CHAPTER 8 NORMS, INNER PRODUCTS AND ORTHOGONALITY

8.1 Norms and Distances

In applied mathematics, Norms are functions which measure the magnitude or length of a vector. They are commonly used to determine similarities between observations by measuring the distance between them. As we will see, there are many ways to define distance between two points.

Definition 8.1.1: Vector Norms and Distance Metrics

A Norm, or distance metric, is a function that takes a vector as input and returns a scalar quantity ($f : \mathbb{R}^n \to \mathbb{R}$). A vector norm is typically denoted by two vertical bars surrounding the input vector, $\|\mathbf{x}\|$, to signify that it is not just any function, but one that satisfies the following criteria:

1. If c is a scalar, then
$$\|c\mathbf{x}\| = |c|\|x\|$$
2. The triangle inequality:
$$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$
3. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$.
4. $\|\mathbf{x}\| \geq 0$ for any vector $\mathbf{x}$

We will not spend any time on these axioms or on the theoretical aspects of norms, but we will put a couple of these functions to good use in our studies, the first of which is the **Euclidean norm** or **2-norm**.

Definition 8.1.2: Euclidean Norm, $\| \star \|_2$

The **Euclidean Norm**, also known as the 2**-norm** simply measures the Euclidean length of a vector (i.e. a point's distance from the origin). Let $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Then,

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

If $\mathbf{x}$ is a column vector, then

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T \mathbf{x}}.$$

Often we will simply write $\| \star \|$ rather than $\| \star \|_2$ to denote the 2-norm, as it is by far the most commonly used norm.

This is merely the distance formula from undergraduate mathematics, measuring the distance between the point $\mathbf{x}$ and the origin. To compute the distance between two different points, say $\mathbf{x}$ and $\mathbf{y}$, we'd calculate

$$\|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Example 8.1.1: Euclidean Norm and Distance

Suppose I have two vectors in 3-space:

$$\mathbf{x} = (1, 1, 1) \text{ and } \mathbf{y} = (1, 0, 0)$$

Then the magnitude of $\mathbf{x}$ (i.e. its length or distance from the origin) is

$$\|\mathbf{x}\|_2 = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3}$$

and the magnitude of $\mathbf{y}$ is

$$\|\mathbf{y}\|_2 = \sqrt{1^2 + 0^2 + 0^2} = 1$$

and the distance between point $\mathbf{x}$ and point $\mathbf{y}$ is

$$\|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{(1-1)^2 + (1-0)^2 + (1-0)^2} = \sqrt{2}.$$

The Euclidean norm is crucial to many methods in data analysis as it measures the closeness of two data points.

Thus, to turn any vector into a **unit vector**, a vector with a length of 1, we need only to divide each of the entries in the vector by its Euclidean norm. This is a simple form of standardization used in many areas of data analysis. For a unit vector $\mathbf{x}$, $\mathbf{x}^T\mathbf{x} = 1$.

Perhaps without knowing it, we've already seen many formulas involving the norm of a vector. Examples 8.1.2 and 8.1.3 show how some of the most important concepts in statistics can be represented using vector norms.

Example 8.1.2: Standard Deviation and Variance

Suppose a group of individuals has the following heights, measured in inches: (60, 70, 65, 50, 55). The mean height for this group is 60 inches. The formula for the **sample standard deviation** is typically given as

$$s = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{n-1}}$$

We want to subtract the mean from each observation, square the numbers, sum the result, take the square root and divide by $\sqrt{n-1}$.
If we let $\bar{\mathbf{x}} = \bar{x}\mathbf{e} = (60, 60, 60, 60, 60)$ be a vector containing the mean, and $\mathbf{x} = (60, 70, 65, 50, 55)$ be the vector of data then the standard deviation in matrix notation is:

$$s = \frac{1}{\sqrt{n-1}}\|\mathbf{x} - \bar{\mathbf{x}}\|_2 = 7.9$$

The **sample variance** of this data is merely the square of the sample standard deviation:

$$s^2 = \frac{1}{n-1}\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2$$

Example 8.1.3: Residual Sums of Squares

Another place we've seen a similar calculation is in linear regression. You'll recall the objective of our regression line is to minimize the sum of squared residuals between the predicted value $\hat{y}$ and the observed value y:

$$\sum_{i=1}^{n}(\hat{y}_i - y_i)^2.$$

In vector notation, we'd let $\mathbf{y}$ be a vector containing the observed data and $\hat{\mathbf{y}}$ be a vector containing the corresponding predictions and write this summation as

$$\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$$

In fact, any situation where the phrase "sum of squares" is encountered, the 2-norm is generally implicated.

Example 8.1.4: Coefficient of Determination, R^2

Since variance can be expressed using the Euclidean norm, so can the **coefficient of determination** or R^2.

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2}$$

Other useful norms and distances

1-norm, $\|\star\|_1$. If $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$ then the 1-norm of $\mathbf{X}$ is

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|.$$

This metric is often referred to as *Manhattan distance*, *city block distance*, or *taxicab distance* because it measures the distance between points along a rectangular grid (as a taxicab must travel on the streets of Manhattan, for example). When $\mathbf{x}$ and $\mathbf{y}$ are binary vectors, the 1-norm is called the **Hamming Distance**, and simply measures the number of elements that are different between the two vectors.

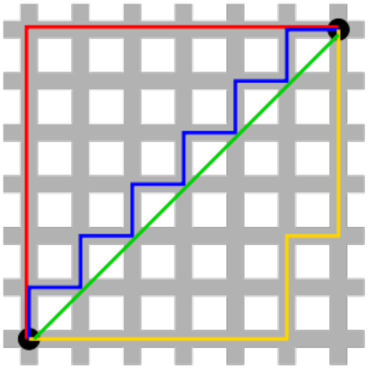


Figure 8.1: The lengths of the red, yellow, and blue paths represent the 1-norm distance between the two points. The green line shows the Euclidean measurement (2-norm).

∞-norm, $\| \star \|_\infty$. The infinity norm, also called the Supremum, or Max distance, is:

$$\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_p|\}$$

8.2 Inner Products

The inner product of vectors is a notion that you've already seen, it is what's called the *dot product* in most physics and calculus text books.

Definition 8.2.1: Vector Inner Product

The inner product of two $n \times 1$ vectors $\mathbf{x}$ and $\mathbf{y}$ is written $\mathbf{x}^T\mathbf{y}$ (or sometimes as $\langle \mathbf{x}, \mathbf{y} \rangle$) and is the sum of the product of corresponding elements.

$$\mathbf{x}^T\mathbf{y} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = x_1y_1 + x_2y_2 + \dots + x_ny_n = \sum_{i=1}^{n} x_iy_i.$$

When we take the inner product of a vector with itself, we get the square of the 2-norm:

$$\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2.$$

Inner products are at the heart of every matrix product. When we multiply two matrices, $\mathbf{X}_{m\times n}$ and $\mathbf{Y}_{n\times p}$, we can represent the individual elements of the result as inner products of rows of $\mathbf{X}$ and columns of $\mathbf{Y}$ as follows:

$$\mathbf{XY} = \begin{pmatrix} \mathbf{X}_{1\star} \\ \mathbf{X}_{2\star} \\ \vdots \\ \mathbf{X}_{m\star} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_{\star 1} & \mathbf{Y}_{\star 2} & \dots & \mathbf{Y}_{\star p} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{1\star}\mathbf{Y}_{\star 1} & \mathbf{X}_{1\star}\mathbf{Y}_{\star 2} & \dots & \mathbf{X}_{1\star}\mathbf{Y}_{\star p} \\ \mathbf{X}_{2\star}\mathbf{Y}_{\star 1} & \mathbf{X}_{2\star}\mathbf{Y}_{\star 2} & \dots & \mathbf{X}_{2\star}\mathbf{Y}_{\star p} \\ \mathbf{X}_{3\star}\mathbf{Y}_{\star 1} & \mathbf{X}_{3\star}\mathbf{Y}_{\star 2} & \dots & \mathbf{X}_{3\star}\mathbf{Y}_{\star p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_{m\star}\mathbf{Y}_{\star 1} & \dots & \ddots & \mathbf{X}_{m\star}\mathbf{Y}_{\star p} \end{pmatrix}$$

8.2.1 Covariance

Another important statistical measurement that is represented by an inner product is **covariance.** Covariance is a measure of how much two random variables change together. The statistical formula for covariance is given as

$$Covariance(\mathbf{x}, \mathbf{y}) = E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{y} - E[\mathbf{y}])] \tag{8.1}$$

where $E[\star]$ is the expected value of the variable. If larger values of one variable correspond to larger values of the other variable and at the same time smaller values of one correspond to smaller values of the other, then the covariance between the two variables is positive. In the opposite case, if larger values of one variable correspond to smaller values of the other and vice versa, then the covariance is negative. Thus, the *sign* of the covariance shows the tendency of the linear relationship between variables, however the *magnitude* of the covariance is not easy to interpret. Covariance is a population parameter - it is a property of the joint distribution of the random variables $\mathbf{x}$ and $\mathbf{y}$. Definition 8.2.2 provides the mathematical formulation for the *sample* covariance. This is our best estimate for the population parameter when we have data sampled from a population.

Definition 8.2.2: Sample Covariance

If $\mathbf{x}$ and $\mathbf{y}$ are $n \times 1$ vectors containing n observations for two different variables, then the **sample covariance** of $\mathbf{x}$ and $\mathbf{y}$ is given by

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

Where again $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are vectors that contain $\bar{x}$ and $\bar{y}$ repeated n times. It should be clear from this formulation that

$$cov(\mathbf{x}, \mathbf{y}) = cov(\mathbf{y}, \mathbf{x}).$$

When we have p vectors, $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$, each containing n observations for p different variables, the sample covariances are most commonly given by the **sample covariance matrix**, $\mathbf{\Sigma}$, where

$$\mathbf{\Sigma}_{ij} = cov(\mathbf{v}_i, \mathbf{v}_j).$$

This matrix is symmetric, since $\mathbf{\Sigma}_{ij} = \mathbf{\Sigma}_{ji}$. If we create a matrix $\mathbf{V}$ whose columns are the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_p$ *once the variables have been centered to have mean 0*, then the covariance matrix is given by:

$$cov(\mathbf{V}) = \mathbf{\Sigma} = \frac{1}{n-1}\mathbf{V}^T\mathbf{V}.$$

The j^{th} diagonal element of this matrix gives the variance $\mathbf{v}_j$ since

$$\begin{aligned} \mathbf{\Sigma}_{jj} = cov(\mathbf{v}_j, \mathbf{v}_j) &= \frac{1}{n-1}(\mathbf{v}_j - \bar{\mathbf{v}}_j)^T(\mathbf{v}_j - \bar{\mathbf{v}}_j) && (8.2) \\ &= \frac{1}{n-1}\|\mathbf{v}_j - \bar{\mathbf{v}}_j\|_2^2 && (8.3) \\ &= var(\mathbf{v}_j) && (8.4) \end{aligned}$$

When two variables are completely uncorrelated, their covariance is zero.

This lack of correlation would be seen in a covariance matrix with a diagonal structure. That is, if $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ are uncorrelated with individual variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2$ respectively then the corresponding covariance matrix is:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \vdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \sigma_p^2 \end{pmatrix}$$

Furthermore, for variables which are independent and identically distributed (take for instance the error terms in a linear regression model, which are assumed to independent and normally distributed with mean 0 and constant variance σ), the covariance matrix is a multiple of the identity matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \vdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}$$

Transforming our variables in a such a way that their covariance matrix becomes diagonal will be our goal in Chapter 11.

Theorem 8.2.1: Properties of Covariance Matrices

The following mathematical properties stem from Equation 8.1. Let $\mathbf{X}_{n\times p}$ be a matrix of data containing n observations on p variables. If $\mathbf{A}$ is a constant matrix (or vector, in the first case) then

$$cov(\mathbf{X}\mathbf{A}) = \mathbf{A}^T cov(\mathbf{X})\mathbf{A} \quad \text{and} \quad cov(\mathbf{X}+\mathbf{A}) = cov(\mathbf{X})$$

8.2.2 Mahalanobis Distance

Mahalanobis Distance is similar to Euclidean distance, but takes into account the correlation of the variables. This metric is relatively common in data mining applications like classification. Suppose we have p variables which have some covariance matrix, $\boldsymbol{\Sigma}$. Then the Mahalanobis distance between two observations, $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \dots & x_p \end{pmatrix}^T$ and $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \dots & y_p \end{pmatrix}^T$ is given by

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\mathbf{y})}.$$

If the covariance matrix is diagonal (meaning the variables are uncorrelated) then the Mahalanobis distance reduces to Euclidean distance normalized by the variance of each variable:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{s_i^2}} = \|\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\mathbf{y})\|_2.$$

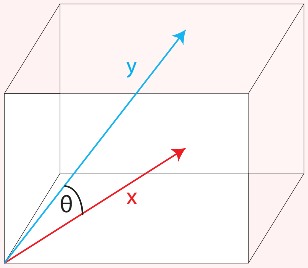
8.2.3 Angular Distance

The inner product between two vectors can provide useful information about their relative orientation in space and about their similarity. For example, to find the cosine of the angle between two vectors in n-space, the inner product of their corresponding unit vectors will provide the result. This cosine is often used as a measure of similarity or correlation between two vectors.

Definition 8.2.3: Cosine of Angle between Vectors

The cosine of the angle between two vectors in n-space is given by

$$\cos(\theta) = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}$$



This angular distance is at the heart of **Pearson's correlation coefficient**.

8.2.4 Correlation

Pearson's correlation is a normalized version of the covariance, so that not only the *sign* of the coefficient is meaningful, but its *magnitude* is meaningful in measuring the strength of the linear association.

Example 8.2.1: Pearson's Correlation and Cosine Distance

You may recall the formula for Pearson's correlation between variable **x** and **y** with a sample size of n to be as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

If we let $\bar{\mathbf{x}}$ be a vector that contains $\bar{x}$ repeated n times, like we did in Example 8.1.2, and let $\bar{\mathbf{y}}$ be a vector that contains $\bar{y}$ then Pearson's coefficient can be written as:

$$r = \frac{(\mathbf{x} - \bar{\mathbf{x}})^T(\mathbf{y} - \bar{\mathbf{y}})}{\|\mathbf{x} - \bar{\mathbf{x}}\|\|\mathbf{y} - \bar{\mathbf{y}}\|}$$

In other words, it is just the cosine of the angle between the two vectors once they have been *centered* to have mean 0.
This makes sense: correlation is a measure of the extent to which the two variables share a line in space. If the cosine of the angle is positive or negative one, this means the angle between the two vectors is 0° or 180°, thus, the two vectors are perfectly correlated or *collinear*.

It is difficult to visualize the angle between two variable vectors because they exist in n-space, where n is the number of observations in the dataset. Unless we have fewer than 3 observations, we cannot draw these vectors or even picture them in our minds. As it turns out, this angular measurement does translate into something we can conceptualize: Pearson's correlation coefficient is the angle formed between the two possible regression lines using the centered data: **y** regressed on **x** and **x** regressed on **y**. This is illustrated in Figure 8.2.

To compute the matrix of pairwise correlations between variables $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_p$ (columns containing n observations for each variable), we'd first center them to have mean zero, then normalize them to have length $\|\mathbf{x}_i\| = 1$ and then compose the matrix

$$\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\mathbf{x}_3|\dots|\mathbf{x}_p].$$

Using this centered and normalized data, the correlation matrix is simply

$$\mathbf{C} = \mathbf{X}^T\mathbf{X}.$$

8.3 Orthogonality

Orthogonal (or perpendicular) vectors have an angle between them of 90°, meaning that their cosine (and subsequently their inner product) is zero.

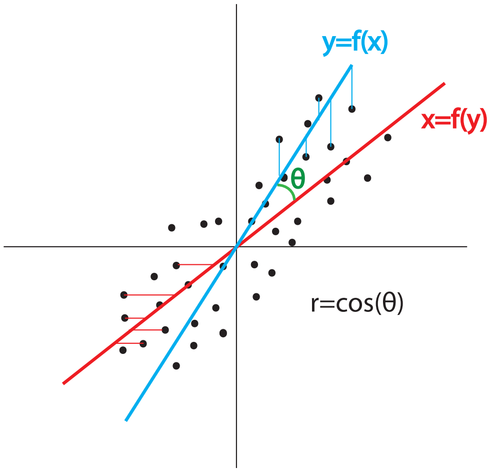


Figure 8.2: Correlation Coefficient r and Angle between Regression Lines

Definition 8.3.1: Orthogonality

Two vectors, $\mathbf{x}$ and $\mathbf{y}$, are **orthogonal** in n-space if their inner product is zero:

$$\mathbf{x}^T\mathbf{y} = 0$$

Combining the notion of orthogonality and unit vectors we can define an orthonormal set of vectors, or an orthonormal matrix. Remember, for a unit vector, $\mathbf{x}^T\mathbf{x} = 1$.

Definition 8.3.2: Orthonormal Sets

The $n \times 1$ vectors $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_p\}$ form an **orthonormal set** if and only if

1. $\mathbf{x}_i^T\mathbf{x}_j = 0$ when $i \neq j$ and
2. $\mathbf{x}_i^T\mathbf{x}_i = 1$ (equivalently $\|\mathbf{x}_i\| = 1$)

In other words, an orthonormal set is a collection of *unit vectors which are mutually orthogonal.*

If we form a matrix, $\mathbf{X} = (\mathbf{x}_1|\mathbf{x}_2|\mathbf{x}_3|\dots|\mathbf{x}_p)$, having an orthonormal set of vectors as columns, we will find that multiplying the matrix by its transpose provides a nice result:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \mathbf{x}_3^T \\ \vdots \\ \mathbf{x}_p^T \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \dots & \mathbf{x}_p \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_2 & \mathbf{x}_1^T\mathbf{x}_3 & \dots & \mathbf{x}_1^T\mathbf{x}_p \\ \mathbf{x}_2^T\mathbf{x}_1 & \mathbf{x}_2^T\mathbf{x}_2 & \mathbf{x}_2^T\mathbf{x}_3 & \dots & \mathbf{x}_2^T\mathbf{x}_p \\ \mathbf{x}_3^T\mathbf{x}_1 & \mathbf{x}_3^T\mathbf{x}_2 & \mathbf{x}_3^T\mathbf{x}_3 & \dots & \mathbf{x}_3^T\mathbf{x}_p \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{x}_p^T\mathbf{x}_1 & \dots & \dots & \ddots & \mathbf{x}_p^T\mathbf{x}_p \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} = \mathbf{I}_p$$

We will be particularly interested in these types of matrices when they are square. If **X** is a square matrix with orthonormal columns, the arithmetic above means that the inverse of **X** is $\mathbf{X}^T$ (i.e. **X** also has orthonormal rows):

$$\mathbf{X}^T\mathbf{X} = \mathbf{X}\mathbf{X}^T = I.$$

Square matrices with orthonormal columns are called orthogonal matrices.

Definition 8.3.3: Orthogonal (or Orthonormal) Matrix

A *square* matrix, **U** with orthonormal columns also has orthonormal rows and is called an **orthogonal matrix**. Such a matrix has an inverse which is equal to it's transpose,

$$\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$$

8.4 Outer Products

The outer product of two vectors $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$, written $\mathbf{x}\mathbf{y}^T$, is an $m \times n$ matrix with rank 1. To see this basic fact, lets just look at an example.

Example 8.4.1: Outer Product

Let $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$ and let $\mathbf{y} = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$. Then the outer product of $\mathbf{x}$ and $\mathbf{y}$ is:

$$\mathbf{x}\mathbf{y}^T = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 3 \\ 4 & 2 & 6 \\ 6 & 3 & 9 \\ 8 & 4 & 12 \end{pmatrix}$$

which clearly has rank 1. It should be clear from this example that computing an outer product will always result in a matrix whose rows and columns are multiples of each other.

Example 8.4.2: Centering Data with an Outer Product

As we've seen in previous examples, many statistical formulas involve the *centered* data, that is, data from which the mean has been subtracted so that the new mean is zero. Suppose we have a matrix of data containing observations of individuals' heights (h) in inches, weights (w), in pounds and wrist sizes (s), in inches:

$$\mathbf{A} = \begin{array}{c} \\ person_1 \\ person_2 \\ person_3 \\ person_4 \\ person_5 \end{array} \begin{array}{c} \begin{array}{ccc} h & w & s \end{array} \\ \begin{pmatrix} 60 & 102 & 5.5 \\ 72 & 170 & 7.5 \\ 66 & 110 & 6.0 \\ 69 & 128 & 6.5 \\ 63 & 130 & 7.0 \end{pmatrix} \end{array}$$

The average values for height, weight, and wrist size are as follows:

$$\bar{h} = 66 \quad (8.5)$$
$$\bar{w} = 128 \quad (8.6)$$
$$\bar{s} = 6.5 \quad (8.7)$$

To center all of the variables in this data set simultaneously, we could compute an outer product using a vector containing the means and a vector of all ones:

$$\begin{pmatrix} 60 & 102 & 5.5 \\ 72 & 170 & 7.5 \\ 66 & 110 & 6.0 \\ 69 & 128 & 6.5 \\ 63 & 130 & 7.0 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 66 & 128 & 6.5 \end{pmatrix}$$

$$= \begin{pmatrix} 60 & 102 & 5.5 \\ 72 & 170 & 7.5 \\ 66 & 110 & 6.0 \\ 69 & 128 & 6.5 \\ 63 & 130 & 7.0 \end{pmatrix} - \begin{pmatrix} 66 & 128 & 6.5 \\ 66 & 128 & 6.5 \\ 66 & 128 & 6.5 \\ 66 & 128 & 6.5 \\ 66 & 128 & 6.5 \end{pmatrix}$$

$$= \begin{pmatrix} -6.0000 & -26.0000 & -1.0000 \\ 6.0000 & 42.0000 & 1.0000 \\ 0 & -18.0000 & -0.5000 \\ 3.0000 & 0 & 0 \\ -3.0000 & 2.0000 & 0.5000 \end{pmatrix}$$

Exercises

1. Let $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ -4 \\ -2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$.

   a. Determine the Euclidean distance between **u** and **v**.

   b. Find a vector of unit length in the direction of **u**.

   c. Determine the cosine of the angle between **u** and **v**.

   d. Find the 1- and ∞-norms of **u** and **v**.

   c. Suppose these vectors are observations on four independent variables, which have the following covariance matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

   Determine the Mahalanobis distance between **u** and **v**.

2. Let

$$\mathbf{U} = \frac{1}{3}\begin{pmatrix} -1 & 2 & 0 & -2 \\ 2 & 2 & 0 & 1 \\ 0 & 0 & 3 & 0 \\ -2 & 1 & 0 & 2 \end{pmatrix}$$

   a. Show that $\mathbf{U}$ is an orthogonal matrix.

   b. Let $\mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$. Solve the equation $\mathbf{Ux} = \mathbf{b}$.

3. Write a matrix expression for the correlation matrix, $\mathbf{C}$, for a matrix of *centered* data, $\mathbf{X}$, where $\mathbf{C}_{ij} = r_{ij}$ is Pearson's correlation measure between variables $\mathbf{x}_i$ and $\mathbf{x}_j$. To do this, we need more than an inner product, we need to normalize the rows and columns by the norms $\|\mathbf{x}_i\|$. For a hint, see Exercise 2 in Chapter **??**.

4. Suppose you have a matrix of data, $\mathbf{A}_{n\times p}$, containing n observations on p variables. Develop a matrix formula for the standardized data (where the mean of each variable should be subtracted from the corresponding column before dividing by the standard deviation). *Hint: use Exercises 1(f) and 4 from Chapter* **??** *along with Example 8.4.2.*

5. Explain why, for any norm or distance metric,

$$\|\mathbf{x} - \mathbf{y}\| = \|\mathbf{y} - \mathbf{x}\|$$

6. Find two vectors which are orthogonal to $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

7. **Pythagorean Theorem.** Show that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if and only if

$$\|\mathbf{x} + \mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2$$

   (Hint: Recall that $\|\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{x}$*)*