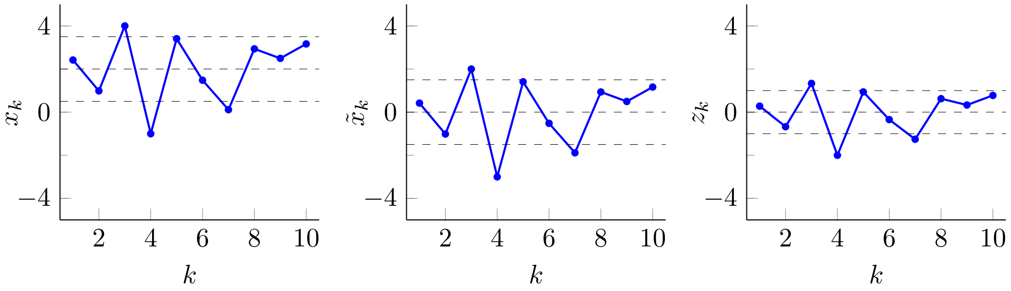

Figure 3.5 A 10-vector x, the de-meaned vector $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$, and the standardized vector $z = (1/\mathbf{std}(x))\tilde{x}$. The horizontal dashed lines indicate the mean and the standard deviation of each vector. The middle line is the mean; the distance between the middle line and the other two is the standard deviation.

Standardization. For any vector x, we refer to $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$ as the de-meaned version of x, since it has average or mean value zero. If we then divide by the RMS value of $\tilde{x}$ (which is the standard deviation of x), we obtain the vector

$$z = \frac{1}{\mathbf{std}(x)}(x - \mathbf{avg}(x)\mathbf{1}).$$

This vector is called the *standardized* version of x. It has mean zero, and standard deviation one. Its entries are sometimes called the *z-scores* associated with the original entries of x. For example, $z_4 = 1.4$ means that x_4 is 1.4 standard deviations above the mean of the entries of x. Figure 3.5 shows an example.

The standardized values for a vector give a simple way to interpret the original values in the vectors. For example, if an n-vector x gives the values of some medical test of n patients admitted to a hospital, the standardized values or z-scores tell us how high or low, compared to the population, that patient's value is. A value $z_6 = -3.2$, for example, means that patient 6 has a very low value of the measurement; whereas $z_{22} = 0.3$ says that patient 22's value is quite close to the average value.

3.4 Angle

Cauchy–Schwarz inequality. An important inequality that relates norms and inner products is the *Cauchy–Schwarz inequality*:

$$|a^T b| \leq \|a\| \, \|b\|$$

for any n-vectors a and b. Written out in terms of the entries, this is

$$|a_1 b_1 + \cdots + a_n b_n| \leq \left(a_1^2 + \cdots + a_n^2\right)^{1/2} \left(b_1^2 + \cdots + b_n^2\right)^{1/2},$$

which looks more intimidating. This inequality is attributed to the mathematician Augustin-Louis Cauchy; Hermann Schwarz gave the derivation given below.

The Cauchy–Schwarz inequality can be shown as follows. The inequality clearly holds if $a = 0$ or $b = 0$ (in this case, both sides of the inequality are zero). So we suppose now that $a \neq 0$, $b \neq 0$, and define $\alpha = \|a\|$, $\beta = \|b\|$. We observe that

$$\begin{array}{rcl} 0 & \leq & \|\beta a - \alpha b\|^2 \\ & = & \|\beta a\|^2 - 2(\beta a)^T(\alpha b) + \|\alpha b\|^2 \\ & = & \beta^2\|a\|^2 - 2\beta\alpha(a^T b) + \alpha^2\|b\|^2 \\ & = & \|b\|^2\|a\|^2 - 2\|b\|\|a\|(a^T b) + \|a\|^2\|b\|^2 \\ & = & 2\|a\|^2\|b\|^2 - 2\|a\|\,\|b\|(a^T b). \end{array}$$

Dividing by $2\|a\|\,\|b\|$ yields $a^T b \leq \|a\|\,\|b\|$. Applying this inequality to $-a$ and b we obtain $-a^T b \leq \|a\|\,\|b\|$. Putting these two inequalities together we get the Cauchy–Schwarz inequality, $|a^T b| \leq \|a\|\,\|b\|$.

This argument also reveals the conditions on a and b under which they satisfy the Cauchy–Schwarz inequality with equality. This occurs only if $\|\beta a - \alpha b\| = 0$, *i.e.*, $\beta a = \alpha b$. This means that each vector is a scalar multiple of the other (in the case when they are nonzero). This statement remains true when either a or b is zero. So the Cauchy–Schwarz inequality holds with equality when one of the vectors is a multiple of the other; in all other cases, it holds with strict inequality.

Verification of triangle inequality. We can use the Cauchy–Schwarz inequality to verify the triangle inequality. Let a and b be any vectors. Then

$$\begin{array}{rcl} \|a+b\|^2 & = & \|a\|^2 + 2a^T b + \|b\|^2 \\ & \leq & \|a\|^2 + 2\|a\|\|b\| + \|b\|^2 \\ & = & (\|a\| + \|b\|)^2, \end{array}$$

where we used the Cauchy–Schwarz inequality in the second line. Taking the squareroot we get the triangle inequality, $\|a+b\| \leq \|a\| + \|b\|$.

Angle between vectors. The *angle* between two nonzero vectors a, b is defined as

$$\theta = \arccos\left(\frac{a^T b}{\|a\|\,\|b\|}\right)$$

where arccos denotes the inverse cosine, normalized to lie in the interval $[0, \pi]$. In other words, we define θ as the unique number between 0 and π that satisfies

$$a^T b = \|a\|\,\|b\| \cos\theta.$$

The angle between a and b is written as $\angle(a, b)$, and is sometimes expressed in degrees. (The default angle unit is *radians*; 360° is 2π radians.) For example, $\angle(a, b) = 60^\circ$ means $\angle(a, b) = \pi/3$, *i.e.*, $a^T b = (1/2)\|a\|\|b\|$.

The angle coincides with the usual notion of angle between vectors, when they have dimension two or three, and they are thought of as displacements from a common point. For example, the angle between the vectors $a = (1, 2, -1)$ and $b = (2, 0, -3)$ is

$$\arccos\left(\frac{5}{\sqrt{6}\,\sqrt{13}}\right) = \arccos(0.5661) = 0.9690 = 55.52^\circ$$

(to 4 digits). But the definition of angle is more general; we can refer to the angle between two vectors with dimension 100.

The angle is a symmetric function of a and b: We have $\angle(a, b) = \angle(b, a)$. The angle is not affected by scaling each of the vectors by a positive scalar: We have, for any vectors a and b, and any positive numbers α and β,

$$\angle(\alpha a, \beta b) = \angle(a, b).$$

Acute and obtuse angles. Angles are classified according to the sign of $a^T b$. Suppose a and b are nonzero vectors of the same size.

- If the angle is $\pi/2 = 90^\circ$, *i.e.*, $a^T b = 0$, the vectors are said to be *orthogonal*. We write $a \perp b$ if a and b are orthogonal. (By convention, we also say that a zero vector is orthogonal to any vector.)
- If the angle is zero, which means $a^T b = \|a\|\|b\|$, the vectors are *aligned*. Each vector is a positive multiple of the other.
- If the angle is $\pi = 180^\circ$, which means $a^T b = -\|a\|\,\|b\|$, the vectors are *anti-aligned*. Each vector is a negative multiple of the other.
- If $\angle(a, b) < \pi/2 = 90^\circ$, the vectors are said to make an *acute angle*. This is the same as $a^T b > 0$, *i.e.*, the vectors have positive inner product.
- If $\angle(a, b) > \pi/2 = 90^\circ$, the vectors are said to make an *obtuse angle*. This is the same as $a^T b < 0$, *i.e.*, the vectors have negative inner product.

These definitions are illustrated in figure 3.6.

Examples.

- *Spherical distance.* Suppose a and b are 3-vectors that represent two points that lie on a sphere of radius R (for example, locations on earth). The spherical distance between them, measured along the sphere, is given by $R\angle(a, b)$. This is illustrated in figure 3.7.
- *Document similarity via angles.* If n-vectors x and y represent the word counts for two documents, their angle $\angle(x, y)$ can be used as a measure of document dissimilarity. (When using angle to measure document dissimilarity, either word counts or histograms can be used; they produce the same result.) As an example, table 3.2 gives the angles in degrees between the word histograms in the example at the end of §3.2.

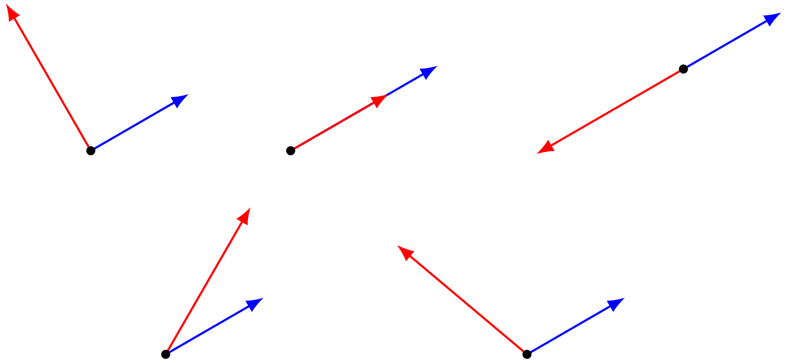

Figure 3.6 *Top row.* Examples of orthogonal, aligned, and anti-aligned vectors. *Bottom row.* Vectors that make an obtuse and an acute angle.

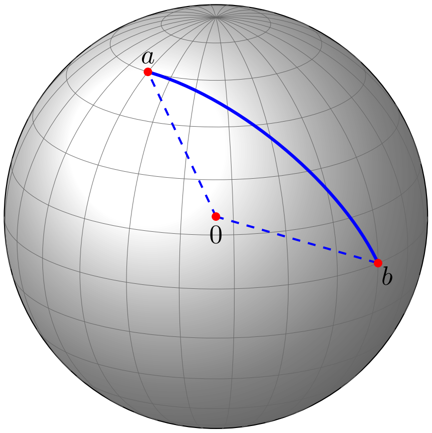

Figure 3.7 Two points a and b on a sphere with radius R and center at the origin. The spherical distance between the points is equal to $R\angle(a,b)$.

	Veterans	Memorial	Academy	Golden Globe	Super Bowl
	Day	Day	Awards	Awards	
Veterans Day	0	60.6	85.7	87.0	87.7
Memorial Day	60.6	0	85.6	87.5	87.5
Academy A.	85.7	85.6	0	58.7	85.7
Golden Globe A.	87.0	87.5	58.7	0	86.0
Super Bowl	87.7	87.5	86.1	86.0	0

Table 3.2 Pairwise angles (in degrees) between word histograms of five Wikipedia articles.

Norm of sum via angles. For vectors x and y we have

$$\|x+y\|^2 = \|x\|^2 + 2x^Ty + \|y\|^2 = \|x\|^2 + 2\|x\|\|y\|\cos\theta + \|y\|^2, \tag{3.6}$$

where $\theta = \angle(x,y)$. (The first equality comes from (3.1).) From this we can make several observations.

- If x and y are aligned ($\theta = 0$), we have $\|x+y\| = \|x\| + \|y\|$. Thus, their norms add.
- If x and y are orthogonal ($\theta = 90°$), we have $\|x+y\|^2 = \|x\|^2 + \|y\|^2$. In this case the norm-squared values add, and we have $\|x+y\| = \sqrt{\|x\|^2 + \|y\|^2}$. This formula is sometimes called the *Pythagorean theorem*, after the Greek mathematician Pythagoras of Samos.

Correlation coefficient. Suppose a and b are n-vectors, with associated de-meaned vectors

$$\tilde{a} = a - \mathbf{avg}(a)\mathbf{1}, \qquad \tilde{b} = b - \mathbf{avg}(b)\mathbf{1}.$$

Assuming these de-meaned vectors are not zero (which occurs when the original vectors have all equal entries), we define their *correlation coefficient* as

$$\rho = \frac{\tilde{a}^T\tilde{b}}{\|\tilde{a}\|\,\|\tilde{b}\|}. \tag{3.7}$$

Thus, $\rho = \cos\theta$, where $\theta = \angle(\tilde{a},\tilde{b})$. We can also express the correlation coefficient in terms of the vectors u and v obtained by standardizing a and b. With $u = \tilde{a}/\,\mathbf{std}(a)$ and $v = \tilde{b}/\,\mathbf{std}(b)$, we have

$$\rho = u^Tv/n. \tag{3.8}$$

(We use $\|u\| = \|v\| = \sqrt{n}$.)

This is a symmetric function of the vectors: The correlation coefficient between a and b is the same as the correlation coefficient between b and a. The Cauchy–Schwarz inequality tells us that the correlation coefficient ranges between -1 and $+1$. For this reason, the correlation coefficient is sometimes expressed as a percentage. For example, $\rho = 30\%$ means $\rho = 0.3$. When $\rho = 0$, we say the vectors are *uncorrelated*. (By convention, we say that a vector with all entries equal is uncorrelated with any vector.)

The correlation coefficient tells us how the entries in the two vectors vary together. High correlation (say, $\rho = 0.8$) means that entries of a and b are typically above their mean for many of the same entries. The extreme case $\rho = 1$ occurs only if the vectors $\tilde{a}$ and $\tilde{b}$ are aligned, which means that each is a positive multiple of the other, and the other extreme case $\rho = -1$ occurs only when $\tilde{a}$ and $\tilde{b}$ are negative multiples of each other. This idea is illustrated in figure 3.8, which shows the entries of two vectors, as well as a scatter plot of them, for cases with correlation near 1, near -1, and near 0.

The correlation coefficient is often used when the vectors represent time series, such as the returns on two investments over some time interval, or the rainfall in two locations over some time interval. If they are highly correlated (say, $\rho > 0.8$),

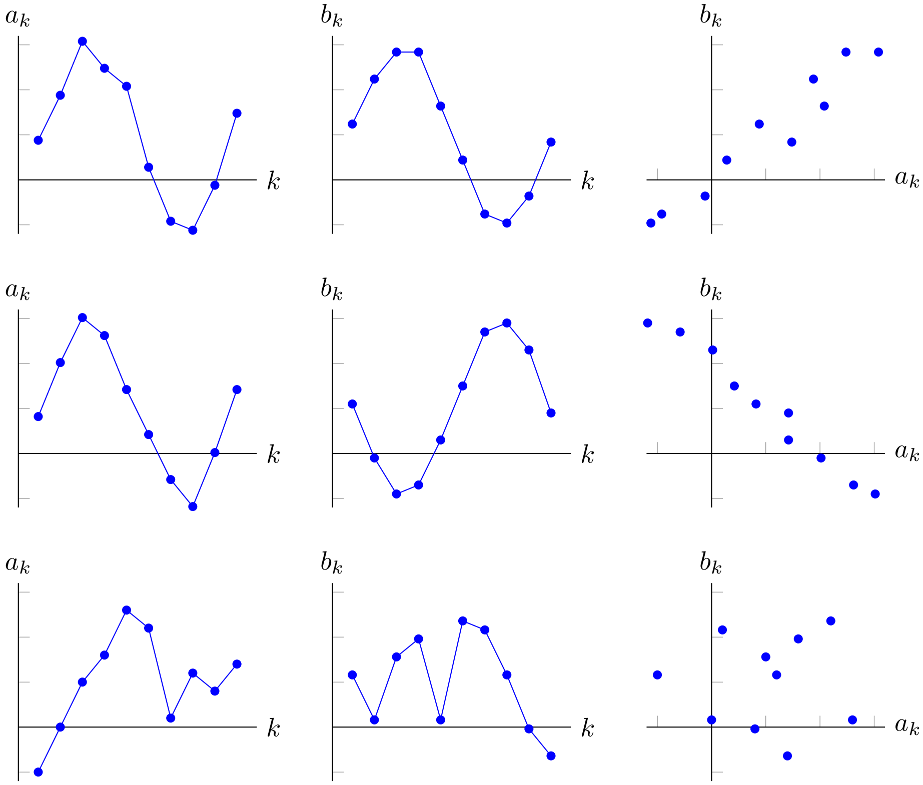

Figure 3.8 Three pairs of vectors a, b of length 10, with correlation coefficients 0.968 (top), -0.988 (middle), and 0.004 (bottom).

the two time series are typically above their mean values at the same times. For example, we would expect the rainfall time series at two nearby locations to be highly correlated. As another example, we might expect the returns of two similar companies, in the same business area, to be highly correlated.

Standard deviation of sum. We can derive a formula for the standard deviation of a sum from (3.6):

$$\mathbf{std}(a+b) = \sqrt{\mathbf{std}(a)^2 + 2\rho\,\mathbf{std}(a)\,\mathbf{std}(b) + \mathbf{std}(b)^2}. \tag{3.9}$$

To derive this from (3.6) we let $\tilde{a}$ and $\tilde{b}$ denote the de-meaned versions of a and b. Then $\tilde{a}+\tilde{b}$ is the de-meaned version of $a+b$, and $\mathbf{std}(a+b)^2 = \|\tilde{a}+\tilde{b}\|^2/n$. Now using (3.6) and $\rho = \cos\angle(\tilde{a},\tilde{b})$, we get

$$\begin{array}{rcl} n\,\mathbf{std}(a+b)^2 & = & \|\tilde{a}+\tilde{b}\|^2 \\ & = & \|\tilde{a}\|^2 + 2\rho\|\tilde{a}\|\|\tilde{b}\| + \|\tilde{b}\|^2 \\ & = & n\,\mathbf{std}(a)^2 + 2\rho n\,\mathbf{std}(a)\,\mathbf{std}(b) + n\,\mathbf{std}(b)^2. \end{array}$$

Dividing by n and taking the squareroot yields the formula above.

If $\rho = 1$, the standard deviation of the sum of vectors is the sum of their standard deviations, *i.e.*,

$$\mathbf{std}(a+b) = \mathbf{std}(a) + \mathbf{std}(b).$$

As ρ decreases, the standard deviation of the sum decreases. When $\rho = 0$, *i.e.*, a and b are uncorrelated, the standard deviation of the sum $a+b$ is

$$\mathbf{std}(a+b) = \sqrt{\mathbf{std}(a)^2 + \mathbf{std}(b)^2},$$

which is smaller than $\mathbf{std}(a)+\mathbf{std}(b)$ (unless one of them is zero). When $\rho = -1$, the standard deviation of the sum is as small as it can be,

$$\mathbf{std}(a+b) = |\,\mathbf{std}(a) - \mathbf{std}(b)|.$$

Hedging investments. Suppose that vectors a and b are time series of returns for two assets with the same return (average) μ and risk (standard deviation) σ, and correlation coefficient ρ. (These are the traditional symbols used.) The vector $c = (a+b)/2$ is the time series of returns for an investment with 50% in each of the assets. This blended investment has the same return as the original assets, since

$$\mathbf{avg}(c) = \mathbf{avg}((a+b)/2) = (\mathbf{avg}(a) + \mathbf{avg}(b))/2 = \mu.$$

The risk (standard deviation) of this blended investment is

$$\mathbf{std}(c) = \sqrt{2\sigma^2 + 2\rho\sigma^2}/2 = \sigma\sqrt{(1+\rho)/2},$$

using (3.9). From this we see that the risk of the blended investment is never more than the risk of the original assets, and is smaller when the correlation of the original asset returns is smaller. When the returns are uncorrelated, the risk is a factor $1/\sqrt{2} = 0.707$ smaller than the risk of the original assets. If the asset returns are strongly negatively correlated (*i.e.*, ρ is near -1), the risk of the blended investment is much smaller than the risk of the original assets. Investing in two assets with uncorrelated, or negatively correlated, returns is called *hedging* (which is short for 'hedging your bets'). Hedging reduces risk.

Units for heterogeneous vector entries. When the entries of vectors represent different types of quantities, the choice of units used to represent each entry affects the angle, standard deviation, and correlation between a pair of vectors. The discussion on page 51, about how the choice of units can affect distances between pairs of vectors, therefore applies to these quantities as well. The general rule of thumb is to choose units for different entries so the typical vector entries have similar sizes or ranges of values.

3.5 Complexity

Computing the norm of an n-vector requires n multiplications (to square each entry), $n-1$ additions (to add the squares), and one squareroot. Even though computing the squareroot typically takes more time than computing the product or sum of two numbers, it is counted as just one flop. So computing the norm takes $2n$ flops. The cost of computing the RMS value of an n-vector is the same, since we can ignore the two flops involved in division by $\sqrt{n}$. Computing the distance between two vectors costs $3n$ flops, and computing the angle between them costs $6n$ flops. All of these operations have order n.

De-meaning an n-vector requires $2n$ flops (n for forming the average and another n flops for subtracting the average from each entry). The standard deviation is the RMS value of the de-meaned vector, and this calculation takes $4n$ flops ($2n$ for computing the de-meaned vector and $2n$ for computing its RMS value). Equation (3.5) suggests a slightly more efficient method with a complexity of $3n$ flops: first compute the average (n flops) and RMS value ($2n$ flops), and then find the standard deviation as $\mathbf{std}(x) = (\mathbf{rms}(x)^2 - \mathbf{avg}(x)^2)^{1/2}$. Standardizing an n-vector costs $5n$ flops. The correlation coefficient between two vectors costs $10n$ flops to compute. These operations also have order n.

As a slightly more involved computation, suppose that we wish to determine the nearest neighbor among a collection of k n-vectors $z_1, \ldots, z_k$ to another n-vector x. (This will come up in the next chapter.) The simple approach is to compute the distances $\|x - z_i\|$ for $i = 1, \ldots, k$, and then find the minimum of these. (Sometimes a comparison of two numbers is also counted as a flop.) The cost of this is $3kn$ flops to compute the distances, and $k-1$ comparisons to find the minimum. The latter term can be ignored, so the flop count is $3kn$. The order of finding the nearest neighbor in a collection of k n-vectors is kn.