Connections between Inner Products, Vector Norms, Angles, and Sample Covariance and Correlation

This doc revisits some fundamental concepts in linear algebra and statistics: **inner products**, **vector norms**, **angles between vectors**, and how these relate to **sample covariance** and **correlation**. By the end, we'll see how the correlation between two data vectors is simply the cosine of the angle between them.

---

1. Inner Products (Scalar Products)

The **inner product** (also called the **scalar product** or **dot product**) is a way to multiply two vectors and get a single number (a scalar). For two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, the inner product is defined as:

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = \sum_{i=1}^{n} u_i v_i$$

This means you multiply corresponding components of the vectors and add up the results. For example, if $\mathbf{u} = [1, 2, 3]$ and $\mathbf{v} = [4, 5, 6]$, their inner product is:

$$\mathbf{u} \cdot \mathbf{v} = (1 \cdot 4) + (2 \cdot 5) + (3 \cdot 6) = 4 + 10 + 18 = 32$$

The inner product[1] has two key properties: 1. It measures how much one vector "points in the direction" of another. 2. It is closely related to the **angle** between the vectors.

---

[1] $\mathbf{u} \cdot \mathbf{v}$ is sometimes written as $< \mathbf{u}, \mathbf{v} >$ and often referred to as the *dot product*.

2. Vector Norms

The **norm** of a vector $\mathbf{u}$, denoted $\|\mathbf{u}\|$, is a measure of its length. For a vector in $\mathbb{R}^n$, the norm is defined as:

$$\|\mathbf{u}\| = \sqrt{\sum_{i=1}^{n} u_i^2}$$

This is essentially the Euclidean distance from the origin to the point represented by $\mathbf{u}$. For example, if $\mathbf{u} = [3, 4]$, its norm is:

$$\|\mathbf{u}\| = \sqrt{3^2 + 4^2} = \sqrt{9 + 16} = 5$$

The norm is related to the inner product because:

$$\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$$

---

3. Angles Between Vectors

The inner product and norm are used to define the **angle** between two vectors. For vectors $\mathbf{u}$ and $\mathbf{v}$, the cosine of the angle θ between them is given by:

$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

This formula connects geometry (angles) with algebra (inner products and norms). Let's break it down: - The numerator $\mathbf{u} \cdot \mathbf{v}$ measures how aligned the vectors are. - The denominator $\|\mathbf{u}\|\|\mathbf{v}\|$ scales the result by the lengths of the vectors.

If $\mathbf{u}$ and $\mathbf{v}$ point in exactly the same direction, $\cos\theta = 1$. If they are perpendicular, $\cos\theta = 0$. If they point in exactly opposite directions, $\cos\theta = -1$.

---

4. Sample Standard Deviation

Now, let's connect these ideas to statistics. Suppose we have two sets of data represented as vectors $\mathbf{x} = [x_1, x_2, \dots, x_n]$ and $\mathbf{y} = [y_1, y_2, \dots, y_n]$.

Define $\bar{x}$ and $\bar{y}$ as the means of $\mathbf{x}$ and $\mathbf{y}$, respectively. If we center the data by subtracting the means:

$$\mathbf{x}_c = [x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}]$$

$$\mathbf{y}_c = [y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y}]$$

The standard deviations are defined as:

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\|\mathbf{x}_c\|}{\sqrt{n-1}}$$

and

$$S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\|\mathbf{y}_c\|}{\sqrt{n-1}}$$

We can see that they are equal to the norms of the centered data vectors, divided by the square root of their number of entries minus one.

5. Sample Covariance and Correlation

The **sample covariance** measures how much $\mathbf{x}$ and $\mathbf{y}$ change together:

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

The **correlation** ρ between $\mathbf{x}$ and $\mathbf{y}$ is a normalized version of covariance, scaled to lie between -1 and 1:

$$\rho = \frac{S_{xy}}{S_x S_y}$$

where S_x and S_y are the standard deviations of $\mathbf{x}$ and $\mathbf{y}$.

---

6. Correlation as the Cosine of the Angle

Here's the key insight: **The correlation ρ is exactly the cosine of the angle between the centered versions of x and y.**

The correlation ρ can be rewritten as:

$$\rho = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\mathbf{x}_c \cdot \mathbf{y}_c}{\|\mathbf{x}_c\|\|\mathbf{y}_c\|}$$

This is exactly the formula for $\cos\theta$! So:

$$\rho = \cos\theta$$

The correlation is negative if the data vectors point in (roughly) opposite directions (their angle is between 180° and 90°, with a maximum of -1 if they are exactly opposite with an angle of 180°), zero if they are orthogonal (their angle is 90°), and positive if the point in similar directions (their angle is betwen 90° and 0°, with a maximum of 1 if they are exactly parallel with an angle of 0°).

Summary

1. The **inner product** measures alignment between vectors.
2. The **norm** measures the length of a vector.
3. The **angle** between vectors is determined by their inner product and norms.
4. The **standard deviation** of a data vector is simply the **norm of the centered data vector**, divided by the square root of their dimension.
5. **Sample covariance** measures how two data vectors vary together and is simply the **inner product of the centered data vectors** divided by their dimension.
6. **Correlation** is the cosine of the angle between the centered data vectors: their inner product divided by the product of their norms.

By understanding these relationships, you can see how geometry (angles) and statistics (correlation) are deeply connected. This is a powerful insight that will help you in both linear algebra and data analysis!

(adapted from DeepSeek R1 output)