10 The Singular Value Decomposition

In section 9, we saw that a matrix transforms vectors in its domain into vectors in its range (column space), and vectors in its null space into the zero vector. No nonzero vector is mapped into the left null space, that is, into the orthogonal complement of the range. In this section, we make this statement more specific by showing how *unit* vectors[45] in the rowspace are transformed by matrices. This describes the action that a matrix has on the *magnitudes* of vectors as well. To this end, we first need to introduce the notion of orthogonal matrices, and interpret them geometrically as transformations between systems of orthonormal coordinates. We do this in section 10. Then, in section 10, we use these new concepts to introduce the all-important concept of the Singular Value Decomposition (SVD). The chapter concludes with some basic applications and examples.

Orthogonal Matrices

Let $\mathcal{S}$ be an n-dimensional subspace of $\mathbf{R}^m$ (so that we necessarily have $n \leq m$), and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be an orthonormal basis for $\mathcal{S}$. Consider a point P in $\mathcal{S}$. If the coordinates of P in $\mathbf{R}^m$ are collected in an m-dimensional vector

$$\mathbf{p} = \begin{bmatrix} p_1 \\ \vdots \\ p_m \end{bmatrix} ,$$

and since P is in $\mathcal{S}$, it must be possible to write $\mathbf{p}$ as a linear combination of the $\mathbf{v}_j$s. In other words, there must exist coefficients

$$\mathbf{q} = \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix}$$

such that

$$\mathbf{p} = q_1\mathbf{v}_1 + \ldots + q_n\mathbf{v}_n = V\mathbf{q}$$

where

$$V = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix}$$

is an $m \times n$ matrix that collects the basis for $\mathcal{S}$ as its columns. Then for any $i = 1, \ldots, n$ we have

$$\mathbf{v}_i^T\mathbf{p} = \mathbf{v}_i^T \sum_{j=1}^{n} q_j\mathbf{v}_j = \sum_{j=1}^{n} q_j\mathbf{v}_i^T\mathbf{v}_j = q_i ,$$

since the $\mathbf{v}_j$ are orthonormal. This is important, and may need emphasis:

If

$$\mathbf{p} = \sum_{j=1}^{n} q_j\mathbf{v}_j$$

[45] Vectors with unit norm.

and the vectors of the basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are orthonormal, then the coefficients q_j are the signed magnitudes of the projections of **p** *onto the basis vectors:*

$$q_j = \mathbf{v}_j^T \mathbf{p} \; . \tag{66}$$

In matrix form,

$$\mathbf{q} = V^T \mathbf{p} \; . \tag{67}$$

Also, we can collect the n^2 equations

$$\mathbf{v}_i^T \mathbf{v}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

into the following matrix equation:

$$V^T V = I \tag{68}$$

where I is the $n \times n$ identity matrix. A matrix V that satisfies equation (68) is said to be *orthogonal*. Thus, a matrix is orthogonal if its columns are orthonormal. Since the *left inverse* of a matrix V is defined as the matrix L such that

$$LV = I \; , \tag{69}$$

comparison with equation (68) shows that the left inverse of an orthogonal matrix V exists, and is equal to the transpose of V.

Of course, this argument requires V to be full rank, so that the solution L to equation (69) is unique. However, V is certainly full rank, because it is made of orthonormal columns.

Notice that $VR = I$ cannot possibly have a solution when $m > n$, because the $m \times m$ identity matrix has m linearly independent [46] columns, while the columns of VR are linear combinations of the n columns of V, so VR can have at most n linearly independent columns.

Of course, this result is still valid when V is $m \times m$ and has orthonormal columns, since equation (68) still holds. However, for square, full-rank matrices ($r = m = n$), the distinction between left and right inverse vanishes. In fact, suppose that there exist matrices L and R such that $LV = I$ and $VR = I$. Then $L = L(VR) = (LV)R = R$, so the left and the right inverse are the same. Thus, for square orthogonal matrices, V^T is both the left and the right inverse:

$$V^T V = V V^T = I \; ,$$

and V^T is then simply said to be the *inverse* of V:

$$V^T = V^{-1} \; .$$

Since the matrix VV^T contains the inner products between the *rows* of V (just as $V^T V$ is formed by the inner products of its *columns*), the argument above shows that the rows of a *square* orthogonal matrix are orthonormal as well. We can summarize this discussion as follows:

[46] Nay, orthonormal.

Theorem 10.1 *The left inverse of an orthogonal $m \times n$ matrix V with $m \geq n$ exists and is equal to the transpose of V:*

$$V^T V = I \, .$$

In particular, if $m = n$, the matrix $V^{-1} = V^T$ is also the right inverse of V:

$$V \textit{ square} \quad \Rightarrow \quad V^{-1}V = V^T V = VV^{-1} = VV^T = I \, .$$

Sometimes, when $m = n$, the geometric interpretation of equation (67) causes confusion, because two interpretations of it are possible. In the interpretation given above, the point P remains the same, and the underlying reference frame is changed from the elementary vectors $\mathbf{e}_j$ (that is, from the columns of I) to the vectors $\mathbf{v}_j$ (that is, to the columns of V). Alternatively, equation (67) can be seen as a transformation, in a fixed reference system, of point P with coordinates $\mathbf{p}$ into a different point Q with coordinates $\mathbf{q}$. This, however, is relativity, and should not be surprising: If you spin clockwise on your feet, or if you stand still and the whole universe spins counterclockwise around you, the result is the same.[47]

Consistently with either of these geometric interpretations, we have the following result:

Theorem 10.2 *The norm of a vector $\mathbf{x}$ is not changed by multiplication by an orthogonal matrix V:*

$$\|V\mathbf{x}\| = \|\mathbf{x}\| \, .$$

Proof.

$$\|V\mathbf{x}\|^2 = \mathbf{x}^T V^T V \mathbf{x} = \mathbf{x}^T \mathbf{x} = \|\mathbf{x}\|^2 \, .$$

Δ

We conclude this section with an obvious but useful consequence of orthogonality. In section 9 we defined the projection $\mathbf{p}$ of a vector $\mathbf{b}$ onto another vector $\mathbf{c}$ as the point on the line through $\mathbf{c}$ that is closest to $\mathbf{b}$. This notion of projection can be extended from lines to vector spaces by the following definition: The *projection* $\mathbf{p}$ of a point $\mathbf{b} \in \mathbf{R}^n$ *onto a subspace* C is the point in C that is closest to $\mathbf{b}$.

Also, for *unit* vectors $\mathbf{c}$, the projection matrix is $\mathbf{c}\mathbf{c}^T$ (theorem 9.7), and the vector $\mathbf{b} - \mathbf{p}$ is orthogonal to $\mathbf{c}$. An analogous result holds for subspace projection, as the following theorem shows.

Theorem 10.3 *Let U be an orthogonal matrix. Then the matrix UU^T projects any vector $\mathbf{b}$ onto* range(U)*. Furthermore, the difference vector between $\mathbf{b}$ and its projection $\mathbf{p}$ onto* range(U) *is orthogonal to* range(U)*:*

$$U^T(\mathbf{b} - \mathbf{p}) = \mathbf{0} \, .$$

[47] At least geometrically. One solution may be more efficient than the other in other ways.

Proof. A point **p** in range(U) is a linear combination of the columns of U:

$$\mathbf{p} = U\mathbf{x}$$

where **x** is the vector of coefficients (as many coefficients as there are columns in U). The squared distance between **b** and **p** is

$$\|\mathbf{b} - \mathbf{p}\|^2 = (\mathbf{b} - \mathbf{p})^T(\mathbf{b} - \mathbf{p}) = \mathbf{b}^T\mathbf{b} + \mathbf{p}^T\mathbf{p} - 2\mathbf{b}^T\mathbf{p} = \mathbf{b}^T\mathbf{b} + \mathbf{x}^T U^T U\mathbf{x} - 2\mathbf{b}^T U\mathbf{x}\ .$$

Because of orthogonality, U^TU is the identity matrix, so

$$\|\mathbf{b} - \mathbf{p}\|^2 = \mathbf{b}^T\mathbf{b} + \mathbf{x}^T\mathbf{x} - 2\mathbf{b}^T U\mathbf{x}\ .$$

The derivative of this squared distance with respect to **x** is the vector

$$2\mathbf{x} - 2U^T\mathbf{b}$$

which is zero iff

$$\mathbf{x} = U^T\mathbf{b}\ ,$$

that is, when

$$\mathbf{p} = U\mathbf{x} = UU^T\mathbf{b}$$

as promised.

For this value of **p** the difference vector $\mathbf{b} - \mathbf{p}$ is orthogonal to range(U), in the sense that

$$U^T(\mathbf{b} - \mathbf{p}) = U^T(\mathbf{b} - UU^T\mathbf{b}) = U^T\mathbf{b} - U^T\mathbf{b} = \mathbf{0}\ .$$

Δ

The Singular Value Decomposition

The following statement draws a geometric picture underlying the concept of Singular Value Decomposition using the concepts developed in the previous Section:

> An $m \times n$ matrix A of rank r maps the r-dimensional unit hypersphere in rowspace(A) into an r-dimensional hyperellipse in range(A).

This statement is stronger than saying that A maps rowspace(A) into range(A), because it also describes what happens to the *magnitudes* of the vectors: a hypersphere is stretched or compressed into a hyperellipse, which is a quadratic hypersurface that generalizes the two-dimensional notion

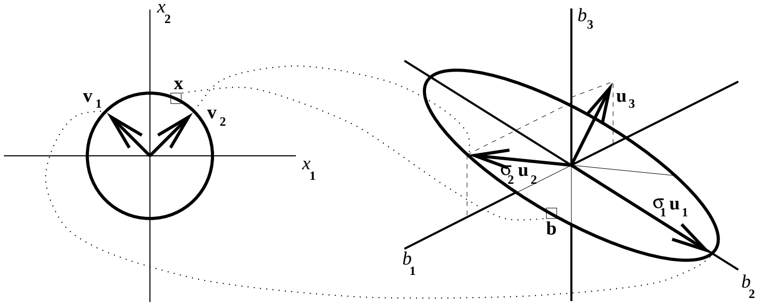

Figure 32: The matrix in equation (70) maps a circle on the plane into an ellipse in space. The two small boxes are corresponding points.

of ellipse to an arbitrary number of dimensions. In three dimensions, the hyperellipse is an ellipsoid, in one dimension it is a pair of points. In all cases, the hyperellipse in question is centered at the origin.

For instance, the rank-2 matrix

$$A = \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{3} & \sqrt{3} \\ -3 & 3 \\ 1 & 1 \end{bmatrix} \tag{70}$$

transforms the unit circle on the plane into an ellipse embedded in three-dimensional space. Figure 32 shows the map

$$\mathbf{b} = A\mathbf{x} \; .$$

Two diametrically opposite points on the unit circle are mapped into the two endpoints of the major axis of the ellipse, and two other diametrically opposite points on the unit circle are mapped into the two endpoints of the minor axis of the ellipse. The lines through these two pairs of points on the unit circle are always orthogonal. This result can be generalized to any $m \times n$ matrix.

Simple and fundamental as this geometric fact may be, its proof by geometric means is cumbersome. Instead, we will prove it algebraically by first introducing the existence of the SVD and then using the latter to prove that matrices map hyperspheres into hyperellipses.

Theorem 10.4 *If A is a real $m \times n$ matrix then there exist orthogonal matrices*

$$\begin{aligned} U &= \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_m \end{bmatrix} \in \mathcal{R}^{m\times m} \\ V &= \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix} \in \mathcal{R}^{n\times n} \end{aligned}$$

such that

$$U^T A V = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_p) \in \mathcal{R}^{m\times n}$$

where $p = \min(m, n)$ and $\sigma_1 \geq \ldots \geq \sigma_p \geq 0$. Equivalently,

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

Proof. Let $\mathbf{x}$ and $\mathbf{y}$ be unit vectors in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and consider the bilinear form

$$z = \mathbf{y}^T A\mathbf{x} \ .$$

The set

$$\mathcal{S} \ = \ \{\mathbf{x}, \mathbf{y} \mid \mathbf{x} \in \mathbf{R}^n, \ \mathbf{y} \in \mathbf{R}^m, \ \|\mathbf{x}\| = \|\mathbf{y}\| = 1\}$$

is compact, so that the scalar function $z(\mathbf{x}, \mathbf{y})$ must achieve a maximum value on $\mathcal{S}$, possibly at more than one point [48]. Let $\mathbf{u}_1$, $\mathbf{v}_1$ be two unit vectors in $\mathbf{R}^m$ and $\mathbf{R}^n$ respectively where this maximum is achieved, and let σ_1 be the corresponding value of z:

$$\max_{\|\mathbf{x}\|=\|\mathbf{y}\|=1} \mathbf{y}^T A\mathbf{x} = \mathbf{u}_1^T A\mathbf{v}_1 = \sigma_1 \ .$$

It is easy to see that $\mathbf{u}_1$ is parallel to the vector $A\mathbf{v}_1$. If this were not the case, their inner product $\mathbf{u}_1^T A\mathbf{v}_1$ could be increased by rotating $\mathbf{u}_1$ towards the direction of $A\mathbf{v}_1$, thereby contradicting the fact that $\mathbf{u}_1^T A\mathbf{v}_1$ is a maximum. Similarly, by noticing that

$$\mathbf{u}_1^T A\mathbf{v}_1 = \mathbf{v}_1^T A^T \mathbf{u}_1$$

and repeating the argument above, we see that $\mathbf{v}_1$ is parallel to $A^T\mathbf{u}_1$.

By theorems 9.8 and 9.9, $\mathbf{u}_1$ and $\mathbf{v}_1$ can be extended into orthonormal bases for $\mathbf{R}^m$ and $\mathbf{R}^n$, respectively. Collect these orthonormal basis vectors into orthogonal matrices U_1 and V_1. Then

$$U_1^T A V_1 = S_1 = \begin{bmatrix} \sigma_1 & \mathbf{0}^T \\ \mathbf{0} & A_1 \end{bmatrix} \ .$$

In fact, the first column of AV_1 is $A\mathbf{v}_1 = \sigma_1\mathbf{u}_1$, so the first entry of $U_1^T A V_1$ is $\mathbf{u}_1^T \sigma_1 \mathbf{u}_1 = \sigma_1$, and its other entries are $\mathbf{u}_j^T A\mathbf{v}_1 = 0$ because $A\mathbf{v}_1$ is parallel to $\mathbf{u}_1$ and therefore orthogonal, by construction, to $\mathbf{u}_2, \ldots, \mathbf{u}_m$. A similar argument shows that the entries after the first in the first row of S_1 are zero: the row vector $\mathbf{u}_1^T A$ is parallel to $\mathbf{v}_1^T$, and therefore orthogonal to $\mathbf{v}_2, \ldots, \mathbf{v}_n$, so that $\mathbf{u}_1^T A\mathbf{v}_2 = \ldots = \mathbf{u}_1^T A\mathbf{v}_n = 0$.

The matrix A_1 has one fewer row and column than A. We can repeat the same construction on A_1 and write

$$U_2^T A_1 V_2 = S_2 = \begin{bmatrix} \sigma_2 & \mathbf{0}^T \\ \mathbf{0} & A_2 \end{bmatrix}$$

so that

$$\begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & U_2^T \end{bmatrix} U_1^T A V_1 \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & V_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & \mathbf{0}^T \\ 0 & \sigma_2 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & A_2 \end{bmatrix} \ .$$

This procedure can be repeated until A_k vanishes (zero rows or zero columns) to obtain

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

[48] Actually, at least at two points: if $\mathbf{u}_1^T A\mathbf{v}_1$ is a maximum, so is $(-\mathbf{u}_1)^T A(-\mathbf{v}_1)$.

where U^T and V are orthogonal matrices obtained by multiplying together all the orthogonal matrices used in the procedure, and

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \ .$$

Since matrices U and V are orthogonal, we can premultiply the matrix product in the theorem by U and postmultiply it by V^T to obtain

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

which is the desired result.

It only remains to show that the elements on the diagonal of Σ are nonnegative and arranged in nonincreasing order. To see that $\sigma_1 \geq \ldots \geq \sigma_p$ (where $p = \min(m, n)$), we can observe that the successive maximization problems that yield $\sigma_1, \ldots, \sigma_p$ are performed on a sequence of sets each of which contains the next. To show this, we just need to show that $\sigma_2 \leq \sigma_1$, and induction will do the rest. We have

$$\begin{aligned} \sigma_2 &= \max_{\|\hat{\mathbf{x}}\|=\|\hat{\mathbf{y}}\|=1} \hat{\mathbf{y}}^T A_1 \hat{\mathbf{x}} = \max_{\|\hat{\mathbf{x}}\|=\|\hat{\mathbf{y}}\|=1} \left[\begin{array}{cc} 0 & \hat{\mathbf{y}} \end{array} \right]^T S_1 \left[\begin{array}{c} 0 \\ \hat{\mathbf{x}} \end{array} \right] \\ &= \max_{\|\hat{\mathbf{x}}\|=\|\hat{\mathbf{y}}\|=1} \left[\begin{array}{cc} 0 & \hat{\mathbf{y}} \end{array} \right]^T U_1^T A V_1 \left[\begin{array}{c} 0 \\ \hat{\mathbf{x}} \end{array} \right] = \max_{\substack{\|\mathbf{x}\| = \|\mathbf{y}\| = 1 \\ \mathbf{x}^T \mathbf{v}_1 = \mathbf{y}^T \mathbf{u}_1 = 0}} \mathbf{y}^T A \mathbf{x} \leq \sigma_1 \ . \end{aligned}$$

To explain the last equality above, consider the vectors

$$\mathbf{x} = V_1 \left[\begin{array}{c} 0 \\ \hat{\mathbf{x}} \end{array} \right] \quad \text{and} \quad \mathbf{y} = U_1 \left[\begin{array}{c} 0 \\ \hat{\mathbf{y}} \end{array} \right] \ .$$

The vector $\mathbf{x}$ is equal to the unit vector $[0 \ \hat{\mathbf{x}}]^T$ transformed by the orthogonal matrix V_1, and is therefore itself a unit vector. In addition, it is a linear combination of $\mathbf{v}_2, \ldots, \mathbf{v}_n$, and is therefore orthogonal to $\mathbf{v}_1$. A similar argument shows that $\mathbf{y}$ is a unit vector orthogonal to $\mathbf{u}_1$. Because $\mathbf{x}$ and $\mathbf{y}$ thus defined belong to subsets (actually sub-spheres) of the unit spheres in $\mathbf{R}^n$ and $\mathbf{R}^m$, we conclude that $\sigma_2 \leq \sigma_1$.

The σ_i are nonnegative because all these maximizations are performed on unit hyper-spheres. The σ_is are maxima of the function $z(\mathbf{x}, \mathbf{y})$ which always assumes both positive and negative values on any hyper-sphere: If $z(\mathbf{x}, \mathbf{y})$ is negative, then $z(-\mathbf{x}, \mathbf{y})$ is positive, and if $\mathbf{x}$ is on a hyper-sphere, so is $-\mathbf{x}$. Δ

We can now review the geometric picture in figure 32 in light of the singular value decomposition. In the process, we introduce some nomenclature for the three matrices in the SVD. Consider the map in figure 32, represented by equation (70), and imagine transforming point $\mathbf{x}$ (the small box at $\mathbf{x}$ on the unit circle) into its corresponding point $\mathbf{b} = A\mathbf{x}$ (the small box on the ellipse). This transformation can be achieved in three steps (see figure 33):

1. Write **x** in the frame of reference of the two vectors $\mathbf{v}_1, \mathbf{v}_2$ on the unit circle that map into the major axes of the ellipse. There are a few ways to do this, because axis endpoints come in pairs. Just pick one way, but order $\mathbf{v}_1, \mathbf{v}_2$ so they map into the major and the minor axis, in this order. Let us call $\mathbf{v}_1, \mathbf{v}_2$ the two *right singular vectors* of A. The corresponding axis unit vectors $\mathbf{u}_1, \mathbf{u}_2$ on the ellipse are called *left singular vectors*. If we define

   $$V = \left[\begin{array}{cc} \mathbf{v}_1 & \mathbf{v}_2 \end{array} \right] ,$$

   the new coordinates ξ of **x** become

   $$\xi = V^T \mathbf{x}$$

   because V is orthogonal.

2. Transform ξ into its image on a "straight" version of the final ellipse. "Straight" here means that the axes of the ellipse are aligned with the y_1, y_2 axes. Otherwise, the "straight" ellipse has the same shape as the ellipse in figure 32. If the lengths of the half-axes of the ellipse are σ_1, σ_2 (major axis first), the transformed vector η has coordinates

   $$\eta = \Sigma \xi$$

   where

   $$\Sigma = \left[\begin{array}{cc} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{array} \right]$$

   is a diagonal matrix. The real, nonnegative numbers σ_1, σ_2 are called the *singular values* of A.

3. Rotate the reference frame in $\mathbf{R}^m = \mathbf{R}^3$ so that the "straight" ellipse becomes the ellipse in figure 32. This rotation brings η along, and maps it to **b**. The components of η are the signed magnitudes of the projections of **b** along the unit vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ that identify the axes of the ellipse and the normal to the plane of the ellipse, so

   $$\mathbf{b} = U \eta$$

   where the orthogonal matrix

   $$U = \left[\begin{array}{ccc} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{array} \right]$$

   collects the left singular vectors of A.

We can concatenate these three transformations to obtain

$$\mathbf{b} = U \Sigma V^T \mathbf{x}$$

or

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

since this construction works for any point **x** on the unit circle. This is the SVD of A.

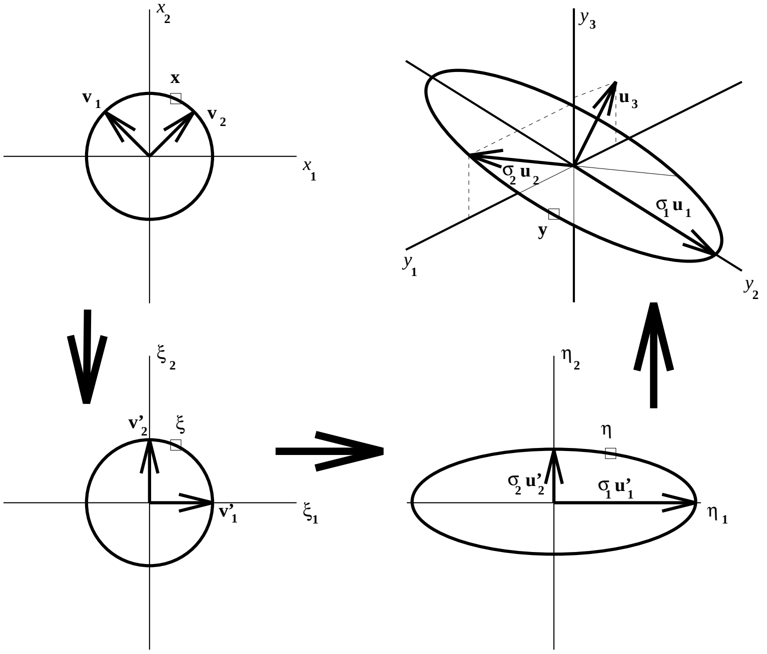

Figure 33: Decomposition of the mapping in figure 32.

The singular value decomposition is "almost unique". There are two sources of ambiguity. The first is in the orientation of the singular vectors. One can flip any right singular vector, provided that the corresponding left singular vector is flipped as well, and still obtain a valid SVD. Singular vectors must be flipped in pairs (a left vector and its corresponding right vector) because the singular values are required to be nonnegative. This is a trivial ambiguity. If desired, it can be removed by imposing, for instance, that the first nonzero entry of every left singular value be positive.

The second source of ambiguity is deeper. If the matrix A maps a hypersphere into another hypersphere, the axes of the latter are not defined. For instance, the identity matrix has an infinity of SVDs, all of the form

$$I = UIU^T$$

where U is any orthogonal matrix of suitable size. More generally, whenever two or more singular values coincide, the subspaces identified by the corresponding left and right singular vectors are unique, but any orthonormal basis can be chosen within, say, the right subspace and yield, together with the corresponding left singular vectors, a valid SVD. Except for these ambiguities, the SVD is unique.

Even in the general case, the singular values of a matrix A are the lengths of the semi-axes of the hyperellipse E defined by

$$E = \{A\mathbf{x} : \|\mathbf{x}\| = 1\} .$$

The SVD reveals a great deal about the structure of a matrix. If we define r by

$$\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = 0 ,$$

that is, if σ_r is the smallest nonzero singular value of A, then

$$\begin{aligned} \text{rank}(A) &= r \\ \text{null}(A) &= \text{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\} \\ \text{range}(A) &= \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\} \ . \end{aligned}$$

The sizes of the matrices in the SVD are as follows: U is $m \times m$, Σ is $m \times n$, and V is $n \times n$. Thus, Σ has the same shape and size as A, while U and V are square. However, if $m > n$, the bottom $(m-n) \times n$ block of Σ is zero, so that the last $m-n$ columns of U are multiplied by zero. Similarly, if $m < n$, the rightmost $m \times (n-m)$ block of Σ is zero, and this multiplies the last $n-m$ rows of V. This suggests a "small," equivalent version of the SVD. If $p = \min(m, n)$, we can define $U_p = U(:, 1:p)$, $\Sigma_p = \Sigma(1:p, 1:p)$, and $V_p = V(:, 1:p)$, and write

$$A = U_p \Sigma_p V_p^T$$

where U_p is $m \times p$, Σ_p is $p \times p$, and V_p is $n \times p$.

Moreover, if $p - r$ singular values are zero, we can let $U_r = U(:, 1:r)$, $\Sigma_r = \Sigma(1:r, 1:r)$, and $V_r = V(:, 1:r)$, then we have

$$A = U_r \Sigma_r V_r^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \ ,$$

which is an even smaller, *minimal*, SVD.

Finally, both the 2-norm and the Frobenius norm

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$$

and

$$\|A\|_2 = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$$

are neatly characterized in terms of the SVD:

$$\begin{aligned} \|A\|_F^2 &= \sigma_1^2 + \ldots + \sigma_p^2 \\ \|A\|_2 &= \sigma_1 \ . \end{aligned}$$

In the next few sections we introduce fundamental results and applications that testify to the importance of the SVD.

The Pseudoinverse

One of the most important applications of the SVD is the solution of linear systems in the least squares sense. A linear system of the form

$$A\mathbf{x} = \mathbf{b} \tag{71}$$

arising from a real-life application may or may not admit a solution, that is, a vector **x** that satisfies this equation exactly. Often more measurements are available than strictly necessary, because measurements are unreliable. This leads to more equations than unknowns (the number m of rows in A is greater than the number n of columns), and equations are often mutually incompatible because they come from inexact measurements (incompatible linear systems were defined in chapter 9). Even when $m \leq n$ the equations can be incompatible, because of errors in the measurements that produce the entries of A. In these cases, it makes more sense to find a vector **x** that minimizes the norm

$$\|A\mathbf{x} - \mathbf{b}\|$$

of the *residual* vector

$$\mathbf{r} = A\mathbf{x} - \mathbf{b} \; .$$

where the double bars henceforth refer to the Euclidean norm. Thus, **x** cannot exactly satisfy any of the m equations in the system, but it tries to satisfy all of them as closely as possible, as measured by the sum of the squares of the discrepancies between left- and right-hand sides of the equations.

In other circumstances, not enough measurements are available. Then, the linear system (71) is underdetermined, in the sense that it has fewer independent equations than unknowns (its rank r is less than n, see again chapter 9).

Incompatibility and underdeterminacy can occur together: the system admits no solution, and the least-squares solution is not unique. For instance, the system

$$\begin{aligned} x_1 + x_2 &= 1 \\ x_1 + x_2 &= 3 \\ x_3 &= 2 \end{aligned}$$

has three unknowns, but rank 2, and its first two equations are incompatible: $x_1 + x_2$ cannot be equal to both 1 and 3. A least-squares solution turns out to be $\mathbf{x} = [1 \;\; 1 \;\; 2]^T$ with residual $\mathbf{r} = A\mathbf{x} - \mathbf{b} = [1 \;\; -1 \;\; 0]$, which has norm $\sqrt{2}$ (admittedly, this is a rather high residual, but this is the best we can do for this problem, in the least-squares sense). However, any other vector of the form

$$\mathbf{x}' = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} + \alpha \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$$

is as good as **x**. For instance, $\mathbf{x}' = [0 \;\; 2 \;\; 2]$, obtained for $\alpha = 1$, yields exactly the same residual as **x** (check this).

In summary, an exact solution to the system (71) may not exist, or may not be unique, as we learned in chapter 9. An approximate solution, in the least-squares sense, always exists, but may fail to be unique.

If there are several least-squares solutions, all equally good (or bad), then one of them turns out to be shorter than all the others, that is, its norm $\|\mathbf{x}\|$ is smallest. One can therefore redefine what it means to "solve" a linear system so that there is always exactly one solution. This minimum norm solution is the subject of the following theorem, which both proves uniqueness and provides a recipe for the computation of the solution.

Theorem 10.5 *The minimum-norm least squares solution to a linear system $A\mathbf{x} = \mathbf{b}$, that is, the shortest vector* $\mathbf{x}$ *that achieves the*

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\| \ ,$$

is unique, and is given by

$$\hat{\mathbf{x}} = V\Sigma^{\dagger}U^T\mathbf{b} \tag{72}$$

where

$$\Sigma^{\dagger} = \begin{bmatrix} 1/\sigma_1 & & & & & & 0 & \cdots & 0 \\ & \ddots & & & & & & & \\ & & 1/\sigma_r & & & & \vdots & & \vdots \\ & & & 0 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 0 & 0 & \cdots & 0 \end{bmatrix}$$

is an $n \times m$ diagonal matrix.

The matrix

$$A^{\dagger} = V\Sigma^{\dagger}U^T$$

is called the *pseudoinverse* of A.

Proof. The minimum-norm Least Squares solution to

$$A\mathbf{x} = \mathbf{b}$$

is the shortest vector $\mathbf{x}$ that minimizes

$$\|A\mathbf{x} - \mathbf{b}\|$$

that is,

$$\|U\Sigma V^T\mathbf{x} - \mathbf{b}\| \ .$$

This can be written as

$$\|U(\Sigma V^T\mathbf{x} - U^T\mathbf{b})\| \tag{73}$$

because U is an orthogonal matrix, $UU^T = I$. But orthogonal matrices do not change the norm of vectors they are applied to (theorem 10.2), so that the last expression above equals

$$\|\Sigma V^T\mathbf{x} - U^T\mathbf{b}\|$$

or, with $\mathbf{y} = V^T\mathbf{x}$ and $\mathbf{c} = U^T\mathbf{b}$,

$$\|\Sigma\mathbf{y} - \mathbf{c}\| \ .$$

In order to find the solution to this minimization problem, let us spell out the last expression. We want to minimize the norm of the following vector:

$$\begin{bmatrix} \sigma_1 & 0 & & \cdots & 0 \\ 0 & \ddots & & \cdots & 0 \\ & \sigma_r & & & \\ \vdots & & 0 & & \vdots \\ & & & \ddots & \\ 0 & & & & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_r \\ y_{r+1} \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} c_1 \\ \vdots \\ c_r \\ c_{r+1} \\ \vdots \\ c_m \end{bmatrix} .$$

The last $m - r$ differences are of the form

$$\mathbf{0} - \begin{bmatrix} c_{r+1} \\ \vdots \\ c_m \end{bmatrix}$$

and do not depend on the unknown **y**. In other words, there is nothing we can do about those differences: if some or all the c_i for $i = r + 1, \ldots, m$ are nonzero, we will not be able to zero these differences, and each of them contributes a *residual* $|c_i|$ to the solution. In each of the first r differences, on the other hand, the last $n - r$ components of **y** are multiplied by zeros, so they have no effect on the solution. Thus, there is freedom in their choice. Since we look for the minimum-norm solution, that is, for the shortest vector **x**, we also want the shortest **y**, because **x** and **y** are related by an orthogonal transformation. We therefore set $y_{r+1} = \ldots = y_n = 0$. In summary, the desired **y** has the following components:

$$\begin{aligned} y_i &= \frac{c_i}{\sigma_i} \quad \text{for } i = 1, \ldots, r \\ y_i &= 0 \quad \text{for } i = r + 1, \ldots, n . \end{aligned}$$

When written as a function of the vector **c**, this is

$$\mathbf{y} = \Sigma^+ \mathbf{c} .$$

Notice that there is no other choice for **y**, which is therefore unique: minimum residual forces the choice of $y_1, \ldots, y_r$, and minimum-norm solution forces the other entries of **y**. Thus, the minimum-norm, least-squares solution to the original system is the unique vector

$$\hat{\mathbf{x}} = V\mathbf{y} = V\Sigma^+\mathbf{c} = V\Sigma^+U^T\mathbf{b}$$

as promised. The residual, that is, the norm of $\|A\mathbf{x} - \mathbf{b}\|$ when **x** is the solution vector, is the norm of $\Sigma\mathbf{y} - \mathbf{c}$, since this vector is related to $A\mathbf{x} - \mathbf{b}$ by an orthogonal transformation (see equation (73)). In conclusion, the square of the residual is

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \|\Sigma\mathbf{y} - \mathbf{c}\|^2 = \sum_{i=r+1}^{m} c_i^2 = \sum_{i=r+1}^{m} (\mathbf{u}_i^T\mathbf{b})^2$$

which is the projection of the right-hand side vector **b** onto the complement of the range of A. Δ

Least-Squares Solution of a Homogeneous Linear Systems

Theorem 10.5 works regardless of the value of the right-hand side vector **b**. When $\mathbf{b} = \mathbf{0}$, that is, when the system is *homogeneous*, the solution is trivial: the minimum-norm solution to

$$A\mathbf{x} = \mathbf{0} \tag{74}$$

is

$$\mathbf{x} = 0 \ ,$$

which happens to be an exact solution. Of course it is not necessarily the only one (any vector in the null space of A is also a solution, by definition), but it is obviously the one with the smallest norm.

Thus, $\mathbf{x} = 0$ is the minimum-norm solution to any homogeneous linear system. Although correct, this solution is not too interesting. In many applications, what is desired is a *nonzero* vector **x** that satisfies the system (74) as well as possible. Without any constraints on **x**, we would fall back to $\mathbf{x} = 0$ again. For homogeneous linear systems, the meaning of a least-squares solution is therefore usually modified, once more, by imposing the constraint

$$\|\mathbf{x}\| = 1$$

on the solution. Unfortunately, the resulting constrained minimization problem does not necessarily admit a *unique* solution. The following theorem provides a recipe for finding this solution, and shows that there is in general a whole hypersphere of solutions.

Theorem 10.6 *Let*

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

be the singular value decomposition of A. Furthermore, let $\mathbf{v}_{n-k+1}, \ldots, \mathbf{v}_n$ be the k columns of V whose corresponding singular values are equal to the last singular value σ_n, that is, let k be the largest integer such that

$$\sigma_{n-k+1} = \ldots = \sigma_n \ .$$

Then, all vectors of the form

$$\mathbf{x} = \alpha_1 \mathbf{v}_{n-k+1} + \ldots + \alpha_k \mathbf{v}_n \tag{75}$$

with

$$\alpha_1^2 + \ldots + \alpha_k^2 = 1 \tag{76}$$

are unit-norm least squares solutions to the homogeneous linear system

$$A\mathbf{x} = \mathbf{0},$$

that is, they achieve the

$$\min_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| \ .$$

Note: when σ_n is greater than zero the most common case is $k = 1$, since it is very unlikely that different singular values have *exactly* the same numerical value. When A is rank deficient, on the other case, it may often have more than one singular value equal to zero. In any event, if $k = 1$, then the minimum-norm solution is unique, $\mathbf{x} = \mathbf{v}_n$. If $k > 1$, the theorem above shows how to express *all* solutions as a linear combination of the last k columns of V.

Proof. The reasoning is very similar to that for the previous theorem. The unit-norm Least Squares solution to

$$A\mathbf{x} = \mathbf{0}$$

is the vector $\mathbf{x}$ with $\|\mathbf{x}\| = 1$ that minimizes

$$\|A\mathbf{x}\|$$

that is,

$$\|U\Sigma V^T\mathbf{x}\| \ .$$

Since orthogonal matrices do not change the norm of vectors they are applied to (theorem 10.2), this norm is the same as

$$\|\Sigma V^T\mathbf{x}\|$$

or, with $\mathbf{y} = V^T\mathbf{x}$,

$$\|\Sigma\mathbf{y}\| \ .$$

Since V is orthogonal, $\|\mathbf{x}\| = 1$ translates to $\|\mathbf{y}\| = 1$. We thus look for the unit-norm vector $\mathbf{y}$ that minimizes the norm (squared) of $\Sigma\mathbf{y}$, that is,

$$\sigma_1^2 y_1^2 + \ldots + \sigma_n^2 y_n^2 \ .$$

This is obviously achieved by concentrating all the (unit) mass of $\mathbf{y}$ where the σs are smallest, that is by letting

$$y_1 = \ldots = y_{n-k} = 0. \qquad (77)$$

From $\mathbf{y} = V^T\mathbf{x}$ we obtain $\mathbf{x} = V\mathbf{y} = y_1\mathbf{v}_1 + \ldots + y_n\mathbf{v}_n$, so that equation (77) is equivalent to equation (75) with $\alpha_1 = y_{n-k+1}, \ldots, \alpha_k = y_n$, and the unit-norm constraint on $\mathbf{y}$ yields equation (76). Δ

Section 10 shows a sample use of theorem 10.6.

SVD Line Fitting

The Singular Value Decomposition of a matrix yields a simple method for fitting a line to a set of points on the plane.

Fitting a Line to a Set of Points

Let $\mathbf{p}_i = (x_i, y_i)^T$ be a set of $m \geq 2$ points on the plane, and let

$$ax + by - c = 0$$

be the equation of a line. If the lefthand side of this equation is multiplied by a nonzero constant, the line does not change. Thus, we can assume without loss of generality that

$$\|\mathbf{n}\| = a^2 + b^2 = 1 \ , \tag{78}$$

where the unit vector $\mathbf{n} = (a, b)^T$, orthogonal to the line, is called the *line normal*.

The distance from the line to the origin is $|c|$ (see figure 34), and the distance between the line $\mathbf{n}$ and a point $\mathbf{p}_i$ is equal to

$$d_i = |ax_i + by_i - c| = |\mathbf{p}_i^T \mathbf{n} - c| \ . \tag{79}$$

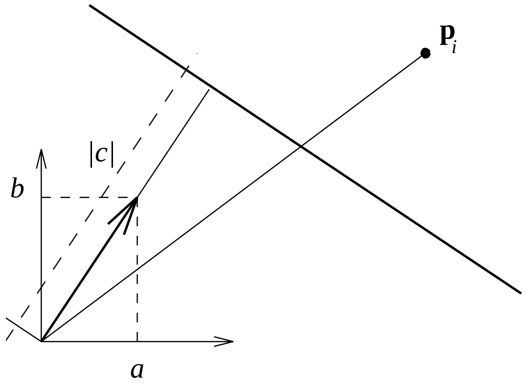

Figure 34: The distance between point $\mathbf{p}_i = (x_i, y_i)^T$ and line $ax + by - c = 0$ is $|ax_i + by_i - c|$.

The best-fit line minimizes the sum of the squared distances. Thus, if we let $\mathbf{d} = (d_1, \ldots, d_m)$ and $P = (\mathbf{p}_1 \ldots, \mathbf{p}_m)^T$, the best-fit line achieves the

$$\min_{\|\mathbf{n}\|=1} \|\mathbf{d}\|^2 = \min_{\|\mathbf{n}\|=1} \|P\mathbf{n} - c\mathbf{1}\|^2 \ . \tag{80}$$

In equation (80), $\mathbf{1}$ is a vector of m ones.

The Best Line Fit

Since the third line parameter c does not appear in the constraint (78), at the minimum (80) we must have

$$\frac{\partial \|\mathbf{d}\|^2}{\partial c} = 0 \; . \qquad (81)$$

If we define the centroid $\mathbf{p}$ of all the points $\mathbf{p}_i$ as

$$\mathbf{p} = \frac{1}{m} P^T \mathbf{1} \; ,$$

equation (81) yields

$$\begin{aligned} \frac{\partial \|\mathbf{d}\|^2}{\partial c} &= \frac{\partial}{\partial c} \left(\mathbf{n}^T P^T - c\mathbf{1}^T \right) \left(P\mathbf{n} - \mathbf{1}c \right) \\ &= \frac{\partial}{\partial c} \left(\mathbf{n}^T P^T P \mathbf{n} + c^2 \mathbf{1}^T \mathbf{1} - 2\mathbf{n}^T P^T c \mathbf{1} \right) \\ &= 2 \left(mc - \mathbf{n}^T P^T \mathbf{1} \right) = 0 \end{aligned}$$

from which we obtain

$$c = \frac{1}{m} \mathbf{n}^T P^T \mathbf{1} \; ,$$

that is,

$$c = \mathbf{p}^T \mathbf{n} \; .$$

By replacing this expression into equation (80), we obtain

$$\min_{\|\mathbf{n}\|=1} \|\mathbf{d}\|^2 = \min_{\|\mathbf{n}\|=1} \|P\mathbf{n} - \mathbf{1}\mathbf{p}^T \mathbf{n}\|^2 = \min_{\|\mathbf{n}\|=1} \|Q\mathbf{n}\|^2 \; ,$$

where $Q = P - \mathbf{1}\mathbf{p}^T$ collects the *centered* coordinates of the m points. We can solve this constrained minimization problem by theorem 10.6. Equivalently, and in order to emphasize the geometric meaning of signular values and vectors, we can recall that if $\mathbf{n}$ is on a circle, the shortest vector of the form $Q\mathbf{n}$ is obtained when $\mathbf{n}$ is the right singular vector $\mathbf{v}_2$ corresponding to the smaller σ_2 of the two singular values of Q. Furthermore, since $Q\mathbf{v}_2$ has norm σ_2, the residue is

$$\min_{\|\mathbf{n}\|=1} \|\mathbf{d}\| = \sigma_2$$

and more specifically the distances d_i are given by

$$\mathbf{d} = \sigma_2 \mathbf{u}_2$$

where $\mathbf{u}_2$ is the left singular vector corresponding to σ_2. In fact, when $\mathbf{n} = \mathbf{v}_2$, the SVD

$$Q = U \Sigma V^T = \sum_{i=1}^{2} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

yields

$$Q\mathbf{n} = Q\mathbf{v}_2 = \sum_{i=1}^{2} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_2 = \sigma_2 \mathbf{u}_2$$

because $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthonormal vectors.

To summarize, to fit a line (a, b, c) to a set of m points $\mathbf{p}_i$ collected in the $m \times 2$ matrix $P = (\mathbf{p}_1 \ldots, \mathbf{p}_m)^T$, proceed as follows:

1. compute the centroid of the points (**1** is a vector of m ones):

$$\mathbf{p} = \frac{1}{m} P^T \mathbf{1}$$

2. form the matrix of centered coordinates:

$$Q = P - \mathbf{1}\mathbf{p}^T$$

3. compute the SVD of Q:

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

4. the line normal is the second column of the 2×2 matrix V:

$$\mathbf{n} = (a, b)^T = \mathbf{v}_2 \ ,$$

5. the third coefficient of the line is

$$c = \mathbf{p}^T \mathbf{n}$$

6. the residue of the fit is

$$\min_{\|\mathbf{n}\|=1} \|\mathbf{d}\| = \sigma_2$$

The following `matlab` code implements the line fitting method.

```
function [l, residue] = linefit(P)
% check input matrix sizes
[m n] = size(P);
if n ~= 2, error('matrix P must be m x 2'), end
if m < 2, error('Need at least two points'), end
one = ones(m, 1);
% centroid of all the points
p = (P' * one) / m;
% matrix of centered coordinates
Q = P - one * p';
[U Sigma V] = svd(Q);
% the line normal is the second column of V
```

```
n = V(:, 2);
% assemble the three line coefficients into a column vector
l = [n ; p' * n];
% the smallest singular value of Q
% measures the residual fitting error
residue = Sigma(2, 2);
```

A useful exercise is to think how this procedure, or something close to it, can be adapted to fit a set of data points in $\mathbf{R}^m$ with an affine subspace of given dimension n. An affine subspace is a linear subspace plus a point, just like an arbitrary line is a line through the origin plus a point. Here "plus" means the following. Let L be a linear space. Then an affine space has the form

$$A = \mathbf{p} + L = \{\mathbf{a} \,|\, \mathbf{a} = \mathbf{p} + \mathbf{l} \text{ and } \mathbf{l} \in L\} \ .$$

Hint: minimizing the distance between a point and a subspace is equivalent to maximizing the norm of the projection of the point onto the subspace. The fitting problem (including fitting a line to a set of points) can be cast either as a maximization or a minimization problem.