm{T}}b$ have the same x_r. ***That vector is x^+***.

The fundamental theorem of linear algebra was in Figure 3.4. Every p in the column space comes from one and only one vector x_r in the row space. *All we are doing is to choose that vector, $x^+ = x_r$, as the best solution to $Ax = b$.*

The pseudoinverse in Figure 6.3 starts with b and comes back to x^+. *It inverts A where A is invertible*—between row space and column space. The pseudoinverse knocks out the left nullspace by sending it to zero, and it knocks out the nullspace by choosing x_r as x^+.

We have not yet shown that there is a matrix A^+ that always gives x^+—but there is. It will be n by m, because it takes b and p in $\mathbf{R}^m$ back to x^+ in $\mathbf{R}^n$. We look at one more example before finding A^+ in general.

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 $A = [-1 \;\; 2 \;\; 2]$. The multiple of that row that satisfies the equation is $x^+ = (-2,4,4)$. There are longer solutions like $(-2,5,3)$, $(-2,7,1)$, or $(-6,3,3)$, but they all have nonzero components from the nullspace. The matrix that produces x^+ from $b = [18]$ is the pseudoinverse A^+. Whereas A was 1 by 3, this A^+ is 3 by 1:

$$A^+ = \begin{bmatrix} -1 & 2 & 2 \end{bmatrix}^+ = \begin{bmatrix} -\frac{1}{9} \\ \frac{2}{9} \\ \frac{2}{9} \end{bmatrix} \qquad \text{and} \qquad A^+[18] = \begin{bmatrix} -2 \\ 4 \\ 4 \end{bmatrix}. \tag{6}$$

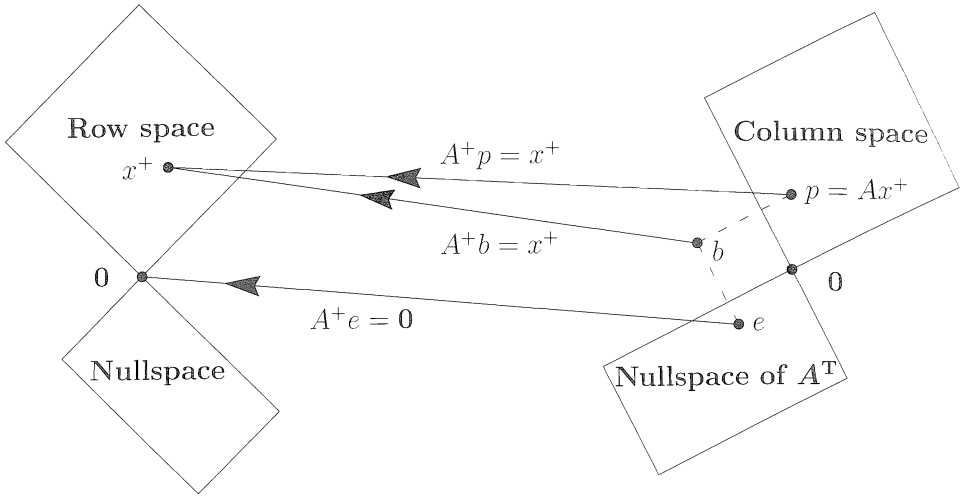

Figure 6.3: The pseudoinverse A^+ inverts A where it can on the column space.

The row space of A is the column space of A^+. Here is a formula for A^+:

$$\textbf{If } A = U\Sigma V^{\mathrm{T}} \textbf{ (the SVD), then its pseudoinverse is } A^+ = V\Sigma^+ U^{\mathrm{T}}. \tag{7}$$

Example 6 had $\sigma = 3$—the square root of the eigenvalue of $AA^{\mathrm{T}} = [9]$. Here it is again with Σ and Σ^+:

$$A = \begin{bmatrix} -1 & 2 & 2 \end{bmatrix} = U\Sigma V^{\mathrm{T}} = \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix}$$

$$V\Sigma^+ U^{\mathrm{T}} = \begin{bmatrix} -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{1}{3} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{9} \\ \frac{2}{9} \\ \frac{2}{9} \end{bmatrix} = A^+.$$

The minimum length least-squares solution is $x^+ = A^+b = V\Sigma^+U^{\mathrm{T}}b$.

Proof. Multiplication by the orthogonal matrix U^{T} leaves lengths unchanged:

$$\|Ax - b\| = \|U\Sigma V^{\mathrm{T}}x - b\| = \|\Sigma V^{\mathrm{T}}x - U^{\mathrm{T}}b\|.$$

Introduce the new unknown $y = V^{\mathrm{T}}x = V^{-1}x$, which has the same length as x. Then, minimizing $\|Ax-b\|$ is the same as minimizing $\|\Sigma y - U^{\mathrm{T}}b\|$. Now Σ is diagonal and we know the best y^+. It is $y^+ = \Sigma^+U^{\mathrm{T}}b$ so the best x^+ is Vy^+:

$$\textbf{Shortest solution} \qquad x^+ = Vy^+ = V\Sigma^+U^{\mathrm{T}}b = A^+b.$$

Vy^+ is in the row space, and $A^{\mathrm{T}}Ax^+ = A^{\mathrm{T}}b$ from the **SVD**. □