Scientific Computing for Biologists

Data as Vectors: Geometry of Bivariate Relationships I

Instructor: Paul M. Magwene

Overview of Lecture

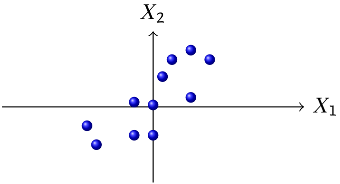
- Variable space/Subject space representations
- Vector Geometry
  - Vectors are directed line segments
  - Vector length
- Vector Arithmetic
  - Addition, subtraction
  - Scalar multiplication
  - Linear combinations of vectors
  - Dot product and projection
- Vector representations of multivariate data
  - Mean as projection in subject space
  - Bivariate regression in geometric terms

Variable Space Representation of a Data Set

Consider a data set in which we've measured variables
$\mathsf{X} = X_1, X_2, \ldots, X_p$, on a set of subjects (objects) $a_1, ..., a_n$.

	X_1	X_2
a_1	0.9	1.4
a_2	1.1	1.7
$\vdots$	$\vdots$	$\vdots$
a_n	0.5	1.55

Such data is most often represented by drawing the objects as points in space of dimension p. This is the *variable space representation* of the data.

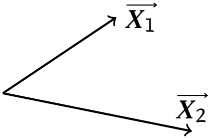


Subject Space Representation of a Data Set

An alternate representation is to consider the variables in the space of the subjects. This is the *subject space* representation.

How do we come up with a useful representation of variables in subject space?

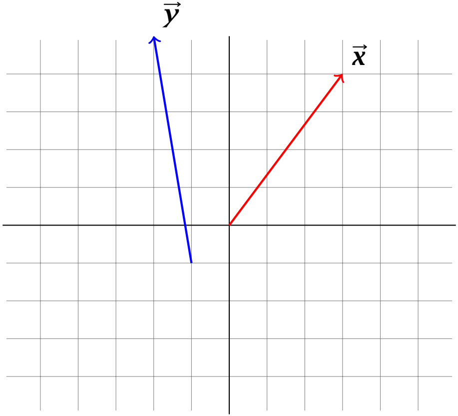
- Let the variables be represented by centered vectors
  - lengths of vectors are proportional to standard deviation
  - angle between vectors represents association or similarity



This representation of variables as vectors in the space of the subjects is the view that we'll develop over the next few lectures.

Vector Geometry

Vectors are directed line segments.

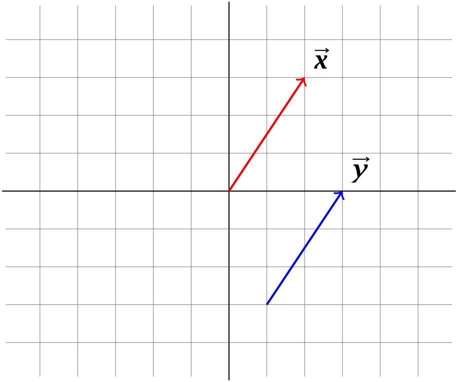


All of the figures and algebraic formulas I show you apply to n-dimensional vectors.

Vector Geometry

Vectors have direction and length:

$$\vec{\boldsymbol{x}} = [x_1, x_2]' = [2, 3]'; \; |\vec{\boldsymbol{x}}| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

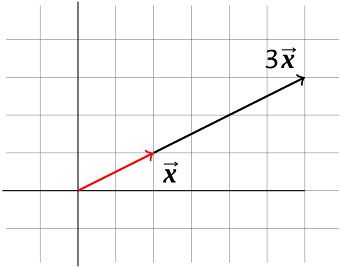


Often starting point is ignored, in which case $\vec{\boldsymbol{x}} = \vec{\boldsymbol{y}}$.

Scalar Multiplication of a Vector

Let k be a scalar.

$$k\vec{\boldsymbol{x}} = \begin{bmatrix} kx_1 \\ kx_2 \\ \vdots \\ kx_n \end{bmatrix}$$

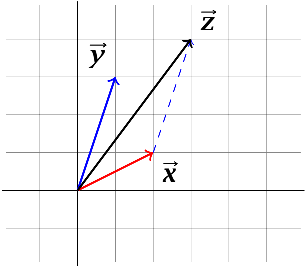


$\vec{\boldsymbol{x}} = [2, 1]'$; $3\vec{\boldsymbol{x}} = [6, 3]'$.

Vector Addition

Let $\vec{\boldsymbol{x}} = [2, 1]'$; $\vec{\boldsymbol{y}} = [1, 3]'$

$$\vec{\boldsymbol{z}} = \vec{\boldsymbol{x}} + \vec{\boldsymbol{y}} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

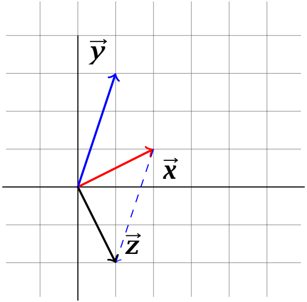


Addition follows the ‘head-to-tail’ rule.

Vector Subtraction

Let $\vec{x} = [2, 1]'$; $\vec{y} = [1, 3]'$

$$\vec{z} = \vec{x} - \vec{y} = \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

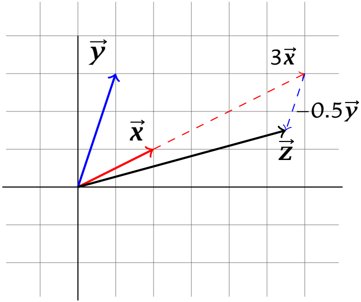


Follow the addition rule for $-1\vec{y}$.

Linear Combinations of Vectors

A linear combination of vectors is of the form $z = b_1\vec{\boldsymbol{x}} + b_2\vec{\boldsymbol{y}}$

$$\vec{\boldsymbol{z}} = 3\vec{\boldsymbol{x}} - 0.5\vec{\boldsymbol{y}} = 3\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 0.5\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

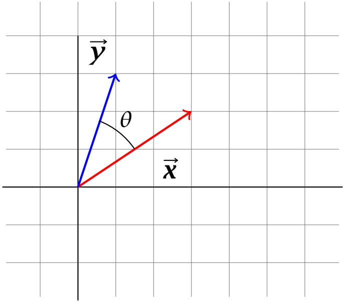


Dot Product

The dot (inner) product of two vectors, $\vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}}$ is a scalar.

$$\begin{aligned} \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}} &= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \\ &= |\vec{\boldsymbol{x}}||\vec{\boldsymbol{y}}| \cos \theta \end{aligned}$$

where θ is the angle (in radians) between $\vec{\boldsymbol{x}}$ and $\vec{\boldsymbol{y}}$



$\vec{\boldsymbol{x}} = [3, 2]', \vec{\boldsymbol{y}} = [1, 3]'; \ \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}} = \sqrt{13}\sqrt{10} \cos \theta = 9$

Useful Geometric Quantities as Dot Product

Length:

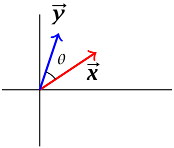
$$
\begin{aligned}
|\vec{x}|^2 &= \vec{x} \cdot \vec{x} = x_1^2 + x_2^2 + \cdots + x_n^2 \\
|\vec{y}|^2 &= \vec{y} \cdot \vec{y}
\end{aligned}
$$

Distance:

$$|\vec{x} - \vec{y}|^2 = \vec{x} \cdot \vec{x} + \vec{y} \cdot \vec{y} - 2\vec{x} \cdot \vec{y}$$

Angle:

$$\cos\theta = \vec{x} \cdot \vec{y} / (|x||y|)$$



Dot Product Properties

Some additional properties of the dot product that are useful to know:

$$
\begin{aligned}
\vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}} &= \vec{\boldsymbol{y}} \cdot \vec{\boldsymbol{x}} \text{ (commutative)} \\
\vec{\boldsymbol{x}} \cdot (\vec{\boldsymbol{y}} + \vec{\boldsymbol{z}}) &= \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}} + \vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{z}} \text{ (distributive)} \\
(k\vec{\boldsymbol{x}}) \cdot \vec{\boldsymbol{y}} &= \vec{\boldsymbol{x}} \cdot (k\vec{\boldsymbol{y}}) = k(\vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}}) \text{ where } k \text{ is a scalar} \\
\vec{\boldsymbol{x}} \cdot \vec{\boldsymbol{y}} &= 0 \text{ iff } \vec{\boldsymbol{x}} \text{ and } \vec{\boldsymbol{y}} \text{ are orthogonal}
\end{aligned}
$$

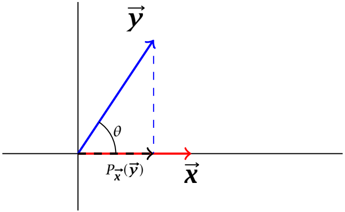
Vector Projection

The projection of $\vec{y}$ onto $\vec{x}$, $P_{\vec{x}}(\vec{y})$, is the vector obtained by placing $\vec{y}$ and $\vec{x}$ tail to tail and dropping a line, perpendicular to $\vec{x}$, from the head of $\vec{y}$ onto the line defined by $\vec{x}$.

$$P_{\vec{x}}(\vec{y}) = \left(\frac{\vec{x} \cdot \vec{y}}{|\vec{x}|}\right) \frac{\vec{x}}{|\vec{x}|} = \left(\frac{\vec{x} \cdot \vec{y}}{|\vec{x}|^2}\right) \vec{x}$$

The component of $\vec{y}$ in $\vec{x}$, $C_{\vec{x}}(\vec{y})$, is the length of $P_{\vec{x}}(\vec{y})$.

$$C_{\vec{x}}(\vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}|} = |\vec{y}| \cos \theta$$

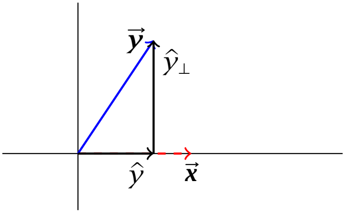


Vector Projection II

$\vec{y}$ can be decomposed into two parts:

1. a vector parallel to $\vec{x}$, $\hat{y} = P_{\vec{x}}(\vec{y})$,
2. a vector perpendicular to $\vec{x}$, $\hat{y}_{\perp}$.

$$\vec{y} = \hat{y} + \hat{y}_{\perp}$$



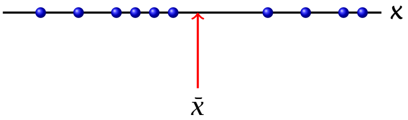
- $\hat{y}_{\perp}$ is *orthogonal* to $\hat{y}$ and $\vec{x}$.
- $\hat{y}$ is the closest vector to $\vec{y}$ in the subspace defined by $\vec{x}$

Vector Geometry of Simple Statistics

Geometry of the Mean in Variables Space

The mean is a single number summary of a set (vector) of values, $\vec{\boldsymbol{x}}$. Mathematically, the mean is the value that minimizes the following quantity:

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



Algebraic and Vector Formulas for the Mean

Let $\vec{\boldsymbol{x}} = [x_1, x_2, \ldots, x_n]'$

Algebraic formula for the mean of $\vec{\boldsymbol{x}}$:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Vector formula for the mean:

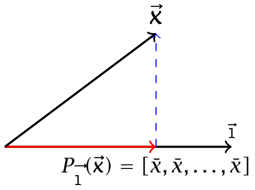
$$\begin{aligned} \bar{x} &= \frac{\vec{1} \cdot \vec{\boldsymbol{x}}}{\vec{1} \cdot \vec{1}} \\ &= \frac{\vec{1} \cdot \vec{\boldsymbol{x}}}{|\vec{1}|^2} \end{aligned}$$

Where $\vec{1} = [1, 1, \ldots, 1]'$ is the 1-vector of dimension n. (Note that the 1-vector is not the same as a unit vector!)

Geometry of the Mean in Subject Space

- The mean can be interpretted in terms of the projection of $\vec{\boldsymbol{x}}$ onto the 1-vector:

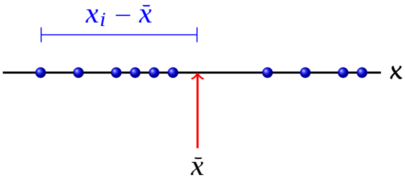
$$
\begin{aligned}
P_{\vec{1}}(\vec{\boldsymbol{x}}) &= \left(\frac{\vec{1}\cdot\vec{\boldsymbol{x}}}{|\vec{1}|^2}\right)\vec{1} \\
&= \bar{x}\vec{1} \\
&= [\bar{x},\bar{x},\ldots,\bar{x}]'
\end{aligned}
$$



Variable Space Geometry of Sample Variance

Sample variance is proportional to the sum of squared deviates about the mean:

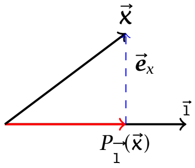
$$S_x^2 = \frac{1}{(n-1)} \sum (x_i - \bar{x})^2$$



Vector Geometry of Sample Variance

- Let $\overrightarrow{\boldsymbol{e_x}} = \vec{\boldsymbol{x}} - \bar{x}\vec{1}$
- The sample variance can be expressed in terms of dots products of $\overrightarrow{\boldsymbol{e_x}}$ with itself:

$$S_x^2 = \frac{\vec{\boldsymbol{e}}_x \cdot \vec{\boldsymbol{e}}_x}{n-1} = \frac{|\vec{\boldsymbol{e}}_x|^2}{n-1}$$



Mean centering

In the previous slide, we considered the vector:

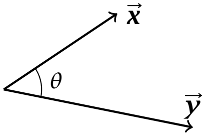
$$\overrightarrow{\boldsymbol{e_x}} = \vec{\boldsymbol{x}} - \bar{x}\vec{1}$$

We can think of $\vec{\boldsymbol{e}}_x$ as a "mean centered" version of $\vec{\boldsymbol{x}}$, i.e. it's the vector we get when we subtract the mean of $\vec{\boldsymbol{x}}$, $\bar{x}$, from every element of $\vec{\boldsymbol{x}}$.

For convenience, I will sometimes state the variables of interest are mean centered and use the notation $\vec{\boldsymbol{x}}$ instead of $\vec{\boldsymbol{e}}_x$ so as to avoid a proliferation of subscripts.

Covariance and Correlation in Vector Geometric Terms

Let X and Y be mean centered variables, and let $\vec{x}$ and $\vec{y}$ be their corresponding mean centered vector representations.



Vector formulas for covariance and correlation:

$$\text{Covariance: } \operatorname{cov}(X,Y) = \frac{\vec{x}\cdot\vec{y}}{n-1}$$

$$\text{Correlation: } \operatorname{corr}(X,Y) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \cos\theta$$

Geometric interpretation of correlation

The correlation between two variables X and Y is equivalent to the cosine of the angle between their mean-centered vector representations!