Chapter 2. Vectors and Vector Spaces

Section 2.3. Centered Vectors and Variances and Covariances of Vectors

Note. In this section, we introduce some operations on vectors which relate to treating them as data sets. In particular, we define variance, covariance, and correlation.

Note. Recall that if $x_1, x_2, \ldots, x_n$ are data points from a population of size n, then

- the mean of the population is $\mu = \dfrac{\sum_{i=1}^n x_i}{n}$,
- the variance of the population is $\sigma^2 = \dfrac{\sum_{i=1}^n (x_i - \mu)^2}{n}$, and
- the standard deviation of the population is $\sigma = \sqrt{\dfrac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$.

Note. If $x_1, x_2, \ldots, x_n$ are data points *sampled* from a population, then

- the sample mean is $\overline{x} = \dfrac{\sum_{i=1}^n x_i}{n}$,
- the sample variance is $s^2 = \dfrac{\sum_{i=1}^n (x_i - \overline{x})^2}{(n-1)}$, and
- the sample standard deviation is $s = \sqrt{\displaystyle\sum_{i=1}^n (x_i - \overline{x})^2/(n-1)}$.

Definition. For a given n-vector x, its *centered counterpart*, denoted x_c, is $x_c = x - \overline{x}$ where $\overline{x}$ is the mean vector of x, $\overline{x} = [\overline{x}, \overline{x}, \ldots, \overline{x}]$. (Notice $\|\overline{x}\|^2 = n\overline{x}^2$.) Any n-vector with entries that sum to 0 is a *centered* vector.

Note. In Exercise 2.14, it is shown that for any $x, y \in \mathbb{R}^n$, $(x+y)_c = x_c + y_c$.

Note. We can interpret vector $\overline{x}$ as

$$\overline{x} = \mathrm{proj}_{1_n}(x) = \frac{\langle 1_n, x\rangle 1_n}{\|1_n\|^2} = \left(\frac{1_n^T x}{n}\right) 1_n$$

(recall that $\|1_n\|^2 = (\sqrt{n})^2 = n$). As shown in the class notes for Section 2.2 (see the Note after the definition of "projection"), $\mathrm{proj}_x(y) \perp (y - \mathrm{proj}_x(y))$ and so $\overline{x} \perp x - \overline{x}$ or $\overline{x} \perp x_c$. So by the Pythagoraean Theorem (see the same note), $\|x\|^2 = \|\overline{x}\|^2 + \|x_c\|^2$. Notice that in terms of vector entries (i.e., scalars) this implies

$$\sum_{i=1}^n x_i^2 = n\overline{x}^2 + \sum_{i=1}^n (x_i - \overline{x})^2 \text{ or } \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^n x_i^2}{n} - \overline{x}^2.$$

With $\{x_1, x_2, \ldots, x_n\}$ as a data set for a population, we have the variance $\sigma^2 = \sum_{i=1}^n (x_i - \overline{x})^2/n$ and mean $\mu = \overline{x}$ so that the familiar formula $\sigma^2 = \sum_{i=1}^n x_i^2/n - \mu^2$ (or $n\sigma^2 = \sum_{i=1}^n x_i^2 - n\mu^2$) is equivalent to the vector equation $\|x_c\|^2 = \|x\|^2 - \|\overline{x}\|^2$.

Definition. For $x \in \mathbb{R}^n$, the *scaled vector*, denoted x_s, is $x_s = \sqrt{n-1}x/\|x_c\|$. The *centered and scaled vector* is $x_{cs} = \sqrt{n-1}x_c/\|x_c\|$. The process of creation of x_{cs} from x is called *standardizing*. The *standard deviation* of x is the scalar quantity $s_x = \|x_c\|/\sqrt{n-1}$ and the *variance* of x is $V(x) = s_x^2 = \|x_c\|^2/(n-1)$.

Note. In terms of the entries of n-vector x we have

$$s_x = \frac{\|x_c\|}{\sqrt{n-1}} = \frac{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}{\sqrt{n-1}},$$

as would be expected for the *sample* standard deviation of a data set. Also, $V(x) = \sum_{i=1}^n (x_i - \overline{x})^2/(n-1)$, as would be expected for the *sample* variance.

Note. For $x \in \mathbb{R}^n$ we have the variance of the scaled vector x_s is

$$\begin{aligned} V(x_s) &= V\left(\sqrt{n-1}\frac{x}{\|x_c\|}\right) = \frac{\left\|\left(\sqrt{n-1}\frac{x}{\|x_c\|}\right)_c\right\|^2}{n-1} \\ &= \frac{\left\|\sqrt{n-1}\frac{x}{\|x_c\|} - \sqrt{n-1}\frac{\overline{x}}{\|x_c\|}\right\|^2}{n-1} \text{ since } \overline{ax} = a\overline{x} \\ &= \frac{\|x-\overline{x}\|^2}{\|x_c\|^2} = \frac{\|x-\overline{x}\|^2}{\|x-\overline{x}\|^2} \text{ since } x_c = x - \overline{x} \\ &= 1. \end{aligned}$$

thus the term "scaled" vector x_s is justified; its sample variance and sample standard deviation are both 1.

Note. Recall for sample (x_i, y_i) for $i = 1, 2, \ldots, n$ of discrete random variable pair (X, Y), the sample covariance is

$$\text{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}).$$

The sample correlation is

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}.$$

Definition. Let x and y be n-vectors. The (sample) *covariance* between x and y is

$$\mathrm{Cov}(x,y) = \frac{\langle x-\overline{x}, y-\overline{y}\rangle}{n-1} = \frac{\langle x_c, y_c\rangle}{n-1}.$$

Note. With 1_n as the summing vector, we have $x-\overline{x} = x-\overline{x}1_n$ and $y-\overline{y} = y-\overline{y}1_n$ (with the careful interpretation of $\overline{x}$ and $\overline{y}$ as scalars *or* vectors), so

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \frac{\langle x-\overline{x}, y-\overline{y}\rangle}{n-1} = \frac{\langle x-\overline{x}1_n, y-\overline{y}1_n\rangle}{n-1} \\
&= \frac{\langle x,y\rangle - \langle \overline{x}1_n, y\rangle - \langle x, \overline{y}1_n\rangle + \langle \overline{x}1_n, \overline{y}1_n\rangle}{n-1} \\
&= \frac{\langle x,y\rangle - \sum_{i=1}^n \overline{x}y_i - \sum_{i=1}^n x_i\overline{y} + n\overline{x}\,\overline{y}}{n-1} \\
&= \frac{\langle x,y\rangle - n\overline{x}\sum_{i=1}^n y_i/n - n\overline{y}\sum_{i=1}^n x_i/n + n\overline{x}\,\overline{y}}{n-1} \\
&= \frac{\langle x,y\rangle - n\overline{x}\,\overline{y} - n\overline{x}\,\overline{y} + n\overline{x}\,\overline{y}}{n-1} \\
&= \frac{\langle x,y\rangle - n\overline{x}\,\overline{y}}{n-1}.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{Cov}(x,x) &= \frac{\langle x-\overline{x}, x-\overline{x}\rangle}{n-1} = \frac{\langle x,x\rangle - 2\langle x,\overline{x}\rangle + \langle \overline{x},\overline{x}\rangle}{n-1} \\
&= \frac{\sum_{i=1}^n x_i^2 - 2\sum_{i=1}^n x_i\overline{x} + \sum_{i=1}^n \overline{x}^2}{n-1} = \frac{\sum_{i=1}^n (x_i^2 - 2x_i\overline{x} + \overline{x}^2)}{n-1} \\
&= \frac{\sum_{i=1}^n (x_i-\overline{x})^2}{n-1} = V(x).
\end{aligned}$$

Theorem 2.3.1. Properties of Covariance.

Let x, y, z be n-vectors and let $a \in \mathbb{R}$. Then:

1. $\mathrm{Cov}(a1_n, y) = 0$,

2. $\mathrm{Cov}(ax, y) = a\mathrm{Cov}(x, y)$,

3. $\mathrm{Cov}(y, y) = V(y)$,

4. $\mathrm{Cov}(x+z, y) = \mathrm{Cov}(x, y) + \mathrm{Cov}(z, y)$, in particular $\mathrm{Cov}(x+y, y) = \mathrm{Cov}(x, y) + V(y)$ and $\mathrm{Cov}(x + a1_n, y) = \mathrm{Cov}(x, y)$.

Definition. Let x and y be n-vectors. The *correlation* between x and y is

$$\mathrm{Corr}(x, y) = \mathrm{Cov}(x_{cs}, y_{cs}) = \left\langle \frac{x_c}{\|x_c\|}, \frac{y_c}{\|y_c\|} \right\rangle = \frac{\langle x_c, y_c\rangle}{\|x_c\|\|y_c\|}.$$

Note. Gentle describes correlation as "a measure of the extent to which the vectors point in the same direction." This is justified with the observation that $\mathrm{Corr}(x, y) = \dfrac{\langle x_c, y_c\rangle}{\|x_c\|\|y_c\|} = \cos\theta$ where θ is the angle between x_c and y_c (though not the angle between x and y; see Exercise 2.15). It then follows that (as we expect) $\mathrm{Corr}(x, y) \in [-1, 1]$.

Note. We can also express $\mathrm{Corr}(x, y)$ as

$$\mathrm{Corr}(x, y) = \frac{\langle x_c, y_c\rangle}{\|x_c\|\|y_c\|} = \frac{\langle x_c, y_c\rangle}{(n-1)\sqrt{\frac{\|x_c\|^2}{n-1}\frac{\|y_c\|^2}{n-1}}}$$

$$= \frac{\langle x_c, y_c\rangle}{(n-1)\sqrt{V(x)V(y)}} = \frac{\langle x - \overline{x}, y - \overline{y}\rangle}{(n-1)\sqrt{V(x)V(y)}} = \frac{\mathrm{Cov}(x, y)}{\sqrt{V(x)V(y)}}.$$

Note. As might be expected, scaling x by a scalar $a \in \mathbb{R}$ affects the correlation only in sign:

$$\text{Corr}(ax, y) = \frac{\langle (ax)_c, y_c \rangle}{\|(ax)_c\| \|y_c\|} = \frac{\langle ax - a\overline{x}, y_c \rangle}{\|ax - a\overline{x}\| \|y_c\|}$$

$$= \frac{a\langle x - \overline{x}, y_c \rangle}{|a| \|x - \overline{x}\| \|y_c\|} = \frac{a}{|a|} \frac{\langle x_c, y_c \rangle}{\|x_c\| \|y_c\|} = \text{sign}(a)\text{Corr}(x, y).$$

Revised: 5/27/2020