bedrooms. These are very different types of features, since the first one is a physical area, and the second one is a count, *i.e.*, an integer. In the example on page 39, we chose the unit used to represent the first feature, area, to be thousands of square feet. With this choice of unit used to represent house area, the numerical values of both of these features range from around 1 to 5; their values have roughly the same magnitude. When we determine the distance between feature vectors associated with two houses, the difference in the area (in thousands of square feet), and the difference in the number of bedrooms, play equal roles.

For example, consider three houses with feature vectors

$$x = (1.6, 2), \quad y = (1.5, 2), \quad z = (1.6, 4).$$

The first two are 'close' or 'similar' since $\|x - y\| = 0.1$ is small (compared to the norms of x and y, which are around 2.5). This matches our intuition that the first two houses are similar, since they both have two bedrooms and are close in area. The third house would be considered 'far' or 'different' from the first two houses, and rightly so since it has four bedrooms instead of two.

To appreciate the significance of our choice of units in this example, suppose we had chosen instead to represent house area directly in square feet, and not thousands of square feet. The three houses above would then be represented by feature vectors

$$\tilde{x} = (1600, 2), \quad \tilde{y} = (1500, 2), \quad \tilde{z} = (1600, 4).$$

The distance between the first and third houses is now 2, which is very small compared to the norms of the vectors (which are around 1600). The distance between the first and second houses is much larger. It seems strange to consider a two-bedroom house and a four-bedroom house as 'very close', while two houses with the same number of bedrooms and similar areas are much more dissimilar. The reason is simple: With our choice of square feet as the unit to measure house area, distances are very strongly influenced by differences in area, with number of bedrooms playing a much smaller (relative) role.

3.3 Standard deviation

For any vector x, the vector $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$ is called the associated *de-meaned* vector, obtained by subtracting from each entry of x the mean value of the entries. (This is not standard notation; *i.e.*, $\tilde{x}$ is not generally used to denote the de-meaned vector.) The mean value of the entries of $\tilde{x}$ is zero, *i.e.*, $\mathbf{avg}(\tilde{x}) = 0$. This explains why $\tilde{x}$ is called the de-meaned version of x; it is x with its mean removed. The de-meaned vector is useful for understanding how the entries of a vector deviate from their mean value. It is zero if all the entries in the original vector x are the same.

The *standard deviation* of an n-vector x is defined as the RMS value of the de-meaned vector $x - \mathbf{avg}(x)\mathbf{1}$, *i.e.*,

$$\mathbf{std}(x) = \sqrt{\frac{(x_1 - \mathbf{avg}(x))^2 + \cdots + (x_n - \mathbf{avg}(x))^2}{n}}.$$

This is the same as the RMS deviation between a vector x and the vector all of whose entries are $\mathbf{avg}(x)$. It can be written using the inner product and norm as

$$\mathbf{std}(x) = \frac{\|x - (\mathbf{1}^T x/n)\mathbf{1}\|}{\sqrt{n}}. \tag{3.4}$$

The standard deviation of a vector x tells us the typical amount by which its entries deviate from their average value. The standard deviation of a vector is zero only when all its entries are equal. The standard deviation of a vector is small when the entries of the vector are nearly the same.

As a simple example consider the vector $x = (1, -2, 3, 2)$. Its mean or average value is $\mathbf{avg}(x) = 1$, so the de-meaned vector is $\tilde{x} = (0, -3, 2, 1)$. Its standard deviation is $\mathbf{std}(x) = 1.872$. We interpret this number as a 'typical' value by which the entries differ from the mean of the entries. These numbers are 0, 3, 2, and 1, so 1.872 is reasonable.

We should warn the reader that another slightly different definition of the standard deviation of a vector is widely used, in which the denominator $\sqrt{n}$ in (3.4) is replaced with $\sqrt{n-1}$ (for $n \geq 2$). In this book we will only use the definition (3.4).

In some applications the Greek letter σ (sigma) is traditionally used to denote standard deviation, while the mean is denoted μ (mu). In this notation we have, for an n-vector x,

$$\mu = \mathbf{1}^T x/n, \qquad \sigma = \|x - \mu\mathbf{1}\|/\sqrt{n}.$$

We will use the symbols $\mathbf{avg}(x)$ and $\mathbf{std}(x)$, switching to μ and σ only with explanation, when describing an application that traditionally uses these symbols.

Average, RMS value, and standard deviation. The average, RMS value, and standard deviation of a vector are related by the formula

$$\mathbf{rms}(x)^2 = \mathbf{avg}(x)^2 + \mathbf{std}(x)^2. \tag{3.5}$$

This formula makes sense: $\mathbf{rms}(x)^2$ is the mean square value of the entries of x, which can be expressed as the square of the mean value, plus the mean square fluctuation of the entries of x around their mean value. We can derive this formula from our vector notation formula for $\mathbf{std}(x)$ given above. We have

$$\begin{aligned}
\mathbf{std}(x)^2 &= (1/n)\|x - (\mathbf{1}^T x/n)\mathbf{1}\|^2 \\
&= (1/n)(x^T x - 2x^T(\mathbf{1}^T x/n)\mathbf{1} + ((\mathbf{1}^T x/n)\mathbf{1})^T((\mathbf{1}^T x/n)\mathbf{1})) \\
&= (1/n)(x^T x - (2/n)(\mathbf{1}^T x)^2 + n(\mathbf{1}^T x/n)^2) \\
&= (1/n)x^T x - (\mathbf{1}^T x/n)^2 \\
&= \mathbf{rms}(x)^2 - \mathbf{avg}(x)^2,
\end{aligned}$$

which can be re-arranged to obtain the identity (3.5) above. This derivation uses many of the properties for norms and inner products, and should be read carefully to understand every step. In the second line, we expand the norm-square of the sum of two vectors. In the third line, we use the commutative property of scalar-vector multiplication, moving scalars such as $(\mathbf{1}^T x/n)$ to the front of each term, and also the fact that $\mathbf{1}^T\mathbf{1} = n$.

Examples.

- *Mean return and risk.* Suppose that an n-vector represents a time series of return on an investment, expressed as a percentage, in n time periods over some interval of time. Its average gives the mean return over the whole interval, often shortened to its *return*. Its standard deviation is a measure of how variable the return is, from period to period, over the time interval, *i.e.*, how much it typically varies from its mean, and is often called the (per period) *risk* of the investment. Multiple investments can be compared by plotting them on a *risk-return plot*, which gives the mean and standard deviation of the returns of each of the investments over some interval. A desirable return history vector has high mean return and low risk; this means that the returns in the different periods are consistently high. Figure 3.4 shows an example.

- *Temperature or rainfall.* Suppose that an n-vector is a time series of the daily average temperature at a particular location, over a one year period. Its average gives the average temperature at that location (over the year) and its standard deviation is a measure of how much the temperature varied from its average value. We would expect the average temperature to be high and the standard deviation to be low in a tropical location, and the opposite for a location with high latitude.

Chebyshev inequality for standard deviation. The Chebyshev inequality (3.2) can be transcribed to an inequality expressed in terms of the mean and standard deviation: If k is the number of entries of x that satisfy $|x_i - \mathbf{avg}(x)| \geq a$, then $k/n \leq (\mathbf{std}(x)/a)^2$. (This inequality is only interesting for $a > \mathbf{std}(x)$.) For example, at most $1/9 = 11.1\%$ of the entries of a vector can deviate from the mean value $\mathbf{avg}(x)$ by 3 standard deviations or more. Another way to state this is: The fraction of entries of x within α standard deviations of $\mathbf{avg}(x)$ is at least $1 - 1/\alpha^2$ (for $\alpha > 1$).

As an example, consider a time series of return on an investment, with a mean return of 8%, and a risk (standard deviation) 3%. By the Chebyshev inequality, the fraction of periods with a loss (*i.e.*, $x_i \leq 0$) is no more than $(3/8)^2 = 14.1\%$. (In fact, the fraction of periods when the return is either a loss, $x_i \leq 0$, or very good, $x_i \geq 16\%$, is together no more than 14.1%.)

Properties of standard deviation.

- *Adding a constant.* For any vector x and any number a, we have $\mathbf{std}(x+a\mathbf{1}) = \mathbf{std}(x)$. Adding a constant to every entry of a vector does not change its standard deviation.

- *Multiplying by a scalar.* For any vector x and any number a, we have $\mathbf{std}(ax) = |a|\,\mathbf{std}(x)$. Multiplying a vector by a scalar multiplies the standard deviation by the absolute value of the scalar.

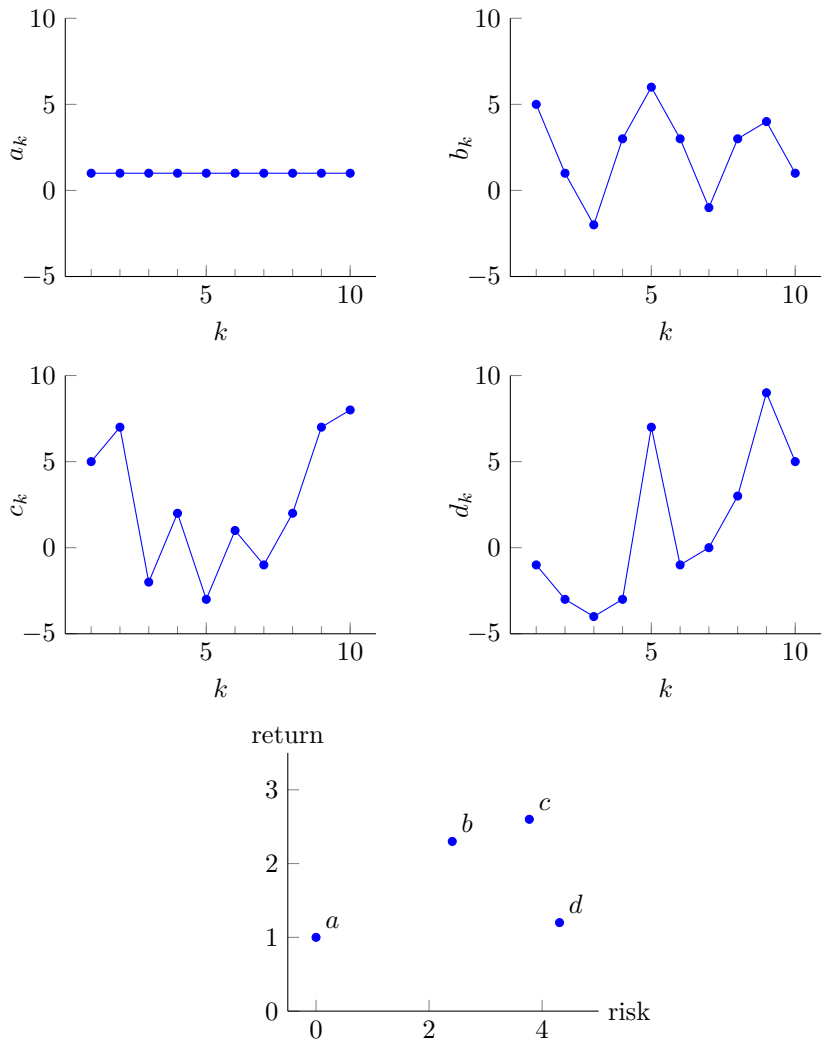

Figure 3.4 The vectors a, b, c, d represent time series of returns on investments over 10 periods. The bottom plot shows the investments in a risk-return plane, with return defined as the average value and risk as the standard deviation of the corresponding vector.

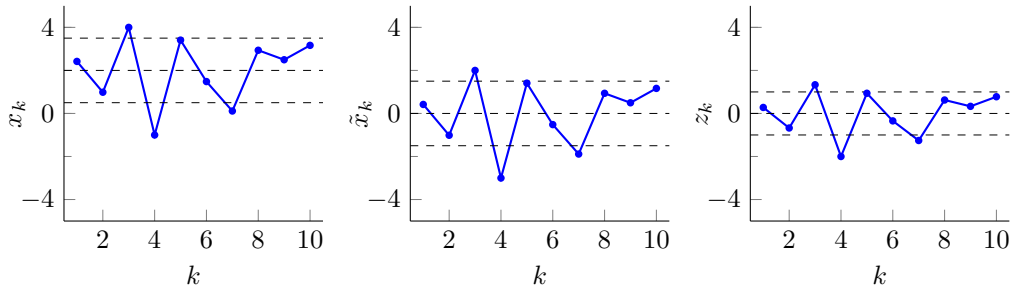

Figure 3.5 A 10-vector x, the de-meaned vector $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$, and the standardized vector $z = (1/\mathbf{std}(x))\tilde{x}$. The horizontal dashed lines indicate the mean and the standard deviation of each vector. The middle line is the mean; the distance between the middle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

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